FROM EXTRACTS TO ABSTRACTS: HUMAN SUMMARY PRODUCTION OPERATIONS FOR COMPUTER-AIDED SUMMARISATION

Laura Hasler BA (Hons)

A thesis submitted in partial fulfilment of the
requirements of the University of Wolverhampton
for the degree of Doctor of Philosophy

April 2007

Signature……..

Date……..

Abstract

This thesis is concerned with the field of computer-aided summarisation, which has emerged at the confluence of the separate but related fields of human and automatic summarisation. Due to the poor quality of the readability and coherence of automatically produced extracts, computer-aided summarisation (CAS) is a viable working option to fully automatic summarisation. CAS allows a human summariser to post-edit automatically produced extracts to improve their readability and coherence. In order to best utilise the concept of computer-aided summarisation, reliable ways of improving the coherence and readability of extracts when transforming them into abstracts must be established.

To achieve this, a corpus-based analysis of the operations a human summariser applies to extracts to transform them into abstracts is presented. The corpus developed here is a corpus of pairs of news texts annotated for important information (i.e., human-produced extracts) and the human-produced abstracts corresponding to these extracts. The creation of this corpus simulates the computer-aided summarisation process to enable a reliable investigation into the operations used. A detailed classification of human summary production operations is proposed, with examples which highlight the common linguistic realisations and functions of the operations identified in the corpus. The classification is then used as a basis for guidelines which can be given to users of computer-aided summarisation systems in order to ensure that the summaries they produce are of a consistently high quality.

The human summary production operations are applied to extracts using the guidelines in order to evaluate them. Evaluation is performed using a metric developed for Centering Theory, a discourse theory of local coherence and salience, which constitutes a new evaluation method. This is appropriate because existing methods of evaluating summaries are unsuitable. A set of both automatic and human-produced extracts and their corresponding abstracts are evaluated, and a comparison is made with evaluations given by a human judge. The evaluation shows that when the operations are applied to extracts using the guidelines, there is an improvement in the readability and coherence of the resulting abstracts.

Acknowledgements

There are a number of people whom I wish to thank for the help they have given me during my PhD research. First of all, I would like to thank my supervisory team, Ruslan Mitkov, Michael Hoey and Constantin Orasan, for their support and comments on my work. A special thank you to Constantin, who spent countless hours reading and discussing my thesis with limitless patience and enthusiasm. He will be surprised to hear that I do not have enough words to thank him properly. My gratitude also goes to my examiners, Sylviane Cardey, Frances Johnson and Tom Dickins, for their constructive questions, discussion and suggestions, as well as to the AHRB, who provided the funding for my PhD via the CAST project.

I owe my thanks to Nikiforos Karamanis, for his useful and enthusiastic discussions about Centering Theory and other aspects of my work, and to Shiyan Ou for reading my chapters on summarisation. Thanks also to Constantin and Le An Ha for their help with the formatting of my thesis. My proofreaders, Aidan Byrne, Emma Costelloe and David Hasler, employed their skills at very short notice and deserve a big thank you. I am grateful to Aidan for his participation in various aspects of the evaluation, and also to the other judges who took part in my experiments: Meena Dhanda, Richard Evans, Mark Jones, Amelia Smith and Gina Sutherland.

The friendship and support of the various members of the research group at Wolverhampton, past and present, has made a difference to my PhD studies – thank you. And last but by no means least, I want to say thank you to my family and friends, for their support and encouragement during the last few years, and for being very neglected recently.

Table of contents

List of tables

List of abbreviations

ANSI - American National Standards Institute

BNC – British National Corpus

CAS – Computer-Aided Summarisation

CAST – Computer-Aided Summarisation Tool

Cb – Backward Looking Center

Cf – Forward Looking Center

Conj - Conjunction

Cp – Preferred Center

CT – Centering Theory

DUC – Document Understanding Conferences

GATE – General Architecture for Text Engineering

KWIC – Key Words In Context

NISO – National Information Standards Organization

NLG – Natural Language Generation

NLP – Natural Language Processing

NP – Noun Phrase

PALinkA – Perspicuous and Adjustable Links Annotator

PP – Prepositional Phrase

ROUGE – Recall-Oriented Understudy for Gisting Evaluation

RST – Rhetorical Structure Theory

S - Subject

SEE – Summary Evaluation Environment

U – Utterance

V – Verb

VP – Verb Phrase

Chapter 1. Introduction

1.1 Overview

Automatic summarisation has been widely investigated since its conception in the 1950s. However, despite extensive research, the quality of the summaries produced is still not comparable in terms of coherence and readability with summaries produced by humans. This is because the focus to date has mainly been on the information content of summaries, and because formalising models which can measure and deal with coherence is difficult. At present, there are well-established methods of *extracting* information from source texts, but little work has been done on how this information is presented as a coherent text. This thesis addresses that issue by investigating the operations a human summariser applies to improve the coherence and readability of extracts by transforming them into abstracts. There are a number of proposed revision operations in automatic extracting, yet these do not account for many of the changes human summarisers make when post-editing an extract, and are very restricted. Research in the 1970s and 1980s on automatic *abstracting* dealt with some of the issues involved, but again, the focus was on information rather than readability.

Because the aim of automatic summarisation is to help users of such systems to save time in one way or another, it seems necessary to account not just for informativeness but also for coherence and readability in a system, so that users do not have to waste time attempting to process incoherent texts. Current summarisation

systems do not, for the main part, address such issues, and this thesis argues that it is necessary to return to the human summarisation process to improve matters. Recently, the concept of *computer-aided* summarisation was developed as a means of integrating the human and automatic summarisation processes (Orasan, Mitkov and Hasler 2003). The approach presents a feasible alternative for users because it recognises the shortcomings of fully automatic processes and allows a human summariser to interact with an automatic system to produce the best possible summary. However, in order to benefit fully from such a system, users (who are not necessarily experts in summarisation or linguistics) need to know exactly how best to improve the automatic output to transform it into a high quality, coherent and readable summary.

Whilst guidelines exist for professional abstractors in the field of human summarisation and are usually available to human annotators in the field of automatic summarisation, there are none which explicitly deal with transforming extracts into abstracts. This transformation is the task carried out by users of computer-aided summarisation systems. Just as guidelines in human and automatic summarisation are necessary to facilitate consistency and quality of the task in hand, they are also necessary in the field of computer-aided summarisation. To date, such a resource does not exist. In order to develop these guidelines for the task of post-editing computer-aided extracts, an analysis of the operations a human summariser applies to the texts is crucial. The investigation in this thesis develops guidelines for this task, based on a corpus analysis of human *summary production* operations, and evaluates the extent to which they, and the analysis they are based on, are useful, by assessing the coherence and readability of the resulting abstracts.

The novelty of this work lies in its interaction between the fields of human and automatic summarisation. Previous work in human summarisation has concentrated on texts such as research papers, particularly scientific research papers. There is little research into the way humans summarise or deal with certain types of texts which are common in the field of automatic summarisation, such as news texts. However, whilst extraction methods can be used relatively reliably in automatic summarisation, there are problems with the coherence of extracts which are often ignored. By using stages identified in the human summarisation process, the best parts of human and automatic summarisation are employed in computer-aided summarisation. Human summarisation is therefore taken as a model by which to produce computer-aided summaries of news texts.

1.2 Aims and Contributions

This thesis provides original contributions to research in the fields of human summarisation and computer-aided summarisation, and within computer-aided summarisation in the area of evaluation. The contributions to computer-aided summarisation are also, by virtue of association, contributions to the field of automatic summarisation.[1] This thesis focuses on improving the readability and coherence of extracts by transforming them into abstracts, thereby addressing a major shortcoming of existing automatic summarisation methods/systems. The **main aim** of the thesis is to identify ways of improving the quality, in terms of coherence

[1] Computer-aided summarisation developed out of automatic summarisation, and therefore is closely related to it.

and readability, of extracts, which can then be applied in the field of computer-aided summarisation. In order to achieve this main aim, **several goals** need to be met.

First of all, it is necessary to introduce the general area of summarisation and issues which need to be considered during any kind of summarisation process. Linked to this is the need to introduce the separate but related fields of human and automatic summarisation, as the thesis addresses both fields. Computer-aided summarisation, which has emerged at the confluence of these two fields, also has to be described as it provides the context in which this research is carried out.

Having established the basic context of the research and identified the shortcomings which need to be addressed, it is then necessary to investigate exactly how the readability and coherence of extracts can be improved. To do this, the operations used by a human summariser to transform extracts into abstracts should be examined. This cannot be completed without a collection of texts which demonstrate such operations, meaning that an appropriate corpus must be developed containing extracts and abstracts produced from these extracts. To develop this corpus, extracts must first of all be obtained. This is achieved by human annotation because the extracts will be of a better quality in terms of informativeness than extracts produced automatically, allowing the focus to remain wholly on the transformations applied. Guidelines are needed to facilitate consistency, and therefore an appropriate set must be formulated. This in turn necessitates a review of existing guidelines.

Once the corpus has been built, it has to be analysed to identify the operations used by the human summariser, and this analysis needs to be presented in an appropriate

way: a classification of human summary production operations. A review of existing related work allows the operations to be compared with the findings of other researchers. It is then necessary to formulate this classification into a set of summary production guidelines which can be given to a human summariser to help them transform extracts into abstracts.

To assess the usefulness of the investigation, and the extent to which the operations identified and the guidelines developed from them can improve the coherence and readability of a text by transforming it from an extract into an abstract, an evaluation is required. To achieve this, existing evaluation methods must be considered, and if no appropriate methods are available, a new method should be proposed and developed. The guidelines need to be employed to create abstracts from extracts, and this should be done on a different set of texts to those used to identify the operations to ensure that they can be applied to other extracts.

To summarise, the **original contributions** of this thesis, presented in the order in which they are addressed, are as follows:

1. A set of guidelines to annotate source texts for important information which results in extracts for a corpus of (extract, abstract) pairs.
2. A corpus of (extract, abstract) pairs to enable the analysis of human summary production operations.
3. A corpus-based classification of operations applied to extracts by a human summariser in order to arrive at readable and coherent abstracts.
4. A set of summary production guidelines derived from the classification.

5. The development of Centering Theory (Grosz, Joshi and Weinstein 1995) as an evaluation method to assess the operations due to the unsuitability of existing evaluation methods for the task.
6. An evaluation of the coherence and readability of abstracts produced using the summary production operations and, by default, the guidelines issued to a human summariser and the operations themselves.

1.3 Structure of the thesis

The thesis comprises three parts. Chapters 2 and 3 provide the background for the remaining chapters, offering an overview of summarisation and the various fields within it. Chapters 4, 5 and 6 constitute the first part of the original contributions: guidelines for summarisation, an annotated corpus, a corpus analysis and classification of human summary production operations, and a set of guidelines for summary production. Chapter 7 presents the second part of the original contributions by evaluating an application of the work completed in the previous three chapters using a new evaluation method.

Chapter 2 gives an introduction to summaries and summarisation. This is achieved via a discussion of *context factors* (Sparck Jones 1999; Tucker 1999), which are used to describe issues that need to be taken into account when creating any kind of summary. The chapter then addresses the field of *human summarisation*, as this is a necessary precursor to any discussion of automatic summarisation. The difference between professional and non-professional summarisers is established, as are the three stages of the summarisation process (Endres-Niggemeyer 1998). These three stages, *document exploration*, *relevance assessment* and *summary production* are

taken as a general model for automatic and computer-aided summarisation discussed in the remainder of the thesis.

Chapter 3 follows on from human summarisation, providing a review of major work in *automatic summarisation*. Automatic abstracting and automatic extracting are addressed, and the positive and negative aspects of the summaries produced by each are detailed. Chapter 3 then introduces the recently developed concept of *computer-aided summarisation* (Orasan, Mitkov and Hasler 2003), which is presented as a feasible alternative to fully automatic methods in the current automatic summarisation climate, as these tend to ignore issues of coherence and readability in the production of summaries. Computer-aided summarisation is also justified in terms of the three stages of human summarisation presented in Chapter 2, and is the framework in which the original contributions of this thesis are developed. Chapters 2 and 3 together provide a justification of investigating the *summary production* stage to improve the readability and coherence of extracts in the rest of the thesis.

Chapter 4 presents the annotated corpus of news texts exploited in Chapters 5 and 6 and a set of guidelines used for the annotation. The corpus contains 43 (extract, abstract) pairs of texts, the extracts being produced by a human annotator using the guidelines formulated in this chapter. To place these guidelines in context and justify their development, an overview of existing summarisation guidelines available to professional abstractors, and a discussion of the need for such guidelines, is given. The use of human-produced extracts as opposed to automatically produced ones as a starting point for the transformation into abstracts is necessary because of the better quality in terms of informativeness, which allows the third stage of the

summarisation process (summary production) to be the sole focus of Chapters 5 and 6. Although the focus of this thesis is on the summary production stage of summarisation, the first two stages also have to be accounted for if the computer-aided summarisation process is to be simulated to fully appreciate the task. Document exploration and relevance assessment are dealt with by the human annotation of texts for important information, resulting in a corpus of extracts which are then used to produce abstracts (summary production).

Chapter 5 introduces the classification of human summary production operations resulting from the analysis of the corpus developed in Chapter 4. Two general types of operation are identified: *atomic* and *complex*, the difference being that complex operations are made up of atomic ones. The chapter focuses on *atomic* operations, and two classes are distinguished: *deletion* and *insertion*, each comprising a number of sub-operations which can be identified by certain *triggers*. It is argued that whilst the function of the units to which operations are applied is vital in deciding when sub-operations are appropriate, it is necessary to classify the sub-operations in terms of their form. This is due to the fact that the field in which this research is conducted is computer-aided summarisation, and surface forms are easier for machines to recognise than the functions of units. To contextualise this investigation, a brief review of existing work related to the operations humans use to create summaries is offered, although these focus on the production of a summary from a source, and not that of an abstract from an extract.

Chapter 6 continues the classification based on the corpus analysis, examining cases of the *complex* summary production operations *replacement*, *reordering* and

merging. Although the sub-operations of *replacement* can be classified in the same way as those of the atomic operations in Chapter 5, this is a much more challenging task for reordering and merging. Indeed, the sub-operations of *reordering* can only be classified in terms of their function. *Merging* is found to be the operation which best captures the essence of an *abstract* as opposed to an *extract*. A discussion of possibilities for future automation of the sub-operations is given, and a novel set of guidelines to facilitate the human post-editing of computer-aided extracts, based on the classification, is presented. To highlight the complexity of the operations applied by the human summariser during summary production, an example of an (extract, abstract) pair from the corpus is also analysed.

Chapter 7 evaluates the extent to which the classification of summary production operations and the guidelines formulated from it are useful to a human summariser when post-editing extracts. The guidelines are used to produce abstracts from extracts, and these are assessed in terms of coherence using a metric developed for Centering Theory (Grosz, Joshi and Weinstein 1995), a theory of local coherence and salience. The development of Centering Theory (CT) as an evaluation method is an original contribution to the area of evaluation, necessary because of the inadequacy of existing evaluation methods in automatic summarisation to fully assess the task undertaken in Chapters 5 and 6. An overview of relevant existing evaluation methods, and an introduction to other computational linguistics work using CT is given to justify this development. Human judgment is also obtained to assess readability, and to 'evaluate' the CT evaluation method. The evaluation shows that, overall, the summary production operations applied to extracts to transform them

into abstracts improves their coherence and readability, making them suitable for employment in the task of computer-aided summarisation.

Chapter 8 revisits the aims of the thesis, and discusses the extent to which they have been achieved. It also presents the main conclusions which can be drawn from the investigation carried out in the previous chapters, and indicates directions for future work.

Chapter 2. Summaries and Human Summarisation

2.1 Overview

This chapter presents a discussion of aspects to be considered during any kind of summarisation process and an account of summarisation from a human abstracting perspective. The use of human summarisation as a model for computer-aided summarisation is further developed in the remaining chapters of the thesis, and this chapter provides an introduction.

In order to properly introduce the general field of summarisation, a review of what needs to be considered during summarisation is necessary. First of all, *summaries* and *summarisation* are briefly discussed, to give the reader some idea of the basic notions dealt with in this thesis. These basic notions are covered in Section 2.2. *Context factors* (Sparck Jones 1999; Tucker 1999; Orasan 2006) for summarisation are then detailed, as a means of identifying the many different aspects of texts, readers and authors which can affect both the human and automatic summarisation processes. This provides a basis to work from when one starts to think about summarising a text, as these factors can be used to describe the parameters that need to be considered when creating summaries of any type and the implications they have for the process of summarisation. Context factors are presented in Section 2.3.

Human summarisation is discussed next, looking at the ways in which humans summarise texts (Pinto Molina 1995; Cremmins 1996; Endres-Niggemeyer 1998). This is necessary as the main aim of this research is to establish means of making summaries more coherent and readable by applying human-style transformations to them. In this chapter, human summarisation is considered as a general model for automatic and computer-aided summarisation, explored in Chapter 3, for two reasons. Firstly, it is the starting point for the automatic processes and secondly, the longer-term aim of most automatic summarisation systems is to produce summaries of a comparable quality to those produced by humans. A more detailed account of using human summarisation as a model for computer-aided summarisation is presented in Chapter 3 and Chapter 4. Human summarisation is discussed in Section 2.4. The chapter finishes with conclusions.

2.2 Summaries and summarisation

Whilst a reasonable first step in a thesis about summarisation might be to define the terms *summary* and *summarisation*, this task is not simple. There are numerous definitions of *summary*, depending on the field in which it is discussed, and even within the same field. This short section suggests what a summary can be and what summarisation can entail. Researchers concerned with *human* summarisation (often termed *abstracting*), i.e., summaries (abstracts)[2] produced by people, tend to define summary rather differently to researchers concerned with *automatic* summarisation, i.e., summaries produced by machines. But we need to know what a summary is

[2] The distinction between *abstracts* and another type of summary, *extracts*, is discussed later in this section and in Section 3.2.

before we can define any kind of summarisation as the process of creating one. Therefore several different definitions from the fields of both human and automatic summarisation are given, allowing the reader an insight into what a summary might be.[3]

As this thesis advocates the idea of human summarisation as a model for computer-aided summarisation, the first definitions of *summary* are taken from human summarisation. A much-used definition within the literature is that of the American National Standards Institute (ANSI), which states that a summary (or abstract) is "an abbreviated, accurate representation of the contents of a document, preferably prepared by its author(s) for publication with it" (American National Standards Institute 1997: i). For ANSI, an abstract should also be highly-structured, concise and coherent. A slightly different view of an abstract or summary is given by Cleveland (1983: 104): "an abstract summarises the essential contents of a particular knowledge record, and it is a true surrogate of the document". A third definition from the field of human summarisation comes from Endres-Niggemeyer (1998: 1), who considers a summary to be "the reduction of mostly textual information to its most essential points", pointing out that people summarise representations not language.

In the field of automatic summarisation, Mani and Maybury (1999: ix) define *summarisation* as "*the process of distilling the most important information from a source (or sources) to produce an abridged version for a particular user (or users) and task (or tasks)*". Hovy (2003: 584) suggests the following definition for a

[3] It should be noted that in the summarisation literature summaries and summarisation are often defined in terms of each other, so that the definitions seem to overlap somewhat. In some cases, only a definition of summary *or* a definition of summarisation is given, but both concepts are mentioned.

summary: "a text produced from one or more texts, that contains a significant portion of the information in the original text(s) and is not longer than half of the original text(s)". Finally, Sparck Jones (1999: 1) defines a summary as: "*a reductive transformation of source text to summary text through content reduction by selection and/or generalization on what is important in the source*", which she states is a general and obvious definition, but one which highlights the difficulty of the task of summarisation, especially automatic summarisation. This definition is particularly appropriate to the work undertaken in this thesis, as it splits the steps of selection and other 'operations' which transform one text into another.

Despite their differences, the definitions given above all point to some kind of reduced representation of the information in a given source (regardless of the number of texts constituting the source). However, human and automatic summaries are very different. The main difference highlighted here is the distinction between *abstracts* and *extracts*. Very generally, humans produce *abstracts* and computers produce *extracts*.[4] An *abstract* can be described as a summary comprising concepts/ideas taken from the source which are then 'reinterpreted' and presented in a different form, whilst an *extract* is a summary consisting of units of text taken from the source and presented verbatim. This is an important distinction within this thesis, as later chapters examine how a human summariser can transform an extract into an abstract using certain operations. Whilst humans nearly always produce abstracts, there have been attempts at creating both abstracts and extracts automatically. A review of the work in the field of automatic summarisation is presented in Chapter 3, where the

[4] It must be pointed out that this is a very general statement, and, as discussed in Chapter 3, it is possible to have both automatic extracts *and* automatic abstracts. However, there is a much larger body of work on automatic extracts than there is on automatic abstracts.

distinction between abstracts and extracts is also discussed more fully.[5] Human summarisation is described in Section 2.4 below. Functions, or uses, and characteristics of summaries are described with the help of context factors in Section 2.3. Throughout the thesis, the term *summary* is used as an umbrella term for *abstracts* and *extracts*, likewise, *summarisation* covers both *extracting* and *abstracting*. *Extract* and *abstract* are only used to refer to those specific types of summary.

2.3 Context factors

Context factors for summarisation were introduced by Sparck Jones (1999; 2001) and further developed by Tucker (1999) as a means of classifying the different parameters that affect the summarisation process and therefore the summaries produced by this process. All summaries, whether created by humans or by automatic means, are subject to these context factors. Whilst Sparck Jones and Tucker both advocate their use in guiding summarisation and evaluation, it is argued here that as well as their obvious importance for these tasks, they also provide a comprehensive way of *describing* summaries, summarisation, and what needs to be taken into account when considering these. Orasan (2006) uses context factors as a means of describing *characteristics* of summaries, adding a factor of his own. He also highlights their importance from a practical perspective, considering them at each stage of the automatic summarisation process. Context factors are therefore

[5] Further discussion of the *extract-abstract* distinction is kept until later in this chapter and Chapter 3, where it is discussed in relation to automatic summarisation.

seen as an appropriate means by which to start an introduction to the area of text summarisation.

Sparck Jones (1999) specifies three different types of context factor: *input factors*, *purpose factors* and *output factors*, which Tucker (1999) further develops. Tucker states that input factors concern the source text only, purpose factors the relationship between the source and summary text, and output factors the summary text only. This means that some of Sparck Jones' output factors become purpose factors in Tucker's classification as she does not define which texts the three types of factors relate to. The differences between the classifications of Sparck Jones, Tucker and Orasan highlight the fact that context factors, in particular purpose and output factors, are somewhat difficult to classify. The following discussion of context factors uses aspects from all three classifications deemed relevant for the discussion of context factors as a tool for describing summaries.

Input factors and output factors are text properties and are seen as task-independent, i.e., they can be applied to tasks other than summarisation. Purpose factors, on the other hand, are very much task-specific as the task and the methods used to achieve this task determine what the relation between the source (input) and summary (output) texts will be. Sparck Jones relates the three types of context factor to each other as "defining a summary *function*, which is: given Input Factor data, to satisfy Purpose Factor requirements, via Output Factor devices" (Sparck Jones 1999: 5).

2.3.1 Input factors

Input factors are concerned with the source text for summarisation. As the texts used in this research are texts written in English to be used in single-document summarisation, this is what source text is taken to mean. The input factors therefore will be assumed to be descriptive of these types of written text. Of course, there are many different kinds of other source texts, and indeed source material, that could be summarised, but discussion of these other kinds is beyond the scope of this thesis. Sparck Jones (1999) stipulates three different classes of input factors, each relating to the *source*: *form*, *subject type* and *unit*, each of which can be further defined by their own factors.

Form

This input factor is further sub-divided into four factors: *text structure*, *scale*, *medium* and *genre*. Tucker (1999) also adds another two factors to those given by Sparck Jones, *style* and *subject matter*.

Structure is concerned with the way a text is organised or arranged. It includes both the *explicit*, or *large-scale* structure, that is, whether the text is divided into elements such as chapters or sections, or contains headings and sub-headings, or is continuous running text. An example where these features are particularly important in summarisation is in the summarisation of research papers and scientific articles as it can be useful to extract information from certain sections of the text, such as the introduction and the conclusion. Another example, more pertinent to this particular research, is the practice of including the title and sub-headings when summarising

news texts, or annotating them for summarisation (see Chapter 4 for more on this type of guidelines for summarisation). Additionally, the *embedded* or *small-scale* structure which refers to things like the rhetorical structure of the text, or linguistic patterns which can be identified, can be considered. Certain rhetorical patterns have proved useful for summarising texts (Marcu 1997), as have linguistic patterns of repetition based on Hoey (1991)'s Lexical Cohesion (Benbrahim and Ahmad 1994; Benbrahim and Ahmad 1995; Barzilay and Elhadad 1997).

Scale refers to the length of the text to be summarised. This is an important factor to consider because it will affect the compression rate during summarisation. Longer and shorter documents will usually require very different compression rates, for example a 30% summary of a newswire text may be desirable whereas a 30% summary of a PhD thesis is not likely to be so.

Medium refers to the language used in the text, and whether it is a sublanguage; there can be different varieties of both of these. The task that the summarisation process or the summary itself is used for will determine whether the source text medium is reflected in the output. The *genre* of a text reflects its communicative function rather than its subject or content: whether it is designed to describe, instruct, criticise and so on will determine whether the text is descriptive, instructional, critical etc. Other authors, however, use the term *genre* in its perhaps more common sense of the *variety* of text (see Swales (1990) for a discussion of genre); Mani (2001) gives examples of source genre such as technical reports, news stories and email messages. Tucker states that his *style* factor is distinct from *genre* because it is concerned with the choice of linguistic constructions and vocabulary used in the text. For example,

reported speech is frequently found in newswire texts and newspaper articles. The *style* factor links to the corpus analysis of (extract, abstract) pairs explored in Chapter 5 and Chapter 6 of this thesis.

Tucker also points out that certain combinations of the above-mentioned form factors are particularly common; comparing the factors often present in novels (quite long, chapter divisions, narrative), scientific papers (shorter, named sections, descriptive, narrative, critical) and newspaper stories (even shorter, continuous narrative or descriptive text, distinctive style) as examples. However, it must be remembered that there are always exceptions and so the factors should be treated as independent of each other.

Subject type

Sparck Jones' *subject type* becomes Tucker's *intended readership*. Tucker argues that because this factor allows a categorisation of the text depending on whether it was intended for a general or specialist reader rather than concerning the subject matter of the source text, *intended readership* is a better name. Even within these two broad groups of reader, different general or specialist readers may still have different amounts of background knowledge or familiarity with the subject matter, which means that the summary or summarisation process must take this into account. In addition, it must be considered that in summarisation the intended readership of the source text is not necessarily the same as the intended readership of the output text, so, for example, a summary may need to contain fewer technical terms than its source depending on who the summary is aimed at. Sparck Jones splits subject type

into *ordinary*, *specialized,* or *restricted*, according to the level of background knowledge of the reader.

Subject matter

Tucker (1999: 13) states that "Since a summary must reflect at least some of the source text's subject matter (i.e., what it is about), this is in a sense the most important source factor." He keeps *subject matter* distinct from the *subject type*/*intended readership* factor as it refers to what the text is actually about rather than the level of terminology or the intended reader. Indeed, in theory, even if a summariser was unaware of any other context factor, for a summary to actually be a summary of a particular text, there needs to be some reproduction or representation of the source text within it (see definitions given in Section 2.2).

Unit

Although this thesis is concerned with single-document summarisation, it is worth mentioning the *unit* factor, as not all work on summarisation focuses on this. Within the field of automatic summarisation, research has been carried out into multi-document summarisation, where more than one text is considered as the source (see, for example, Radev and McKeown (1998), Goldstein et al. (2000), Barzilay (2003)). This factor relates to whether there are *single* or *multiple* sources, and Sparck Jones notes that if several documents are intended as some kind of collection for a specific purpose, then as they are no longer distinct entities this constitutes a single unit and the collection should be treated as such. In multi-document summarisation, this factor is important for issues such as redundancy and changes in information over

time. It is not covered by Tucker, as he is concerned only with single document summarisation.

2.3.2 Purpose factors

According to Sparck Jones (1999), purpose factors are the most important factors. Tucker (1999) further defines them as being concerned solely with the relationship between the source and the output texts, i.e., what needs to be considered in the process of creating a summary from a source text. The three original types of purpose factor given by Sparck Jones are *situation*, *audience* and *use*. In order to relate these factors to ones which directly describe requirements for the summary itself, Tucker also adds *summary type* and *coverage*. These two factors are similar to Sparck Jones' *style* and *material* output factors, and Tucker argues that because they are concerned more with the relationship between source and summary than the summary alone it is more appropriate to classify them as purpose factors. This is also the view taken by Orasan (2006), and the one taken here.

Situation, audience and use

Situation is the context in which a summary will be used. This can either be *tied*, where the context is known in advance thus allowing the summary to be tailored to it, or *floating*, where there is no specific context for a summary to be used in and therefore a generic summary needs to be produced. The *audience* factor provides information about who is going to read the summary, and this can be *targeted* or *untargeted* depending on the audience's domain knowledge and language skill. As

mentioned above (Section 2.3.1), the reader of the summary will not necessarily be the same as the reader of the source and this must be taken into account, as well as the fact that different audiences can need different summaries of the same text. *Situation* and *audience* appear to be very similar factors, and can be linked to another purpose factor, *coverage*, discussed later in this section. It seems impossible to separate them completely because the aspects of summarisation that they take into account are so similar. For example, *targeted* audience cannot be considered to be distinct from *tied* situation, and both of these correspond to *narrow scope* (see **Coverage** below). Instead of complicating matters by trying to classify factors into so fine-grained a distinction which is not always obvious, it may be more appropriate to consider these factors as the same thing and distinguish only between *generic* and *user-*/*topic-*/*query-focused* summaries, common terms in the automatic summarisation literature. These terms are discussed further under **Coverage**.

Use relates to what the summary will be used for, or its function. Just as a summary must be appropriate for its reader, it must also be appropriate for its intended use. A summary intended for one task can be completely inappropriate if it is used for a different task: consider how different a summary would need to be if it was intended to replace the source rather than to indicate whether the reader should read the source. Examples of general uses for summaries are *retrieving*: the summary is used to access or indicate an appropriate or interesting document, *substituting*: the summary is used in place of the full document, *previewing*: the summary is used to give an idea of the structure or content of a document to be read, *categorising*: the summary is used to assign a category to a text, and *refreshing*: the summary is used to remind the reader of something already read. Sparck Jones (2001) discusses

different possible summaries of the same source text as determined by different purpose factors, paying special attention to the impact these factors have on evaluation.

Summary type

This factor relates to the type of information a summary contains. Tucker describes four categories of summary according to the information present in them: *descriptive*, *evaluative*, *indicative* and *informative*. A *descriptive* summary can describe both the form and content of a source text. An *evaluative* summary, as its name suggests, offers some kind of critical response to the source, thereby evaluating it in some way. The *summary type* factor is similar to the *style* output factor indicated by Sparck Jones, who identifies three of the same types of summary as Tucker, *indicative*, *informative* and *critical*,[6] and an additional one, *aggregative*, in which varied or multiple sources are summarised in relation to each other.

The most frequently used distinction for summary type in automatic summarisation is *indicative-informative*, the distinction originally made by Borko and Bernier (1975), as these are the types of summary most often produced automatically. Indeed, descriptive and evaluative summaries are not always considered to be 'summaries' and are not usually considered within the field of automatic summarisation. This is because they do not always contain only information from the source, and evaluative summaries will contain opinions, which are not addressed in

[6] The terms *critical* and *evaluative* for summary types are used interchangeably here; different authors may prefer one term or the other, but they refer to the same type of information present in a summary. Mani (2001), who makes a three-way distinction between types of summary, uses the term *critical evaluative* for this type, in addition to the traditional *indicative* and *informative* labels.

automatic summarisation. An *indicative* summary is designed to allow the reader to judge whether or not the (full) source text is worth reading, and therefore presents only the main topics or points which appear in the source text. As it is only intended to achieve this purpose, an indicative summary cannot replace the source text itself because it does not provide a detailed enough level of information. An *informative* summary, on the other hand, should be able to be used as a surrogate of the original source text. This type of summary should include the important information present in the source to be conveyed to the reader. Of course, the very notion of trying to select what is important can be extremely difficult and is dependent on a number of other factors, such as audience, situation and use.

The American National Standards Institute (ANSI) (American National Standards Institute 1997) provides guidelines for abstractors as to what kinds of source texts indicative and informative summaries should be created for, suggesting that source texts with different structures may be better represented by different types of summary (see Section 4.2.1 for more details on the ANSI guidelines). Although several distinct types of summary have been described here, as Mani (2001: 9) points out, the types are not mutually exclusive: a summary can contain features of more than one summary type, and in particular, informative summaries can be seen as a subset of indicative ones, as they serve both indicative and informative functions.

Coverage

This factor, proposed by Tucker corresponds to Sparck Jones' *material* output factor. It classifies a summary in terms of its *coverage* of the source text, or how much of

the information from the source text is present in the summary. Tucker specifies two "dimensions" to this factor: *scope* and *depth*. *Scope* refers to the focus of the information taken from the source. It can be *narrow* or *broad*, depending on whether it focuses only on one aspect of the source or if it is more representative of the source text content as a whole, and can depend on the situation and audience factors as well as the source text itself. As mentioned above, this factor is very difficult to distinguish completely from *situation* and *audience*. These are similar to Sparck Jones' *partial* and *covering* distinctions within her *material* output factor. Elsewhere in the summarisation literature, these two aspects of Tucker's *scope* factor also tend to depend on the user, but are described as *user-* or *topic-* or *query-focused* and *generic*, respectively (see Mani (2001)). *Depth* refers to the amount of detail present in the summary, whether it provides a detailed account of the required information or just an outline. This dimension also relates to the *summary type* factor, as informative summaries are likely to be much more detailed than indicative ones, for example. However, there can be different degrees of depth within each different summary type, so the two factors should be kept distinct from each other.

Relation to source

Orasan (2006) adds one more purpose factor to those of Sparck Jones and Tucker, that of *relation to source*. He argues that although this factor does influence the way a summary is presented, it also has important implications regarding the choice of summarisation method in automatic systems. The view taken here is that in addition, or perhaps due, to these implications, the fact that whether a summary is an *extract* or an *abstract* falls into the category of *purpose* factors because it is concerned with the

relationship between the source text and the output. As explained briefly in Section 2.2, and elaborated in Section 3.2, an *extract* is created by presenting units from the source text without any modification, whereas an *abstract* contains some text which is not present in the same form in the source. The present research is concerned with the transformation of extracts into abstracts, therefore this distinction is an extremely important one and these terms are used frequently throughout the thesis.

2.3.3 Output factors

The final group of context factors is that of output factors. These are factors which are only concerned with the *output text*, or summary, and Tucker (1999: 14) describes them as "additional textual requirements for the summary itself, arising from the purpose factors" and points out that because they are text properties, as are input factors, they are practically the same as input factors (see Section 2.3.1). Some output factors may also depend on the source text and the purpose of the summary to a certain extent. In this section, Tucker's classification will be followed, and Sparck Jones' covered in relation to this. Tucker's output factors are split into two groups, *form* factors, which are further divided, as they are in input factors, and *subject matter*. Sparck Jones states that there are at least three main output factors: *material*, *format* and *style*. *Material* and *style* are mentioned briefly in Section 2.3.2, under Tucker's purpose factors of *summary type* and *coverage*.

Form

As discussed above (Section 2.3.1), the *form* factor (for output as well as input factors) is divided by Tucker into five factors of its own: *structure*, *scale*, *medium*,

genre and *style*. The *form* factor will not be described in much detail here as the relevant information regarding it has already been covered in Section 2.3.1.

Orasan (2006) states that the *structure* of a summary is usually regarded as independent of that of the source text, but that it may still be influenced by some of the conventions of its genre. Both Tucker and Sparck Jones (her version of structure is *format*) use continuous running text and text appearing under separate headings (for example, for a summary of a scientific paper) as possible examples of summary structures, and Tucker gives a further example of bullet points listing topics. The examples of headings in a summary of a scientific paper, and clearly marked problem/solution sections in some medical abstracts, also exemplify Orasan's observation concerning genre conventions.

Scale again refers to the length of the text, in this case, the length of the summary. This can be more or less dependent on the source, depending on whether the summary length is determined as a percentage of the source (the *compression* or *condensation rate*) or as a fixed number of words. There is no standard guideline for a summary's length (see Chapter 4), although it can be assumed that it is less than the length of the source text. Different authors present different views on the scale of the output text. For example, Hovy (2003: 584)'s definition of a summary states that the output should be "no longer than half of the original text(s)." However, this is a very general guideline regarding the length of the summary, and can obviously vary widely depending on the type and length of the source. ANSI (American National Standards Institute 1997) gives specific maximum lengths depending on the source document, ranging from a single page or 300 words for long documents to 30 words

for editorials and letters to the editor. Papers and articles have a recommended maximum length of 250 words. Borko and Bernier (1975) on the other hand, argue against imposing an arbitrary length limit as this can affect quality, but as a guide they do suggest that a summary should be approximately 10% of the length of the source.

The summary *medium* tends to be the same as that for the source text, although it can differ if it is necessary that the summary be transformed into or from a sublanguage, or translated into a different natural language (multilingual summarisation). Tucker does not say much about the genre of the summary, other than that it should be appropriate for its summary type. He also points out that for an informative summary, the genre may depend on the genre of the source. With regard to *style*, this factor is not only dependent on the style of the source text, but also on the *intended readership* or *audience* and *use* of the summary. As mentioned above (Section 2.3.1), this is not necessarily the same as that for the source, meaning that it can be more appropriate to use different linguistic constructions and vocabulary to those appearing in the source. As an example, consider a highly specialised biochemistry research paper whose summary is aimed at secondary school pupils. It is important to note here that style in Tucker's output factors is very different from style in Sparck Jones' output factors, where it is taken to indicate whether a summary is indicative, informative, critical or aggregative (see Section 2.3.2).

Subject Matter

This output factor depends on the *subject matter of the source text*, the *summary type* and the *coverage* of the summary. Tucker notes that the subject matter is not always only the same as the subject matter of the source. In an informative summary, it is the same because the summary presents the important information from the source. However, in other types of summary, such as indicative, descriptive and evaluative, the subject matter is primarily the source itself because these types of summary tell the reader something about the source text as a text. In these cases, the subject matter of the source is the secondary subject matter of the summary. He also points out that subject matter is a wider notion than content because even when the subject matter of the source and the summary are the same, the summary can still contain background information or world-knowledge which the source does not. In terms of coverage, if a summary has a *narrow scope*, it will only contain part of the subject matter of the source text as it will be focused on a particular aspect (or aspects) of the source. This factor also relates to the purpose factor *relation to source*, because, by definition, an extract will have the same subject matter as the source, whereas an abstract may not.

2.4 Human summarisation

The previous section discussed issues that need to be taken into account during summarisation and described summaries in terms of the factors involved in any summarisation process. This section looks at a sub-type of summarisation: human summarisation, because no kind of automatic summarisation can be considered without an overview of the way humans summarise texts. The long-term aim of automatic summarisation is to produce summaries which are of a comparable

standard to those produced by professional human summarisers, regardless of the methods employed to achieve this. However, the process of summarisation performed by humans has not enjoyed as extensive an investigation in recent years as it might have done in relation to its automatic counterpart. Endres-Niggemeyer (1998) provides the most recent and comprehensive study of human summarisation, which includes her observations of six expert abstractors during the summarisation process. Other authors have concentrated on studying aspects of abstracts such as the types of operations abstractors use in order to transform material from the source into suitable summary material, within the wider fields of both human summarisation (Cremmins 1996) and automatic summarisation (Jing and McKeown 1999; Chuah 2001a; Chuah 2001b). As well as being discussed in this chapter, the work of Cremmins (1996) is covered under the heading of guidelines for summarisers (Section 4.2.3). The operations observed by Jing and McKeown (1999) and Chuah (2001a; 2001b) are described in Section 5.2, along with the operations identified by Cremmins (1996) and Endres-Niggemeyer (1998). Operations identified as a result of a corpus analysis of extracts and abstracts for the present research are also discussed in detail in Chapter 5 and Chapter 6. Also within the field of automatic summarisation, the structure of scientific abstracts has been investigated (Salager-Meyer 1990; Liddy 1991; Orasan 2001). In addition, there is work which focuses on the more psychological aspects of the summarisation process in recalling information (Kintsch and van Dijk 1978; van Dijk 1979), although this body of work is beyond the scope of this thesis.

The discussion of human summarisation starts by distinguishing professional from non-professional summarisation, and focuses on the way professional abstractors

perform their tasks. As professionals, insights into their knowledge and how they apply it are more likely to better inform future developments in the field of automatic summarisation than observations of non-professionals. Professional summarisers and summarisation are discussed in Section 2.4.1. In Sections 2.4.2 and 2.4.3, alternative models of the general processes and strategies human summarisers use during summarisation is examined. Studies related to the structure of abstracts written by humans are described in Section 2.4.4. This review provides a background against which to develop the ideas in the remainder of the thesis concerning the transformation of text extracted from a source into a human-style abstract. It also provides something against which to compare the methods and strategies employed in automatic summarisation.

2.4.1 Professional vs non-professional summarisation

Human summarisers perform a number of different types of summarisation and related tasks such as the indexing and classification of texts. There are all kinds of different ways of presenting information that can be considered a summary. As well as more 'organised' summarisation such as film reviews and weather reports, people in everyday situations perform summarisation tasks, be it telling a friend the plot of a book they read, offering their opinion on a film they saw or texting someone the football results. These types of summarisation situation are what Endres-Niggemeyer (1998) terms *casual* or *everyday* summarisation and distinguishes from *professional* summarisation because the summary producer and user have very different expectations. There are also important aspects in the production of 'professional summaries' such as training, strategies and resources to take into account.

Professional summarisers or abstractors are people who are paid to create abstracts of texts, often working for an abstracting service. As with other professions, abstracting involves elements of training and guidance to which the layperson does not necessarily have access. Professionals can "summarize the same information with greater competence, speed, and quality than non-professionals" (Endres-Niggemeyer 1998: 98). Their *information environment* is very different from that of non-professionals, and incorporates explicit resources and strategies with which to summarise, including standards and guidelines, as well as experience gained from on-the-job training and technical support. They deal, for the main part, with technical and professional documents in a variety of media and in a professional situation. Professional summarisers also have to work quickly to make their enterprise cost-effective.

Professionals do not always produce abstracts only from scratch. Mani (2001) lists several activities that professional summarisers may typically undertake, such as editing author-supplied abstracts to conform to certain guidelines, tailoring abstracts for different audiences, writing abstracts in different languages, improving the fluency of abstracts and editing/revising machine-produced extracts to give coherent abstracts.[7] The field of professional abstracting also covers the related areas of indexing and classification, where similar strategies and methods are used to different ends. These areas of indexing and classification, however, do not fall within the remit of this thesis. Borko and Bernier (1975) and Rowley (1988) provide good

[7] Although this type of editing and revision would relate very well to the research in this thesis, unfortunately no further details are given by Mani (2001). However, this lack of detail does justify the focus of the present work.

overviews of professional abstracting and its related areas of indexing and classifying abstracts, editing and proofreading from a practical/organisation-oriented perspective.

Somewhere between *professional* abstracting and Endres-Niggemeyer's *everyday* or *casual* summarising are summaries of texts produced by their own authors. The most widespread example of this type of summary, especially in relation to automatic summarisation, is abstracts of research papers submitted to conferences, workshops, journals and the like. Abstracts written by authors of papers are recognisable summaries, even though the author is probably not by profession an abstractor, whereas a book review or a weather report or a news headline may not always be deemed such. Endres-Niggemeyer would class these as non-specialist summaries because they are not produced by professional summarisers. However, because this thesis is concerned with automatic and computer-aided summarisation, and this type of abstract is frequently used for research on aspects of automatic summarisation, it is important to distinguish them from other types of non-specialist summarisation. For ease of reference, these will be termed *author summaries*, being produced by *author summarisers* during *author summarisation*. Much of the research into human summarisation within the field of automatic summarisation is based on this type of abstract (for example, Salager-Meyer (1990), Liddy (1991), Orasan (2001), Chuah (2001a; 2001b)) as they are relatively easy to obtain for research purposes, and they usually appear with their source text, allowing a comparison to be made. These author summaries may seem at first glance to be a very good type of abstract: who better to inform others of the most important content of their paper in a condensed version than the author themselves? They know their own work, and the ideas and

results present in their paper; they know the point of their own paper. As Cremmins (1996) points out, author summaries, when properly written, can be of high technical quality as the author is likely to be a subject expert, as well as bringing cost savings to customers of abstracting services.

However, author summaries, particularly those associated with conferences, do have a number of flaws (Mani 2001). They can often be rushed, and due to space restrictions can be abruptly cut off, meaning that the information contained in them is not really representative of the full text. This can also give rise to an abstract which is not easy or nice to read. Intentionally or unintentionally, authors can present material in their abstract which is not an unbiased representation of the source; their own idea of focus or importance in the full text can differ from that of their audience. The discussion of context factors in Section 2.3 pointed out that the audience or user of a summary plays an important part in its construction, and if the expectations of the summary writer and summary reader differ, there can be misunderstandings leading to a summary which is not particularly useful, or is even misleading, to the reader. Self-promotion can also come into play. In addition, there is no uniformity across different journals or conferences with respect to the format or content of abstracts.

This is in contrast to the field of professional summarising, where summarisers have guidelines and references to ensure uniformity and the production of the best possible abstract. If summaries produced without guidelines can be so problematic, and professional abstractors use guidelines to ensure they produce high quality and accurate summaries, then it is only fair that human summarisers using a computer-aided summarisation system have access to such resources too. Chapter 4 discusses

the need for guidelines in summarisation and presents a set of guidelines formulated for the summarisation of news texts, and Section 6.6 presents a set of guidelines for the post-editing of automatic extracts produced by a computer-aided summarisation system. As the professional abstractor is not necessarily an expert in the field of whatever material they have to summarise, they have to obtain information from the material itself, rather than any biased background knowledge or other interest they may have. This is essentially then an act of *reconstructing* information content from the material available and should produce a more representative summary, and in this sense can be seen as similar to what automatic summarisation tries to achieve. A*uthor summarisation* is also different from other non-specialist, or everyday, summarisation because the author is concerned with creating a summary as a distinct task and is aware of this task.

Regardless of these differences between professional and non-professional summarisers, Endres-Niggemeyer (1998) points out that the *core subtask* of any kind of summarisation is the *condensation* of information. The difference between professional, everyday and author summarisers lies in their reasons for summarising, their awareness of the strategies and resources available to them and past training and experience: an everyday summariser may not even realise they are producing a summary, whereas a professional summariser is always aware of their professional role when performing the summarisation task. This is not to say that the underlying cognitive processes of assimilating information, selecting, transforming, reformulating and presenting executed by the professional and non-professional summariser during summarisation are different; both will typically perform the three main subtasks of analysis, condensation and presentation, but they will go about

them in different ways and will have a different level of awareness of strategies employed to reach their goal of a summary (if indeed they recognise their task as such). Chapter 4 gives an overview of some of the different types of guidelines and advice available to professional summarisers, developing a set for the human extraction of news texts used later in this thesis.

2.4.2 The summarisation process: stages and strategies

Having made the distinction between different kinds of summarisers and summarisation, the ways in which summarisers actually produce summaries is now examined. As noted above, professional, everyday and author summarisers perform the same core subtasks in summarisation, which is achieved via certain cognitive processes, although their levels of awareness and the resources available to them can differ greatly. Endres-Niggemeyer (1998) observed six expert (professional) abstractors[8] as they performed the task of summarisation on different texts. Each expert performed nine summarisation processes, summarising three short documents (conference papers or journal articles) and three long documents (reports or monographs), as well as completing three indexing and classifying tasks. The experts summarised the documents in their natural working environment and were allowed to choose their own documents. Insights are gained into the way they summarise by using thinking-aloud protocols to record every step of the process. From these protocols, three different stages in the professional summarisation process are

[8] Two of the expert summarisers were Harold Borko and Edward Cremmins, both of whom have published books in the field of abstracting.

identified: *document exploration*, *relevance assessment* and *summary production*, as well as different strategies employed to reach the goal of summarisation.

Endres-Niggemeyer established the use of common and individual strategies for summarising texts. Strategies are single activities which the experts use during the summarisation process to help them achieve their goal of creating a good summary. Such strategies range from underlining information and improving the style of the abstract to dealing with interruptions in the workplace and self-encouragement, and are classified into four broad categories, each of which is further broken down into a number of sub-categories and individual summariser strategies. Endres-Niggemeyer terms the collection of the 552 separate strategies used by the expert summarisers the *intellectual toolbox* and classifies the four broad categories as follows: Metacognition, Control of working processes, Basic intellectual activities and literacy, and Professional skills: abstracting, indexing, classifying and descriptive cataloging. The fourth category, professional skills, is of most relevance to the work in this thesis, as it is concerned with the presentation of the information in the abstract. It is further split into the sub-categories of information acquisition and information presentation, both of which are then further sub-divided. The sub-category of *information presentation* is of most interest here. Examples of separate information presentation strategies are *author*: Use the author's own words, *no-rep*: Avoid repetitions, *active*: Use active sentences where possible, *direct*: Be as direct as possible in your expression. The relevant aspects of this sub-category of strategies are discussed in more detail in Chapter 5 and Chapter 6 during the corpus analysis of human summary production operations.

In the whole *intellectual toolbox*, 83 strategies were shared by all the expert summarisers, 60 by five of the experts, 62 by four, 79 strategies by three, 101 by two of the abstractors and 167 strategies were found to be individual. Different summarisers used different numbers of strategies for different reasons, for example, one expert preferred speed over sophistication due to the nature of his normal working environment and therefore used fewer individual strategies than the others. Endres-Niggemeyer (1998: 130) draws the general conclusion that "all experts share common methods knowledge to a large degree, but that each one has, in addition, a small private stock of methods which serve purposes special to her or his tasks or reflect an individual working style". She also points out the fact that shared strategies were observed despite each summariser choosing their own documents and working in their natural context. This means that the observations are likely to hold for other professional summarisers. For a detailed account of these strategies, the reader is referred to the original text by Endres-Niggemeyer (1998). Each of the three stages of summarisation is further discussed below.

Document exploration

Document exploration is the first stage identified by Endres-Niggemeyer in the summarisation process. In this stage, the summariser explores the layout and organisation of the document to locate important information. The outline items of documents such as titles, headings and tables of contents, as well as introductions and conclusions, all contribute to the summariser's search. If the document is well-structured or well-organised, then this outline acts as a summary in itself, providing the abstractor with a representation of the document theme (i.e., what the document

is about, see **Relevance assessment** below) which they can expand with information later. Professional abstractors are well aware, through experience, of the typical structures of different document types. This means that they know where to look for the important information in specific types of document. When they begin to explore the document, this prior knowledge, or *scheme* is activated, and the abstractor then just needs to expand it with information relevant to the document under consideration at the present time. Endres-Niggemeyer terms these empty schemes *document surface representation* and *document scheme representation*, which lie dormant in the mind of the summariser awaiting activation by an exploration of a specific document type so that they can be filled with relevant information from a specific document to produce a summary. This document exploration and search for the location of important information is a time-saving device: there is no point in the abstractor reading the whole document if they can access the information they need by carrying out an initial exploration using these outline items.

In terms of the context factors discussed above (Section 2.3), abstractors involved in document exploration are showing an active awareness of the input factor *form* (see Section 2.3.1) in using the document structure as a starting point in summarisation. The other input factors also affect this stage of the summarisation process as the abstractor explores the source document, but *form* is the one being actively exploited and is therefore seen as characteristic of this first stage.

Relevance assessment

Endres-Niggemeyer identifies the second stage of the summarisation process as relevance assessment. During this stage, the summariser assesses information in the document to see if it is relevant to the summary. This involves recognising the *thematic structure* of the document, or the *document theme*. According to Endres-Niggemeyer (1998: 150), "a theme is an organized semantic structure that pervades the text and makes it coherent, i.e. the well-known macrostructure", in other words, what the document is *about*. She sees this stage as reflecting one of the classical main tasks of the summariser, that of *solving the aboutness problem*. A key factor in whether information is relevant is its relation to the core thematic structure of the document. The theme has a *root* (i.e., the main topic or point of the document) to which other information in the document is related. In order to accumulate relevant information for the summary, the summariser reconstructs the thematic structure of the document in their mind, identifies candidate statements in the document, checks these for relevance to the theme root and attaches them with semantic relations to the thematic structure. The summariser expands the theme in this way until there is sufficient material in the representation with which to produce the summary. When the theme is complete, it is a structured representation of the document's meaning. The summariser recognises the theme by looking at various aspects of the document, usually starting with the title. This activity is aided by the fact that authors usually indicate the core meaning of their document in some way, for example, through repetition and rephrasing, rhetorical emphasis, layout features such as italics, or privileged position. It is generally accepted that things occurring at the beginning of documents, sub-divisions within the document, and paragraphs are important/relevant. Professional summarisers are aware of this and therefore are able

to identify the core theme of the document quickly by examining certain aspects of the text, particularly information which appears at the beginning of sections.

Relevance assessment is strongly linked to the purpose factors described above (see Section 2.3.2) because the summariser needs to assess the relevance of the information in the source in terms of the task in hand. The summariser must be aware of the *situation*, *audience* and *use*, in order to pitch the summary at the appropriate level, as well as the *summary type* (i.e., indicative, informative etc.) because different types of summaries will consider different types of information relevant. *Coverage* is also an important factor here, as different information will be relevant depending on the focus of the summary.

Summary production

The third and final stage of summarising is summary production. This is where the actual creation of the summary as a discrete unit takes place and mainly involves *cutting and pasting* material from the original document using sentence patterns typical of the domain. Because summarisers are encouraged not to 'invent' anything for inclusion in their summaries to make them as representative of the source as possible, they follow the original document as closely as possible. Their text production style is therefore described as "copying relevant text items from the original document, and reorganizing them to fit into a new structure, often with the help of standard sentence patterns" (Endres-Niggemeyer 1998: 155). At this point in the summarisation process, the summariser exploits their theme knowledge and document structure knowledge to place relevant content items in a predefined

structure or mental representation activated during summarisation, that is, in the *summary representation*. The abstractor plans the text in order to present appropriate semantic content and a linear organisation which corresponds to the outline of the source document, and gives it a linguistic surface form. This stage involves the rewriting and editing of the summary before it reaches its final form; the summary representation must allow for this kind of revision. The standard transformation operations in the summary production stage include selection, copying, assessment, reformulation, and semantic reorganisation. Chapter 5 and Chapter 6 present a detailed account of similar operations observed in a corpus of (extract, abstract) pairs.

Context factors also affect the summary production stage. A summary as a text must adhere to certain summarisation conventions, depending on its use and audience. In this sense, both purpose factors (see Section 2.3.2) and output factors (Section 2.3.3) are important in this stage: output factors, particularly those belonging to the *form* factor, because they determine the surface form of the summary, and purpose factors because the output factors will vary according to the intended use, audience etc. of the summary. Although Orasan (2006)'s *relation to source* factor is more about automatic summarisation because it is concerned with whether the summary is an abstract or an extract, it is also relevant to the summary production stage of human summarisation. This is due to the fact that abstractors have to decide how much and which material to cut and paste from the source, and how much and which material to rephrase.

2.4.3 Other models of the human summarisation process

Although Endres-Niggemeyer's work is the most comprehensive account of the human summarisation process, it is not the only model. This section investigates alternatives proposed by two other researchers (Pinto Molina 1995; Cremmins 1996).

Cremmins (1996) posits an approximate four-stage model of abstracting, involving *focusing on basic features (content analysis)*, *identifying relevant information*, *extracting, organizing and reducing relevant information* and *information refinement*. This four-stage model is complemented by Cremmins' three-stage reading for abstracting model, which states three different types of reading abstractors should use. *Exploratory-to-Retrieval Reading* is the first stage, where the abstractor *explores* the document in order to identify information which could be *retrieved* for use in the abstract. The second stage of reading is *Responsive-to-Inventive Reading*, which involves the abstractor *responding* to the information identified, and *inventively* synthesising it into a draft abstract. The third stage of reading is *Connective* (*Value-to-meaning*) *Reading*, where the abstractor reviews the draft abstract, adding *value* to it and making more *meaning* of it. Cremmins' reading model is not discussed further here, as the focus of this thesis is on transformations and operations applied to text rather than the reading that precedes this. However, it is an important part of Cremmins' work and cannot be ignored.

The first of Cremmins' four stages, *focusing on basic features*, is where the abstractor determines the most important characteristics of the document by scanning it without performing in-depth reading. In relation to Endres-Niggemeyer (1998)'s stages, Cremmins' first stage corresponds to her first stage of document exploration.

Here, the abstractors identify the type of document first of all, as this will determine the representation of the document as well as how the abstractor will go about exploring, assessing and summarising it. Once the abstractor knows the document type and its representation, they can start to identify relevant information for the summary and continue with the rest of the summarisation process. This is linked to Endres-Niggemeyer's description of empty document representation schemes waiting to be triggered by a specific document type and filled with information relevant to the specific document to be summarised. The abstractor needs to identify the document type in order to activate their prior knowledge about where important information is in different types of documents. The abstractor then checks for sub-divisions such as section headings within the document which can indicate the location of important information. They also decide on issues such as the amount of generalising that will be needed to represent methods, discussions, conclusions or results, or how difficult certain information, such as conclusions, will be to separate out from other information in the text. According to Cremmins, at the end of this stage, by classifying the form and content of the document, the abstractor has determined the type of abstract to be written, its relative length and the degree of difficulty.

The second stage of abstracting in this model is *identifying relevant information* which can be seen as equivalent to Endres-Niggemeyer's second stage of relevance assessment. Cremmins notes that this stage can either happen simultaneously with stage one, especially by skilled abstractors, or individually as a separate stage. It involves reading rapidly to identify passages in the text which contain information that is potentially relevant to the summary. Conventional functional headings, for

example Methods and Findings, and cue words or phrases[9] such as *In this paper we* and *Results suggest* play an important role here, helping to signify the location of important information. The position of information within paragraphs is also exploited. Important sentences for abstracts tend to appear at the beginnings and ends of paragraphs, the first sentence often being the topic sentence, and the last often providing some kind of summary of the material in the paragraph. These means of finding important information for inclusion in the summary are exploited in automatic summarisation systems too. Section 3.4 describes these and other methods in the context of automatic summarisation in more detail. This stage results in the identification of a representative amount of relevant information to be extracted from the document.

Extracting, organizing and reducing[10] *relevant information* is Cremmins' third stage of abstracting. Having explored the document using its structure and identified the important or relevant information, the abstractor then extracts this information, holding it in a pre-established mental format corresponding to a standard format of an abstract for that particular document type. This is again similar to the mental representations, in this case *summary representation*, discussed by Endres-Niggemeyer (1998). The abstractor then checks this information for validity and relevance again, before mentally condensing and consolidating it and writing or typing the abstract according to this format. Both Cremmins and Endres-Niggemeyer point out that the reduction of the text is crucial. This stage results in the preparation of a concise and unified abstract, but one which is as yet unedited.

---

[9] These types of cue phrases are also termed *indicating phrases* by other researchers in the field of automatic summarisation (Paice 1981).
[10] Cremmins also uses the term *condensing* in places.

The fourth and final of Cremmins' stages of abstracting is *information refinement*, where the unedited summary produced in stage three is edited and reviewed, either by the abstractor or by editorial/technical reviewers, resulting in the completion of a 'good' final abstract. Cremmins' third and fourth stages are related to the third stage, summary production, identified by Endres-Niggemeyer (1998). It could be argued that Cremmins' stage three and Endres-Niggemeyer's stage two also overlap. However, the view held here is that stage two of both models correspond due to the result of the steps taken in this stage, i.e., suitable information is identified but not yet presented in a summary format, and the processes of exploration used to identify this information.

Pinto Molina (1995) describes the general abstracting process, advocating the development of a four-stage model for abstracting based on selection and interpretation. The four key steps of abstracting are viewed as a set of techniques as opposed to a theoretically predetermined set and depend strongly upon user needs and abstractor knowledge. The first step is *reading-understanding*, where the text is read and understood,[11] which counts as a first interpretation. In the second step, *selection*, only the relevant information is retained from the text. Here, the abstractor uses operations such as contraction and reduction to eliminate irrelevant material. *Interpretation* is the third step, where the abstractor performs a second interpretation of the text using reasoning and inference. This stage depends much more on the aim of the summary than the first interpretation does, and results in an *interpreted*

[11] Note that this is in contrast to other work (e.g. Endres-Niggemeyer (1998)), where it is claimed that abstractors do not need to fully understand a text in order to summarise it, instead exploiting various structural information and expectations about the document to be summarised.

meaning (i.e., the abstractor's interpretation rather than author's meaning or the textual meaning). The fourth step is *synthesis*, in which the abstract is synthesised, retaining the schematic structure of the document and taking into account the type of abstract which needs to be produced.

Pinto Molina (1995)'s model is not as close to Endres-Niggemeyer's and Cremmins' models as they are to each other, but there are some similarities which can be identified. Pinto Molina's step one involves reading and understanding the text itself, and therefore corresponds in this way to the first stage of Endres-Niggemeyer's and Cremmins' models, document exploration and focusing on basic features. The second step of selection is similar to the second step in the other models, resulting in suitable information being identified for inclusion in the summary. Pinto Molina's third step does not at first seem to fit into the other two models at all. Indeed, there is no obvious corresponding stage, but the idea of reinterpreting the selected information can be seen as being related to summary production and extracting, organising and reducing relevant information if thought of as an intermediate step between identifying relevant information and producing the actual summary; as a representation of relevant parts of the source in the abstractor's own terms. The final step in this model, *synthesis*, corresponds to the final stage of summary production in Endres-Niggemeyer's model, and to the third and fourth stages of Cremmins' model.

2.4.4 The structure of abstracts

The importance of document structure in the summarisation process has been discussed above. Research on the structure of abstracts, the results of which comply

with the view of the importance of structures, schemes and patterns in documents when attempting to process them, is described in this section. Liddy (1991) focuses on predictable information components which can be ordered to form the basic structure of a scientific abstract. Salager-Meyer (1990) considers the lexical realisations and patterns of different moves[12] in medical abstracts. In a similar study, Orasan (2001) looks at lexical patterns in moves in scientific abstracts. As these last two papers are concerned more with lexical realisations within abstracts than an exploration or formulation of their actual structure, they are only briefly described here.

Liddy (1991) examined the role of predictable schema-like structures in abstracts of empirical scientific research. Like Endres-Niggemeyer (1998), she assumed that abstractors use genre-specific schemes to explore and exploit the structure of a text, in this case, scientific research articles (see Endres-Niggemeyer's *document surface* and *document scheme* representations in Section 2.4.2 above). Liddy conducted a three-phase study using 12 professional abstractors from two different organisations in order to ascertain the extent of these predictable structures, both in the minds of professional abstractors and in abstracts themselves. The abstractors were given four tasks to complete in phase one, which provided her with evidence of the scheme in the minds of the abstractors. These tasks involved listing components of information they believed to be present in empirical scientific abstracts, ranking them according to their typicality in abstracts, possible organisations of the components and describing the meaningful relations they thought existed between them. The

[12] A *move* is a part of a text which has a specific purpose, for example, to summarise previous research in the introduction section of a research article (see Swales (1990) for more on moves in research articles).

abstractors from the first organisation gave 34 components, with 25 coming from the second, 15 of which were shared between both abstracting organisations. In phase two, she linguistically examined 276 abstracts to confirm the components given in phase one and found that not all of them occurred frequently, as well as finding some components which did not appear in the list given by the abstractors. Phase three involved validating this basic structure by having four abstractors examine 68 abstracts. Validation was at the level of 86%.

After the first two phases, Liddy was left with 15 of the most frequently occurring components in both the abstractors' views and the scientific abstracts themselves. These components were generally divided into three groups, each with a different level of typicality. The first group is *prototypical*, the basis of the structure of an empirical abstract. This group contains the information components subjects, results, purpose, conclusions, methodology, references and hypothesis. The second group was *typical*, encompassing the essential characteristics of empirical abstracts and included the components procedures, relation to other research, implications, conditions, data collection, research topic, discussion and sample selection technique. The third group contained other aspects of scientific abstracts which *elaborate* the basic structure of the 15 components listed above. These groups were ordered according to the linguistic analysis of the abstracts and the abstractors' opinions, to describe the structure of scientific empirical abstracts. The basic structure, including prototypical and typical components, is as follows: relation to other research, purpose, methodology, results, conclusions, references.

In another study related to the structure of summaries, Salager-Meyer (1990) investigated the linguistic realisation of the different discoursal moves in medical abstracts. A corpus of 77 abstracts from medical papers, reviews and reports was used. Four compulsory and two optional moves were considered to be present in a well-structured abstract, and the paper investigates the standard lexical formulae which signal these moves and the boundaries between them. This relates to the way abstractors skim through a document to locate important information using the document structure and cue phrases. Salager-Meyer considers the following moves: statement of the problem (or background information), purpose, methods (or procedure), results (or findings), conclusions, suggestions/recommendations. There are also two additional moves of data synthesis and case presentation, which were specific to certain types of document. Salager-Meyer focuses on the tenses, hedging devices and connectors used in particular moves, as well as some standard patterns such as *This paper reports*.

Orasan (2001) explores the lexical, syntactic and discourse level characteristics of scientific abstracts of journal papers and papers published in conference proceedings. N-grams and word frequency lists are examined for patterns in the abstracts and it was found that the type of abstract (journal or conference) does not influence the list of n-grams to a great extent. The word *paper* is used as an example to highlight the discovery of patterns in the corpus. At the discourse level, Orasan considers five moves which should be present in an abstract: Introduction, Problem, Solution, Evaluation, Conclusion, but notes that an abstract with only three of these moves, Problem, Solution and Evaluation, is also acceptable. However, only 58% of the texts in his corpus of 67 abstracts demonstrated this structure. Patterns similar to

Paice (1981)'s *indicating phrases* which could be useful to abstractors skimming the document in first and second stages of summarisation identified by Endres-Niggemeyer (1998) and Cremmins (1996) were found in the Problem, Method and Conclusion sections. Similar to Salager-Meyer, Orasan presents findings about how verbs and tenses typically convey information in the sections of a scientific abstract.

2.5 Conclusions

The first purpose of this chapter was to introduce the wider field of summarisation as a whole by examining the parameters involved in the process of creating any kind of summary. The context factors discussed in Section 2.3 proved an efficient way of doing this. First of all however, some basics of summaries and summarisation were briefly presented (Section 2.2) in order to suggest what a summary might be and what the process of summarisation may entail.

The second purpose of the chapter was to present a review of relevant work in the field of human summarisation. A distinction was made between professional and non-professional summarisation (Section 2.4.1) as this offers the reader a background to the guidelines for human summarisers discussed in Chapter 4. Stages and general strategies used by professional summarisers were described in Sections 2.4.2 and 2.4.3, which provides a useful framework for considering automatic and computer-aided summarisation (see Chapter 3). Finally, work on the structure of abstracts was discussed (Section 2.4.4), as this is an example of research by the automatic summarisation community on human summarisation and ties in with Endres-Niggemeyer (1998)'s work on the stages of summarisation as well as the

guidelines for summarisers discussed in Chapter 4. The next chapter discusses the related but separate field of automatic summarisation.

Chapter 3. Automatic Summarisation

3.1 Overview

The previous chapter provided an insight into the way humans summarise texts, thus paving the way for a discussion of *automatic summarisation*, which attempts to produce a similar result using computers instead of people. This chapter deals with automatic summarisation, considering the types of summaries that can be produced and some of the problems associated with them. Section 3.2 presents some basic notions within automatic summarisation as an introduction to the field and to emphasise some of the differences between human and automatic summarisation. This chapter does not attempt to give an exhaustive overview of all automatic summarisation systems developed to date, or of all the methods that these systems use. Instead, it focuses on major trends in automatic summarisation and the quality of summaries that are produced automatically in terms of coherence and readability. The review is split into two main parts: automatic abstracting[13] efforts and automatic extracting efforts. The positive and negative aspects of both types of summarisation are highlighted, providing a rationale for the original research in this thesis, which attempts to improve extracts by applying human-style transformations to them thereby creating abstracts.

---

[13] The term automatic abstracting is sometimes used in the literature, although not in this thesis, to refer to the automatic production of extracts as well as abstracts. As mentioned in Chapter 2, automatic summarisation is used here as an umbrella term for both abstracts and extracts which are produced automatically, allowing the distinction between an extract and an abstract to be maintained.

Section 3.3 examines different attempts at creating abstracts automatically, looking at the use of structured templates (for example, DeJong (1982)), which are filled with relevant information and the various uses of concepts (e.g. Paice and Jones (1993), Hahn and Reimer (1999), Hovy and Lin (1999)). Section 3.4 deals with the methods most commonly used to produce extracts which were first developed in the 1950s and 1960s (Baxendale 1958; Luhn 1958; Edmundson 1969) and have since been adapted and revised (e.g. Kupiec, Pederson and Chen (1995), Hovy and Lin (1999), Teufel and Moens (1999)). Other methods developed more recently, such as discourse-based methods (for example, Marcu (1997), Barzilay and Elhadad (1997)) and some surface rejection rules (e.g. Mitkov (1995)) are also briefly discussed. Common problems of automatic extracts related to coherence and readability are discussed in Section 3.4.7, along with the various solutions proposed for them to date.

Finally, this chapter covers *computer-aided summarisation* (*CAS*) (Orasan, Mitkov and Hasler 2003) as a trade-off between the notions of human and automatic summarisation already discussed. It is argued that CAS provides a feasible way forward for the further development of automatic summarisation given the current limitations of the field. This notion of *computer-aided* as opposed to *fully automatic* summarisation is important because it provides the framework for the research in this thesis and enables the reader to understand more fully the parameters within which the work was undertaken. Computer-aided summarisation, along with a discussion of other computer-assisted NLP tasks, is presented in Section 3.5. The chapter finishes with conclusions.

3.2 Basic notions in automatic summarisation

As mentioned in Chapter 2 (Sections 2.2 and 2.3.2), a basic distinction is often made in the field of automatic summarisation between *abstracts* and *extracts*. These two types of summary are discussed separately in this chapter, to enable the distinction to be made clearer, as automatically produced summaries are often classed as one or the other. Making this distinction now is important because it facilitates the presentation of the remaining work, which deals with how human summarisers transform an extract into a more coherent and readable abstract. The quote from Hovy (below) embodies exactly this aspect of summarisation as it is developed in this thesis. Problems with automatic abstracts and automatic extracts are discussed later in the chapter, in Sections 3.3 and 3.4, respectively.

> "**Extracts** are summaries created by reusing portions (words, sentences etc.) of the input text verbatim, while **abstracts** are created by re-generating the extracted content." (Hovy 2003: 584)

> "An extract is a summary consisting entirely of material copied from the input… an abstract is a summary at least some of whose material is not present in the input." (Mani 2001: 6)

To elaborate Hovy's and Mani's definitions of abstracts and extracts, abstracting is a type of summarisation where the important information in a document is identified and often abstracted to a more general level, and then reformulated into a summary that contains the same essential information but has a different linguistic realisation to the original. An abstract is likely to be more coherent and readable than an extract

of the same document because it is not restricted to the linguistic realisation of the source which is often presented out of context in an extract. This means that abstracts are less likely to contain phenomena such as dangling anaphors[14] or repetitive sentences. However, automatic abstracting systems have the disadvantage of being domain-specific due to the need for a large amount of knowledge to enable the system to abstract away from the original and reformulate the text in an attempt to mimic the result of the human summarisation process (see Section 2.4 for more on human summarisation). Cremmins (1996: 18) highlights the nature of abstracting well, although he is describing the human summarisation process:

> "When applied properly, the advanced techniques that are available to humans allow them not only to identify representative *sentences* but also to identify representative *information* in whatever form or location it appears in the materials to be abstracted and to format this information logically, reduce it coherently, and refine it concisely."

Extracting, on the other hand, is a type of summarisation where the important information in a document is identified, taken from the source and presented to the user or reader verbatim, in its surrounding sentence or other textual unit. The separate units selected are often presented as an extract in the same order that they appear in the source. As discussed below, this poses problems regarding the coherence and readability of this type of summary. Extracting systems have the advantage, however, of being able to summarise texts from different domains, with little or no additional tuning.

---

[14] A *dangling anaphor* is an anaphor whose antecedent is not included in a text.

As with human summarisation (Pinto Molina 1995; Cremmins 1996; Endres-Niggemeyer 1998) there have been related versions of a number of potentially overlapping steps or stages identified in the automatic summarisation process. Mani (2001) describes the three steps of *analysis*, *transformation* and *synthesis*, and Hovy (2003) *topic identification*, *interpretation* and *summary generation* stages. These two versions of the three stages are very similar, with each stage resulting in the same kind of summary and embodying the processes required to achieve this. Both agree that not all three stages appear in today's summarisation systems, citing the case of extraction, which can often be viewed as only utilising the first step of analysis or topic identification. If some form of *surface smoothing*, or improvement of the extracted sentences to make a more coherent discourse, is attempted, then extracting can also include the third stage of generation or synthesis. Abstracting necessarily employs at least the second stage in addition to the first (otherwise the result would still be an extract), and usually the third as well, in order to create an end result of a readable summary. These three steps are discussed below in relation to existing research in automatic summarisation. Sparck Jones (1999) also assumes a three-stage basic process model, although hers is slightly different from those of Hovy and Mani, being more concerned with the text and its representations than with the outcome of stages in terms of extracts and abstracts. Sparck Jones' three stages are: source text *interpretation* to source text representation, source representation *transformation* to summary text representation, and summary text *generation* from summary representation.

In relation to the context factors discussed in Section 2.3, the most relevant factors in terms of this automatic summarisation section are the purpose factors *relation to*

source, already discussed in Section 2.3.2, and *summary type*. Automatic summarisation systems usually produce summaries which can be classed as falling between indicative and informative summaries. Automatically produced summaries cannot currently be used as a surrogate for the full source document as methods are still not sufficiently reliable, and so are not informative abstracts. But neither do systems produce a truly indicative summary as their output is often an extract filling a specified compression rate or length with text presented to the user without any modification. They can therefore include more information than just a description of the main topics of the document. Having said this, most automatically produced summaries are termed *indicative* in the literature, as one of their main uses is to indicate to the user whether the full source text is worth reading.

3.3 Automatic abstracting

Human summarisation has so far been considered a necessary precursor to, and consequently a model for, automatic summarisation, and therefore it makes sense to deal first with automatic abstracts. However, it should be noted that in the automatic summarisation literature, extracts are usually discussed before abstracts as they are considered to require 'simpler' methods of production and are the most common kind of summary. In fact, the majority of work in automatic summarisation has focused on extraction rather than abstraction, and extracts were the first kind of summary to be produced automatically.

The second stage of the automatic summarisation process, *interpretation* or *transformation*, is the stage which distinguishes abstracting from extracting. As

Hovy (2003: 588) states, “During interpretation, the topics identified as important are fused, represented in new terms, and expressed in a new formulation, using words or concepts not found in the original text”. The major difference between abstracting and extracting is that in abstracting, some degree of inference can be made and background concepts which do not appear in the source document can be referenced. This is the reason that automatic abstracting is considered closer to human summarisation than automatic extracting is. At a practical level, it allows a much better use of compression as the abstract is not as restricted by the realisation of the information in the source. This is particularly useful for applications involving multi-document summarisation due to the potential volume of the source and problems of redundancy. Summarising different documents about the same topic by different authors means that similar information can be represented in a variety of ways.

Abstracting attempts to create summaries by mimicking humans in the sense that it does not rely on just the information presented in the source but can also bring other knowledge into play. It requires domain or world knowledge, and tends to work with *concepts*, therefore requiring some form of knowledge base to exploit. The amount of knowledge needed makes it currently impractical to apply automatic abstracting on a large scale and so only small-scale domain-specific applications have been developed to date. The major drawbacks of automatic abstracting are its domain-specificity and labour-intensive nature: “Interpretation remains blocked by the problem of domain knowledge acquisition. Before summarization systems can produce abstracts, this problem will have to be solved” (Hovy 2003: 589). There are a number of different ways to abstract a document, but Mani (2001: 129) notes that there are three general steps in these abstracting methods:

1. Build a semantic representation for sentences in a text.
2. Carry out selection, aggregation, and generalization operations on the semantic representations to create new ones. In the course of doing so, a discourse-level representation for the document may be leveraged. A knowledge base containing background concepts may also be used.
3. Render the new representations in natural language.

3.3.1 Existing abstracting techniques

One of the best-known examples of an automatic abstracting system uses structures such as templates which have slots that are filled with predefined types of information found in the source document. This copies human summarisation in that it uses a document-level representation (Endres-Niggemeyer (1998)'s document surface and scheme representations) and then fills this with salient information (document theme). DeJong's FRUMP system (DeJong 1982) has approximately 60 *sketchy scripts* which contain possible events during a particular situation and are instantiated by various explicit or implicit references in the source. FRUMP skims the source text attempting to match actors and objects involved in the script events. As an example, here is the script for a demonstration:

1. The demonstrators arrive at the demonstration location.
2. The demonstrators march.
3. Police arrive on the scene.
4. The demonstrators communicate with the target of the demonstration.
5. The demonstrators attack the target of the demonstration.
6. The demonstrators attack the police.
7. The police attack the demonstrators.

8. The police arrest the demonstrators.

(DeJong 1982: 150-151)

A major disadvantage is that if there is no slot concerning a particular event, however important it is in the source, it will not be considered and not be included in the abstract. There may be cases, even within the same domain, where a source document contains situations which are not covered in the scripts: only texts which contain situations described in the scripts can be summarised.

Paice and Jones (1993) match context patterns (which define stylistic and conceptual roles associated with the patterns) with salient information in a source text in the area of crop husbandry/agriculture. Patterns found in the source instantiate high-level concepts, such as *PEST* or *SPECIES*, the slots for which are then filled using *filler strings* from the source that name the specific conceptual role. The patterns are weighted because concepts can be referenced in more than one pattern, enabling the system to find the most appropriate concept for that particular occurrence. Output templates are used to generate the summary, the information being inserted into empty slots. Mani (2001) points out that although the abstract may seem effective because it is superficially well-formed, it is not always possible for the reader to judge its fidelity to the source. The work is viewed as falling between extracting and abstracting.

Rau, Jacobs and Zernik (1989) developed the SCISOR system which uses expectations about role-fillers, events and world-knowledge as well as a substantial grammar and lexicon to produce summaries in the domain of corporate mergers and acquisitions. The summaries are presented using a separate generator, which

produces predefined generalised sentences. The authors advocate the toleration of gaps in linguistic knowledge and the automatic acquisition of lexical information from their source texts as a way to combat the lack of extensive linguistic coverage which is so detrimental to automatic abstracting systems. Because of its manner of presentation and conceptual role-fillers, this system can be seen as similar to the work of Paice and Jones (1993) described above. However, SCISOR produces *user-focused* summaries in the form of answers to user queries.

SUSY (Fum, Guida and Tasso 1982) creates abstracts tailored to the user's needs in the domain of technical articles on computer operating systems. The system uses a number of selection rules and is based on semantics as well as on the psycholinguistic research of Kintsch (1974) and Kintsch and van Dijk (1978) in order to simulate human abstraction. Similar to human summarisation, a text schema is used to help the system identify only the most relevant information, and a summary schema is used to enable the user to specify their requirements for the abstract's organisation. This is also linked to the use of a combination of text structure and text meaning that the authors advocate. The extracted information is presented as a summary via rules for 'correct' sentence construction.

Hahn and Reimer (1999) describe operators that work to abstract and condense knowledge representation structures of a text based on activity and connectivity patterns between concepts in a knowledge base using the TOPIC system (Reimer and Hahn 1988). The idea behind this work is that of the salience of an object in discourse; the more salient the object, the more reason it has to be included in the summary. Salience is determined by a threshold and is based on the number of

occurrences of a concept (which can later be generalised) as well as inferences from the source. However, textual summaries cannot yet be produced using these operators, and the user is presented with only the conceptual representation. As with other abstracting systems, this is domain-dependent and new knowledge bases need to be developed for any new domains to be abstracted.

Hovy and Lin (1999)'s paper on SUMMARIST describes using an existing resource, WordNet (Fellbaum 1998), to carry out generalisations instead of creating their own domain-specific knowledge base of concepts as other researchers have done. They posit the notion of *concept fusion* or *topic interpretation*, where two or more extracted topics are merged into one or more unifying concepts, helping to reduce the text into an abstract. This is achieved by counting concepts and generalising them using WordNet, which results in *topic signatures* – sets of keywords and relative strengths of association, each related to a single headword. Summary generation is discussed, with the intention of employing full sentence planning and generation using an existing sentence planner and generator, but had not been implemented at the time of publication, when SUMMARIST was still producing extracts.

3.3.2 Synthesis for automatic abstracting

The third and final stage of automatic summarisation is *synthesis* or *generation*. Mani (2001) discusses a range of synthesis methods for abstracting, the first being *pretty printing*, where the text from the template or other representation is made more user-friendly. *Graphical output*, for example, as used in Hahn and Reimer (1999), provides the user with underlying summary representations such as graphs. The

problem here is that such representations can be difficult to understand and may therefore not be as useful as a textual summary – a possible solution is to use graphical output in combination with text. *Extraction* involves extracting the source segment for a semantic or logical representation given in a summary, but this can lead to a lack of coherence as in automatic extracts themselves.

Generation for synthesis is the most challenging option, but it can also give the best results as it offers an improvement in readability, clarity, and coherence via lexical choice and text planning strategies, as well as the option to perform further aggregation and generalisation operations. A discussion of natural language generation (NLG) is beyond the scope of this thesis, but see Reiter and Dale (2000) for an overview. Examples of summarisation systems using NLG are STREAK and PLANDOC (McKeown, Robin and Kukich 1995), which generate summaries of baseball games and telephone network planning activity, respectively.

3.4 Automatic extracting

The previous section highlighted the fact that whilst abstracts are more similar to human summaries, usually of a higher quality in terms of coherence and are not completely restricted by the linguistic realisation of the source text, there is a problem with the amount of world or domain knowledge necessary to enable abstracting systems to work efficiently. For this reason, automatic extracting is often performed instead as it allows a relatively 'easier' way to summarise. However, because extracts are very strictly dependent on the form of the source, they can sound unnatural or incoherent. Hovy (2003) claims that most systems today only embody

the first stage of summarisation, *topic identification* or *analysis*. This is in keeping with other reviews of the field which also state that the majority of existing systems produce extracts rather than abstracts, mainly due to the domain-specificity and consequent restrictions of automatic abstracting systems.

One of the first classic studies in automatic summarisation was carried out in the late 1960s, the methods employed still being used in some form today. Edmundson (1969) used a combination of *cue words*, *title words*, *key words* and *sentence location* to produce extracts. Most existing extraction systems use a combination of different modules based on those used by Edmundson, such as *position/location*, *cue* or *indicating*[15] *phrases*, *word* or *phrase frequency/key words* and *title* or *query overlap*, to create an extract. In addition, other types of information such as discourse-level information relating to cohesion and coherence, as well as various surface rejection rules, can be used. Each module gives a (positive or negative) weight to a specified textual unit and then the weights are added to give an overall score for each unit. The highest scoring unit is selected, and then the next, and so on, usually until a certain compression rate or word count is reached.

In automatic extracting, a sentence is the most common unit for which weights are added to produce an overall score. It is also possible to extract paragraphs (Salton et al. 1997), clauses or clause equivalents (Marcu 1997), or even phrases or words to produce very short summaries or to give brief overviews of a document (Boguraev and Kennedy 1999; Banko, Mittal and Witbrock 2000). It has been argued that extracting paragraphs results in more coherent and understandable summaries

[15] *Indicating phrases* are also sometimes called *indicator phrases*.

because phenomena such as dangling anaphors and discourse ruptures[16] are less likely. By extracting a whole paragraph, a sentence is not taken completely out of context and could therefore also convey more accurate information than a standalone 'important' sentence. However, this understanding and coherence comes at a price: as many automatic systems produce summaries in compliance with a certain compression rate, extracting whole paragraphs uses up this compression more quickly than extracting sentences alone. Extracting units at a sub-sentential level on the other hand, uses the compression rate more effectively, packing more information into less space, but can result in far more incoherent, or even ungrammatical, extracts due to the fact that parts of sentences have been taken from a source and presented together. Sentences are the most popular unit for extraction as they provide a balance between these two extremes. They are not as long as paragraphs and so do not fill the compression rate as quickly, and they are not as fine-grained as clauses or words and so do not pose as many problems regarding their combination and coherence. This is not to say that extracting sentences does not have its drawbacks. These drawbacks are discussed below, after a brief description of automatic extraction methods.

3.4.1 Position/location

This type of extraction module uses information regarding the regularities of text structure, based on the premise that text located in certain places in a document tends to contain important information. As discussed in Section 2.4, human summarisers also exploit text structure in this way when skimming documents during

[16] A *discourse rupture* occurs when discourse markers or connectives appear in a text without their proper context, for example, if a sentence starting with *Secondly* is extracted, but not the one starting with *Firstly*.

summarisation. In certain genres, titles, section headings or text following certain section headings, first paragraphs of documents and first and last sentences paragraphs contain more important information than text positioned in other places. This is still true in automatic summarisation, but researchers have tended to split titles and headings into a separate extraction module, combined with words overlapping with the user's query if applicable. These aspects of the text are dealt with below, under **Title or query overlap**. In a sentence extraction module exploiting position/location information, more weight is given to those sentences appearing in the first paragraph, under a certain heading, at the beginning or end of a paragraph, etc. (depending on the type of document being summarised) so that these are more likely to be extracted.

Many summarisation systems utilise this method in some form. Baxendale (1958) made the first attempt, finding that important sentences were located at the beginning or end of paragraphs. Edmundson (1969) based his location method on the hypothesis that firstly, sentences occurring under certain headings are *positively relevant*, and secondly, that topic sentences tend to appear either very early or very late in a document or a paragraph. A well-cited example of the position/location method is that of a lead summary, where simply the first *n* sentences of a text are extracted and presented to the user as a summary, outperforming other methods for news articles and newswire texts (e.g. Brandow, Mitze and Rau (1995), Alonso i Alemany and Fuentes Fort (2003), Orasan, Pekar and Hasler (2004)). Lin and Hovy (1997) and Hovy and Lin (1999) discuss a genre-dependent ranked list of sentence positions which provide the most important information for inclusion in summaries.

Amongst others, Teufel and Moens (1997) and Kupiec, Pederson and Chen (1995) also use position/location in combination with other methods.

3.4.2 Cue or indicating phrases

Edmundson (1969) identified *bonus* and *stigma* words in his summarisation experiments. The identification of these words was based on frequency and each group contained certain types of words. *Bonus words* were words with a frequency above a certain threshold which helped to indicate that a sentence is important, for example, *significant*. According to Edmundson, bonus words consisted of comparatives, superlatives, relative interrogatives, causality terms, value terms, and adverbs of conclusion. In contrast, *stigma words* had a frequency below a certain threshold and included anaphoric expressions, hedging expressions, belittling expressions, and insignificant detail expressions. Stigma words helped to show that a sentence was not important enough to be extracted during summarisation, for example, *impossible*, *hardly*. These groups of words have since been expanded to include phrases which explicitly signal importance, termed *indicating phrases* (Paice 1981), and include phrases such as *In this paper we show*, *In conclusion*, *This research is concerned with*. Bonus words and indicating or cue phrases are weighted positively and stigma words are weighted negatively in order to determine suitability for extraction. Teufel and Moens (1997) assign a judgment ranging from 'very likely' to 'very unlikely' to indicating phrases on the basis of the likelihood that the sentences containing them would be included in a summary. Pollock and Zamora (1975)'s ADAM system uses cue words as one of the main ways to determine

sentence selection or rejection. Indicating phrases are also used by, amongst others, Kupiec, Pederson and Chen (1995) and Hovy and Lin (1999).

3.4.3 Word or phrase frequency/key words

This type of method was originated by Luhn (1958) whose premise was that highly frequent words occurring in close proximity to other highly frequent words and separated from these other words by non-significant words indicate a sentence which is likely to be representative of the source. Luhn was concerned with content words and also with the company that these words keep, i.e., frequent words which appear in combination with certain other words are more likely to be important. Edmundson (1969) took content words above a given frequency threshold as *key words* which were given the weight of their document frequency and therefore more frequent words were more likely to be extracted. Stoplists are useful in this method, as the words included in them (determiners, conjunctions, prepositions, etc.) are often frequent but hold little or no content. There are various ways of obtaining the word frequencies for a document, and this method is widely used in automatic summarisation (e.g. Brandow, Mitze and Rau (1995), Kupiec, Pederson and Chen (1995), Teufel and Moens (1997), Hovy and Lin (1999)).

3.4.4 Title or query overlap

This method identifies words in a document which also appear in the title, headings or user's query where applicable. Sentences containing these words have more weight than those sentences which do not. This is based on the idea that an author will provide a relatively accurately descriptive title and headings, and that a user's

query will accurately reflect their interest. In summaries where a user's query is taken into account, a *user-focused* summary is produced rather than a generic one. This method has been utilised by Edmundson (1969), Teufel and Moens (1997) and Strzalkowski et al. (1999), among others.

3.4.5 Discourse-related methods

Instead of looking at individual aspects of texts, discourse-related methods exploit aspects of discourse as a whole and can often result in a more coherent extract although they are more labour-intensive than the methods already discussed. They can be broadly split into two categories: those based on coherence and those based on cohesion. Some of the best-known work using the underlying discourse structure of a text in terms of coherence for automatic summarisation is that by Marcu (1997; 1999). Based on Rhetorical Structure Theory (RST) (Mann and Thompson 1988), Marcu's work takes the notions of nucleus and satellite along with relations between spans of text in order to build rhetorical structure trees and produce extracts which both contain salient information and are coherent. Rhetorical structure information is also used by Ono, Sumita and Miike (1994) and Miike et al.(1994) for Japanese.

Although it seems appropriate due to its label as a theory of local coherence and salience, Centering Theory (CT) (Grosz, Joshi and Weinstein 1995), has rarely been used in automatic summarisation. Orasan (2003b; 2006) uses its *continuity principle*[17] (Kibble and Power 2000) to try to improve local coherence in extracts by taking into account other sentences rather than assessing each sentence in isolation.

[17] Centering Theory's *continuity principle* requires that two consecutive utterances have at least one entity in common.

Improvements in the informativeness and cohesion of summaries are reported; however, the quality of their discourse structure is low. Hasler (2004a) investigated the usefulness of Centering transitions for summarisation, but found that these alone were not enough to account for coherence in extracts. However, preliminary experiments using Centering Theory for the evaluation of extracts in the same paper proved more fruitful. Chapter 7 contains more information on CT in general, and on its development for use in the evaluation of summaries.

Text structure, or *document scheme* information is another aspect of discourse coherence which can be used in summarisation. This can be viewed as being related to the location method described above, as it takes advantage of the layout of the text at document level. Section 2.4.4 describes the work by Liddy (1991) on the role of predictable schema-like structures in abstracts of empirical scientific research. Teufel and Moens (1999) use similar document scheme information such as *background*, *solution/method*, *conclusion/claim*, for automatic extraction and the labelling of rhetorical roles of sentences in scientific texts. The labelled sentences are used to fill an argumentative template with slots based on the document scheme. Other uses of document schema information in automatic summarisation involve creating abstracts and are discussed above (Section 3.3).

Alonso i Alemany and Fuentes Fort (2003) use discourse markers in an attempt to improve the performance of a summariser based on lexical chains. Their work therefore integrates aspects of coherence and cohesion in the same summariser, adding rhetorical and argumentative structure information to lexical cohesion

information. However, their results show that only a slight improvement was achieved and their system does not outperform a lead summary baseline.

Cohesion-based extraction exploits aspects of a text such as lexical repetition (in its various forms), anaphora, coreference and semantic association, with the hypothesis that the more connected or linked a unit or entity in a text is, the more important it is. Hoey (1991)'s work on lexical chains which identifies the number of links and bonds between words has provided a basis for researchers in automatic summarisation. Benbrahim and Ahmad (1994; 1995) developed an automatic system based on Hoey's lexical cohesion using links and bonds between sentences, followed by Barzilay and Elhadad (1997) who take into account noun compounds as well as individual words. Other work has focused on cohesion graphs, where nodes are linked to others with which they share a semantic relation (Skorochod'ko 1971; Salton et al. 1997; Mani and Bloedorn 1999). Those nodes (often words) which are most 'bushy', or are most connected, i.e., share most links with other nodes, are considered to be most salient and the sentences (or other units) containing them are therefore better candidates for extraction. Boguraev and Kennedy (1999) use anaphora and coreference resolution to help create *capsule overviews* of documents.

3.4.6 Other methods used

In addition to those methods described above, there are a number of other ways of determining information which is suitable for inclusion in an extract. A number of researchers use various *surface rejection rules* which work by excluding text exhibiting certain features or located in certain places in a document. These rules can

also help with the problem of coherence and readability in extracts. One example is the exclusion of sentences containing fewer than a given number of words (Kupiec, Pederson and Chen 1995; Teufel and Moens 1999), as short sentences are thought to be less likely to contain important information. Mitkov, Le Roux and Descles (1994) propose a set of rejection rules for the sublanguage of elementary geometry which includes the elimination of text containing examples and text falling within brackets, subordinate clauses and quotation marks.

Guidelines given by Hasler, Orasan and Mitkov (2003) for the human annotation of a summarisation corpus contain similar instructions, although not all subordinate clauses should be ignored. They add that adjuncts specifying dates, times and places should not be annotated unless vital to the extract, along with tables, figures, examples, direct speech and reporting clauses. However, an analysis of the annotated corpus proved that not all text in direct speech or quotation marks was ignored by the annotators, showing that it can be important for a summary and therefore cannot always be ignored without losing vital information. There are similar observations regarding adjuncts specifying dates, times and places. Pollock and Zamora (1975) delete introductory and parenthetical clauses and phrases ending in *that* or beginning with *in* from the final version of their extracts as they claim that these can be removed without losing information.

3.4.7 Problems and possible solutions

Whilst the extraction methods described above can, when combined, produce a relatively acceptable summary in terms of information extracted, these extracts are

still a long way from being of as high a quality as abstracts produced by humans can be. In addition to the fact that an automatic system will not always select the same important information as a human abstractor would when summarising the same text, automatic extraction systems often produce incoherent or unnatural-sounding extracts. This is because the information is extracted verbatim from the source and then presented to the user usually without any modification. Even if simple changes are made to the extracted text to make it more coherent (see **Possible solutions: synthesis/generation for extracts** below), it can still read very differently to a human-written abstract of the same text.

Problems

Minel, Nugier and Piat (1997) describe two protocols for the assessment of the quality of automatic summaries, the FAN protocol focusing on the quality of summaries independently from their source and the information contained in them, and MLUCE on how useful summaries are when used to perform certain tasks. The FAN protocol incorporated four criteria against which to assess a summary: number of anaphora deprived of referents, rupture of textual segments organised by linear integration markers, presence of tautological sentences and legibility of the extract. A total of 27 automatic summaries were evaluated, comprising scientific and general press articles, book extracts and notes. The first two criteria highlighted the most problems: 52% of texts demonstrated at least one case of an anaphoric reference without its antecedent, and 63% of texts contained discourse ruptures. In terms of tautological sentences, only 8% of texts contained one instance, and for legibility, 26% of texts were classified as ‘mediocre’ or ‘very bad’, with the rest ‘good’ or

'very good'. These results suggest that dangling anaphors and discourse ruptures could be the most widespread problems in automatic extracts. Although the MLUCE protocol was primarily designed to assess the usefulness of summaries in given tasks, its results also contain some indication about the frequency of coherence- and readability-related problems with extracts. When asked to classify the extracts in terms of overall quality (based on being able to determine the field and the 'proven idea' of the summary, and on the logical linking of arguments), human judges assigned a label of *not very clear* to 10 extracts, *incomprehensible* to 5, *fairly clear* to 6 and *clear* to 6. More than half of the texts were classified as *not very clear* or *incomprehensible*, giving some idea of the level of quality of automatically produced extracts.

Similar to the results of Minel, Nugier and Piat (1997), Mani (2001) cites three main coherence problems with extracts due to their nature of taking sentences out of their context and presenting them as a summary. The first problem is that of dangling anaphors, where a pronoun or other referring expression is extracted but its antecedent is not. This makes the extract not only incoherent but also possibly incomprehensible or even misleading. For example, if the correct antecedent is not present in the extract, the reader may assign another 'antecedent' which *is* present to the referring expression. The second problem is gaps, which are explained as breaking the connection between ideas in a text by, for example, focusing on one thing and then switching abruptly to another topic. The last problem is structured environments such as itemised lists, bulleted text, tables and logical arguments. These cannot be split without causing problems, and including only parts of them will often mislead the reader. In terms of the texts investigated in this thesis, the only

potentially problematic case within the structured environments set is logical arguments as the corpus does not contain texts with lists, tables or bulleted text (see Section 4.5 for a description of the corpus).

Jing (2001) points out that extracts often contain *extraneous phrases*. Because longer sentences typically obtain higher weights during the selection process, these sentences are considered important. It is not always the case that all of the information contained in long sentences is relevant to the summary, and simple extraction does not remove these irrelevant phrases. She also highlights the fact that dangling logical and rhetorical connectives[18] can make sentences incoherent or incomprehensible and that out-of-context sentences placed together can convey unintended meanings to the reader. Similar to Jing's *extraneous phrases*, Orasan (2006) automatically determines extracts which are most similar to human produced abstracts and shows that they still contain unnecessary information. He argues that the only way to combat this problem is to produce an abstract rather than an extract.

Possible solutions: synthesis/generation for extracts

As in the human abstracting process, revision operations can be applied to automatically produced extracts in order to improve their coherence (and sometimes their informativeness). Revision occurs within the third stage of automatic summarisation, that of *generation* or *synthesis* and can consist of *shallow smoothing* or *full revision*, both of which are described below. Mani (2001) claims that these revision methods do not produce an abstract as they are only concerned with

[18] *Dangling logical and rhetorical connectives* are very similar to discourse ruptures, mentioned above.

'rearranging and smoothing' the source text. However, it has already been shown that an extract is often defined as text presented to the user from the source without any modification, and that human abstractors perform some of the same revision operations that Mani discusses whilst being instructed to stick as closely to the original wording as possible. Therefore the view held here is that revised extracts can be seen as abstracts, providing that some transformations or operations have been applied which somehow change the extracted text from its corresponding segments in the source.

Nanba and Okumura (2000) describe a number of problems and their solutions regarding shallow coherence smoothing. A dangling anaphor is either replaced by its antecedent, or deleted if the previous sentence (which may contain the antecedent) is not included in the summary. Dangling connectives should either be deleted or another conjunction added if the related sentence is within three sentences of the dangling conjunction. Complex sentences are dealt with by splitting the sentence into simpler sentences. Pronominalisation, omission and the addition of demonstratives are given as ways to combat redundant repetition, for example, of proper names. The problem of a lack of or extraneous adverbial particles is resolved by adding or deleting the particles in question.

Johnson et al. (1993) detail rules incorporated in their summarisation system to improve the cohesion of extracts by identifying anaphora using sentence selection and rejection rules.[19] This set of rules rejects all sentences containing anaphoric references, leaving the user with only non-anaphoric sentences which should also

[19] Johnson et al. (1993) also include a selection rule based on Paice (1981)'s *indicting phrases*.

introduce key concepts from the source. As well as rejection rules for different types of anaphoric references such as anaphoric quantifiers, subject pronouns and subject noun phrases before the main verb, the authors describe rejection rules for connectives, comparatives and demonstratives. Brandow, Mitze and Rau (1995) also exclude all sentences containing anaphors. Additionally, in an attempt to minimise problems with gaps they include in their extracts sentences which occur in between two extracted sentences, as well as the first sentence of a paragraph if the second or third sentence is selected. Orasan (2006) assesses the influence of pronouns on the automatic extraction of scientific articles. He argues that sentences containing pronouns should not be indiscriminately rejected during extraction, because the elimination of sentences containing certain pronouns leads to information loss. Paice (1981) proposes the *aggregation process* to combat the problem of dangling anaphors. When a sentence containing an anaphoric reference is extracted for a summary, sentence(s) which precede that sentence in the source are also extracted, as they are likely to contain the antecedent. Pollock and Zamora (1975) use rules similar to guidelines given by abstracting services regarding the standardisation of spelling, abbreviations and chemical compound names and formulae, which can help with readability and understanding (see Section 4.2 for more on guidelines for professional abstractors).

Whilst the methods used to deal with the typical problems in extracts are relatively simple, a number of them can fill the compression rate quickly due to the inclusion of material which would normally be excluded. One alternative to this is to extract only parts of sentences and put these together grammatically in a way similar to human abstractors. Jing and McKeown (1999) advocate a *cut and paste* approach to

extraction on the basis that this is often used by humans. As human abstractors do not necessarily cut and paste whole sentences, Jing and McKeown argue that a summariser does not need to do this either. Instead, they claim that only important fragments of source document sentences need to be identified and then woven together in such a way that they produce a grammatical sentence. Section 5.2.1 gives more information on the human operations identified to create grammatical summary sentences using the cut and paste technique, which are further exploited in Jing (2001).

Mani, Gates and Bloedorn (1999) use *full revision* to improve automatic extracts by first creating a 'draft' and then subjecting the material in this draft to three revision operations: sentence *compaction*, sentence *aggregation*, and sentence *smoothing*, which involves *coordination reduction* and *reference adjustment*. The main idea behind this combination of operations is to fit more information into less space, as compaction creates more compression space and aggregation adds more information. Sentence *compaction* operations eliminate parentheticals, sentence-initial prepositional phrases, and adverbial phrases satisfying certain lexical tests (for example *in particular*, *accordingly*, *in conclusion* etc.). Sentence *aggregation* combines constituents from two sentences, one of which must already be in the draft extract, into one new sentence, based on coreference. Sentence *smoothing* applies to a single sentence to improve its style. Within this operation, *coordination reduction* simplifies coordinated constituents, covering subject ellipsis, relative clause reduction, and relative clause coordination. *Reference adjustment* improves

discourse-level coherence and occurs last. It includes name aliasing, pronoun expansion, and indefinitization.[20]

The problems and possible solutions in automatic extracting are very closely related to the investigation in this thesis. Chapter 5 and Chapter 6 present a classification, based on a corpus analysis, of operations applied by a human summariser to extracts. The function of these operations is to produce abstracts from the extracts by improving their readability and coherence during the *summary production* stage of summarisation. A set of guidelines is developed from the classification, and is used to apply the operations to a different collection of texts in order to assess their usefulness in improving the quality of summaries (see Chapter 7).

3.5 Computer-aided summarisation

Chapter 2 and the earlier part of this chapter looked at human and automatic summarisation respectively. This section deals with *computer-aided* summarisation, a concept which falls between the two and shares ground with both. Computer-aided summarisation (CAS) is a compromise between fully human and fully automatic summarisation and can help overcome the problems by utilising the positive aspects of the other two types. It has been developed at the University of Wolverhampton (Orasan, Mitkov and Hasler 2003) as a feasible alternative to automatic summarisation in an attempt to produce summaries of higher quality than those

[20] Mani (2001: 83) defines *name aliasing* as "substitution of a proper name with a name alias if the name is mentioned earlier" (e.g. *Barry Jones… Jones*) and *indefinitization* as "replacement of a definite NP with a coreferential indefinite if the definite occurs without a prior indefinite" (e.g. replacing an initial mention *the report* with *a report*).

produced by fully automatic methods in a faster time than human summarisers can perform the task from scratch. The basic idea is that a summary of a text is produced by the interaction of a human summariser with automatic summarisation methods in order to produce a summary of the highest possible quality in a minimum amount of time.

Whilst the idea of some form of automated 'help' for human summarisers may have been around for some time (Craven 1993; Mitkov 1995; Craven 1996; Narita 2000; Narita, Kurokawa and Utsuro 2002), the more specific notion of *computer-aided summarisation* which combines automatic extracting and human post-editing, has only recently been explored in depth (Orasan, Mitkov and Hasler 2003; Orasan and Hasler forthcoming). This more specific concept was first mentioned by Mitkov (1995) as *computer-assisted abstracting*.

3.5.1 Computer assistance in natural language processing tasks

Mani (2001) distinguishes human and machine capabilities in summarisation, pointing out that it is likely that they will not be identical, but will overlap. This suggests that their activities could complement each other and that each has their own place in the field of summarisation. Therefore a combination of human and automatic methods (computer-aided summarisation) is a reasonable option as this will lead to the most effective production of summaries. He notes that computers are better at sifting through large amounts of data whilst humans are better at making inferences based on context and using real-world knowledge. Mani also notes that there is a spectrum which ranges from *Machine Assisted Human Summarization* to

Fully Automatic Summarization, with *Human Assisted Machine Summarization* in between. Other fields of computational linguistics and natural language processing also make use of computer-aided, or computer-assisted, as opposed to fully automatic technology. Indeed, Mani's spectrum reiterates a similar spectrum noted in machine translation (Hutchins and Somers 1992).

Kay (1980, reprinted 1997) proposed the development of *cooperative man-machine systems* as a solution to the unrealistic task of fully automatic high quality translation, allowing the computer and the human translator to perform the translation tasks they are best at. He advocates the use of the computer in translation, if used properly, as something which helps humans by "taking over what is mechanical and routine" and allows the productivity of the human translator to increase, as well as making their work "more rewarding, more exciting, more human" (Kay 1997: 3). It was this approach, with its idea of best utilising human and computer abilities, which inspired Orasan, Mitkov and Hasler (2003) to apply similar ideas for summarisation and develop computer-aided *summarisation*. In summarisation, the computer-aided (or cooperative man-machine, to use Kay's term) approach leaves the searching through the full document for information to the computer and the reduced-effort task of linking the units coherently, and adding or removing important or redundant information from the summary, to the human.

More than twenty years after Kay's proposal to develop cooperative man-machine systems for translation the concept is still popular, as fully automatic high quality translation still does not seem to have been achieved. An automatically translated text can be revised, or post-edited, by a human translator, or the human translator can

use tools such as translation memories and on-line dictionaries to help them translate more quickly and/or easily. The computer-aided approach is also employed in other tasks, such as the generation of multiple-choice tests from electronic instructional documents (Mitkov and Ha 2003). The idea behind Mitkov and Ha's work is to save both time and money during the construction of such tests by automating at least some of the question generation task. Their system generates questions, correct answers and distractors[21] and then gives the user the option to post-edit these outputs as required. Their evaluation showed that the computer-aided approach worked very well, with the average time taken to produce a test question falling by 74% (from 6 minutes 55 seconds when produced manually, to 1 minute 48 seconds) with no decline in quality.

Semi-automatic annotation tools also use computer assistance rather than fully automatic means to achieve a goal more efficiently than would otherwise be the case. They range from tools which help a human annotator in the annotation process to tools which annotate automatically and then need a human to post-edit the automatic annotation. Here again the distinction between machine-aided human approaches and human-aided machine approaches is evident. These tools have been used in areas of computational linguistics such as coreference annotation, discourse theory annotation (Orasan 2003a), and semantic annotation (Cunningham et al. 2002).

[21] A *distractor* (or *distracter*) is one of the possible answers to a multiple-choice question that is not the correct answer.

3.5.2 Writing abstracts with computer assistance

During the 1990s, Craven's work on computer-assisted summarisation (Craven 1996; 1998; 2000) investigated whether a basic computer-aided approach could help human abstractors. Craven focuses on the automatic extraction of keywords and phrases from documents which could be useful when presented to a human abstractor trying to summarise the document. This is classified as machine assisted human summarisation (see Section 3.5.1): the use of computer-based tools to assist a human attempting a task. Craven advocates a *hybrid system* where some tasks are performed by humans and others by software, his focus being on "providing writers of conventional abstracts with various computerized tools to assist them" (Craven 2000: 2).[22] The motivation for this is that at the time of writing computer assistance was available for the related field of indexing, but not for abstracting. He reports on a prototype computerised abstractor's assistant which presents users with words and phrases from the full text they may want to use in an abstract. These words and phrases are automatically determined on the basis of their frequency, and stop-words are omitted.

To test how useful this tool was in helping human abstractors create summaries, and to test other hypotheses related to the abstractors' backgrounds, Craven carried out an experiment, where 60 abstractors created a 250 word (maximum) summary for three documents each using Craven (1988)'s TEXNET system. The time limit was one hour, the texts were taken from three different fields (education, computer-mediated communication and information science) and no guidelines were given to

[22] The electronic version of this article, available at http://www3.interscience.wiley.com/cgi-bin/abstract/72001844/START, was used here.

the abstractors, regardless of their level of familiarity with summary-writing. As part of the evaluation, the abstractors were asked to complete a questionnaire about their experience in abstracting and their reaction to the tool. 37% of the abstractors in this experiment assessed the keywords or phrases presented to them as 'quite useful' or 'very useful' when writing their abstracts, from a scale which also included 'not very useful' and 'not at all useful'. This suggests that it is necessary to extract more than these units from a text to help a human summariser create an abstract from it.

3.5.3 Accessing templates and samples for abstracting

Narita (2000) and Narita, Kurokawa and Utsuro (2002) developed a tool which provides Japanese software engineers with a template or model on which to base their own English summary of a document. The tool presented in Narita, Kurokawa and Utsuro (2002), BEAR, is a more complete, web-based version of Narita (2000)'s *Abstract Helper*, with more search options and templates available. The aim behind the development of this abstracting tool is to improve summaries of research papers in the field of information engineering written in English by Japanese software engineers who are intermediate or advanced learners of English as a second language. The tool works on the basis that a summary is one paragraph in length and contains one *topic sentence*, which tells the reader what the text is about, the other sentences being related to it in certain logical or coherent relationships. As with Craven's work, no automatic summarisation methods are employed; instead the tool accesses a corpus of human-produced abstracts which have been analysed for their rhetorical structure in order to help the abstractor. Likewise, this falls into the category of machine assisted human summarisation, as described in Section 3.5.1.

BEAR provides an organisational template for the human abstractor to flesh out with their own material, helping them in the process by providing examples from a corpus, which forms a major part of the tool. BEAR contains four modules to help the abstractor. The first is *rhetorical template selection*, which allows the user to select an appropriate organisational template from the corpus. The *component sentence construction* module allows the user to search for samples of sentences using either sentence pattern keywords (KWIC: Key Words In Context) or sentence roles (for example, introductory, topic (obligatory), verifying or closing sentences). This module also provides access to collocations, grammatical constructions and other sentences, as well as on-line lexical look-up and spell checking. The third module is *feedback message generation* (reported as being under construction), and the fourth is *sentence concatenation* which deals with the output format (i.e., the format of the summary), presenting the sentences in a coherent abstract.

Although Narita, Kurokawa and Utsuro (2002) state that the development of the tool is not yet complete, they have tested parts of it with 27 software engineers. They gathered user feedback via questionnaire about user satisfaction and perceived utility of the software using a 5-point scale for assessment, and also obtained free-form comments. Whilst they note that the *rhetorical template selection* was not evaluated as positively as had been hoped, with the evaluations of the four software components falling on the upper half of the evaluation scale, they conclude that the tool was evaluated positively.

3.5.4 Computer-aided summarisation at the University of Wolverhampton

In a working paper in 1995, Mitkov first described plans to develop a "computer-assisted and user-friendly abstracting tool" (Mitkov 1995: 6) which identifies and highlights sentences considered to be important in terms of content for the user. Once the computer has performed these tasks, the human abstractor accepts or rejects the selected sentences as they see fit, and perhaps adds new sentences, before connecting the text together into cohesive paragraphs. He terms this approach *semi-automatic* and argues that it will make abstracting faster and cheaper as it does not rely on fully human summarisation which is time-consuming and labour-intensive. It is also argued that since fully automatic abstracting is not completely reliable all of the time, this idea of computer-assisted abstracting is a successful compromise.

This idea of selecting sentences containing important information in a text automatically and then allowing a user to edit these as necessary is the basis for the work completed more recently at the University of Wolverhampton. The CAST project (Orasan, Mitkov and Hasler 2003)[23] developed a computer-aided summarisation tool (CAST) based on this notion of human post-editing of automatic summaries.[24] Unlike the related work described above (Sections 3.5.2 and 3.5.3), *computer-aided summarisation* is classed as human assisted machine summarisation. In CAST, the human summariser interacts with automatic summarisation methods in an accessible environment in order to facilitate the task of summarisation. CAST is a

[23] More up to date information about this project can be found at http://clg.wlv.ac.uk/projects/CAST/.
[24] The summaries produced automatically by CAST are *extracts* as opposed to abstracts. Depending on the type of post-editing applied by the user, the final summary can be an extract *or* an abstract.

user-friendly tool which integrates several established automatic summarisation methods and allows a user to run them, combining, filtering and post-editing the results. It uses automatic extraction methods to produce a summary which the user can choose to be presented to them in the form of an actual summary or as suggested important sentences highlighted in the full text. Different 'important' sentences can be highlighted at the same time depending on the number of automatic methods the user selects to produce the extract, thereby offering the user a better range of what could be considered important information. The tool is also able to highlight potential problems in extracted sentences, such as dangling pronouns, which will be detrimental to the quality of the summary if they are not resolved. Experiments involving a professional summariser have shown that CAST reduced the time taken to produce summaries by approximately 20% on average, without any decrease in the quality of the abstract (Orasan and Hasler forthcoming).

Whilst the work of Craven (1996; 1998; 2000), Narita (2000) and Narita, Kurokawa and Utsuro (2002) discussed above indicates how useful the computer-aided approach for summarisation could be and therefore justifies further investigation of it, this type of approach is also linked to research in human summarisation. The work of Endres-Niggemeyer (1998) described in Section 2.4.2 provides the theoretical grounding for the idea of human post-editing in CAST in terms of the three-stage model of document exploration, relevance assessment and summary production. The first two stages correspond to the automatic summarisation in CAST, which uses automatic methods to identify important information in the source and present this, either in the form of a summary or as highlighted units within the full text, to the user. The third stage, which in Endres-Niggemeyer's analysis involves cutting and

pasting operations and reorganisation of the text, corresponds to the human summariser's acceptance or rejection and organisation of the important information produced automatically by CAST in the first two stages, and their post-editing of the extract.

Cremmins (1996)'s four stages of abstracting can also be used to locate the operations carried out in CAST in a theoretical paradigm of human summarisation, although it does not fit quite as well as the model developed by Endres-Niggemeyer. Stages one and two, which involve focusing on the basic features of the task and identifying information[25] correlate with the automatic production of an extract or the identification and highlighting of important information in the text by CAST. Stages three and four, the extraction, organisation and reduction of relevant information and the subsequent refining of this information can be seen as similar to the human post-editing applied to CAST's output. The extraction of relevant information in Cremmins' third stage is either done by CAST itself if the user requests an extract, or by the user if they prefer to see identified relevant information highlighted in the full text. However, even in the second case, the computer-aided summarisation tool still points out the relevant information to the user based on automatic extraction methods, the information is just displayed in a different way.

3.6 Conclusions

The purpose of this chapter was to offer a review of previous work in the field of automatic summarisation considered to be relevant for this thesis in order to provide

[25] These two stages are sometimes carried out simultaneously.

a rationale for the application of human-style abstracting changes to extracts to improve their readability and coherence. Section 3.2 presented some basic notions in automatic summarisation as an introduction to the field. The two types of summary created automatically, abstracts and extracts, were discussed in Sections 3.3 and 3.4, respectively. Section 3.5 discussed computer assistance for NLP tasks in general, and in summarisation more specifically. It set up the notion of *computer-aided summarisation* as a viable alternative to fully automatic summarisation and grounded this proposal in Endres-Niggemeyer (1998)'s three stage model of human summarisation.

Whilst the systems and research described above do not exhaust the existing work in the field, they do provide examples of the different ways that summaries can be produced automatically. They also provide an insight into the main limitations of automatic summarisation: the fact that extracting can produce incoherent text and that abstracting is domain-specific and therefore systems cannot easily be ported to other domains. Indeed, the brief discussion of DeJong's FRUMP system (DeJong 1982) showed that there can even be problems of coverage within the system's domain. Despite the superior coherence and often informativeness of automatic abstracts, it seems that the production of automatic extracts is still preferred because relatively less effort is required to create them.

Therefore a viable option for improving the quality of 'automatic' summarisation at present is to consider using some kind of computational assistance (Section 3.5) to utilise the best parts of automatic summarisation and human summarisation to create summaries. *Computer-aided summarisation* (Section 3.5.4) which contains an

element of human post-editing, is a workable alternative to fully automatic summarisation. The basic idea is to combine the positive aspects of automatic and human summarisation (automatic document exploration and relevance assessment with human summary production) in order to produce high quality summaries in a relatively fast time. In order to exploit computer-aided summarisation to its full potential, human post-editing operations which transform an extract into an abstract need to be analysed, classified and evaluated. The remainder of this thesis creates a middle ground between extracts and abstracts by doing just that. Some of these transformations could then be implemented in a system to improve the extracts before they are presented to the user, meaning that the user would have to spend less time editing the summary in the computer-aided summarisation environment. This approach could be seen as shallow coherence smoothing; however, the operations applied by human abstractors often amount to more than just this. Chapter 5 and Chapter 6 illustrate this via a corpus analysis and classification of operations applied by a human summariser and Chapter 7 evaluates them. The first step is to create a corpus of (extract, abstract) pairs to allow such an analysis. The next chapter presents a set of guidelines for the human annotation of important sentences in documents, i.e., for the creation of extracts, and the corpus developed for the investigation.

Chapter 4. Guidelines and annotated corpus for summarisation

4.1 Overview

Chapter 2 and Chapter 3 introduced related fields of summarisation, discussing human and automatic summarisation, respectively. Chapter 3 also presented the concept of computer-aided summarisation (CAS) as a viable alternative to a fully automatic process due to the shortcomings of current automatic methods. Computer-aided summarisation was also presented as the context in which the investigation in the remainder of this thesis is conducted. The main purpose of this chapter is to introduce the corpus developed for the analysis of human summary production operations classified in Chapter 5 and Chapter 6. To do this, it is first necessary to describe the guidelines used to produce the extracts in that corpus. This in turn requires a brief overview of existing summarisation guidelines available to summarisers, as these are taken as the basis for the guidelines developed in this thesis. Whilst there are a number of existing annotated corpora in the field of automatic summarisation, it is difficult to obtain the sets of guidelines that were used to produce them. Overviews of guidelines are usually available in conference papers reporting on annotated corpora, but guidelines as a distinct document issued to annotators are not often available. For this reason, the review of existing guidelines is limited to those available to professional abstractors. Section 4.2 provides this

review, examining the ANSI guidelines (American National Standards Institute 1997) and work by Rowley (1988), Cremmins (1996) and Borko and Bernier (1975).

Section 4.3 discusses the three stages of human summarisation as identified by Endres-Niggemeyer (1998) specifically in relation to the computer-aided summarisation of news texts and guidelines to facilitate this process. A set of guidelines for the human annotation of news texts is presented in Section 4.4, along with interesting observations and an analysis of their suitability for the annotation of news texts to produce extracts for the corpus exploited in this thesis. Section 4.5 describes the corpus developed for the investigation of human summary production operations in Chapter 5 and Chapter 6, including the texts used and how the extracts and their corresponding abstracts were produced. The chapter finishes with conclusions.

4.2 Existing guidelines for human summarisation

As discussed in Section 2.4.1, professional summarisers have at their disposal a number of resources to help ensure consistency and quality in the abstracting process. One type of resource is guidelines or standards for abstracting. These may be general guidelines for summarising endorsed by practice in the field, they may be national or area/field standards, or they may be organisation-specific. A brief overview of some of the different examples of guidelines available is given here, ranging from organisation-specific instructions to more general books about abstracting. The guidelines described below are the basis for the guidelines

developed for the human summarisation of news texts within a computer-aided summarisation environment (see Section 4.4).

4.2.1 ANSI: *Guidelines for Abstracts*

The ANSI *Guidelines for Abstracts* (American National Standards Institute 1997) are one example of national and area/field standards. They are based on those originally developed in 1971 and revised in 1979, and give guidance for the preparation of abstracts by authors and editors of texts on experimental work or descriptive or discursive studies. These guidelines offer advice on the purpose, location and authorship of abstracts, along with recommendations for specific documents ranging from journals, monographs, books and proceedings to restricted-access documents, patents and standards. They also include information about content and style, as well as examples of different types of abstracts and recommended reading on different aspects of abstracting and related activities. The most relevant sections of the ANSI guidelines to this thesis are sections 6 and 7, which deal with types of abstract and their content, and style, respectively.

The first of these relevant sections, section 6, states that informative abstracts are generally used for investigations, inquiries and surveys, i.e., documents which tend to follow a predefined structure, and should include the purpose, methodology results and conclusions given in the original document. Indicative abstracts, on the other hand, are deemed more appropriate for less-structured or longer documents, or documents which do not tend to contain methodologies or results sections. The examples given include editorials, books, conference proceedings and bibliographies.

The guidelines state that a complete abstract contains specific elements: purpose, methodology, results, conclusions and collateral and other information.

Section 7 of the ANSI guidelines gives details about the style of abstracts, stating that an abstract must be intelligible to the reader without them having to access the source document. In general terms, footnotes and references should be avoided, and the balance and emphasis of the original should be retained. The abstractor should use transitional words and phrases to ensure coherence, and the abstract should be concise. Where there is no abstract length already specified by others involved in producing the abstract, ANSI gives specific maximum lengths depending on the source document. These vary from a single page or 300 words for long documents such as theses, to 30 words for editorials and letters to the editor. The guidelines advise that an abstract should generally be written as a single paragraph except in the case of structured abstracts, where the main points of the source are presented in different labelled paragraphs in the abstract. In addition, complete sentences should be used, and active rather than passive verbs are preferred, although the passive voice can be used for emphasis, as long as this reflects the source. Unfamiliar terms, acronyms, abbreviations etc., should be avoided or defined the first time they appear in the abstract, and non-textual materials such as tables and equations should only be included as a last resort.

4.2.2 Rowley: *Abstracting and Indexing*

Rowley (1988) provides an introduction to the key practices in abstracting and indexing intended mainly for students of library and information science, as well as

others who might be involved in organising and exploiting information, such as managers, computer scientists and administrators. *Abstracting and Indexing* is based on experience and observations in the field. Pointing out that the main aim of an abstract is as a time-saving device for the reader, she offers advice which often takes into account the practicalities of abstracting and indexing services or organisations. In this sense, her work is similar to that of Borko and Bernier (1975) discussed below.

A 5-step abstracting procedure, of reading for content, writing notes, drafting a rough abstract, checking, and writing the final summary, is outlined, although Rowley notes that an experienced abstractor will not necessarily carry out all five steps individually, rather they will often merge steps together or perform them simultaneously. These steps find equivalents in the work by Cremmins (1996) and Endres-Niggemeyer (1998) (see Section 2.4), although their steps or procedures are given as observations of abstractors and the abstracting process whereas Rowley issues them as a set of guidelines to be adhered to by abstractors.

In her guidelines for the style and content of abstracts, Rowley acknowledges that style, content and length depend on the nature of the original document, for example, its length, scope, language, availability and author's style. The anticipated use of the abstract and the requirements of its users, as well as the resources of the abstracting agency including staff, budget, guidelines and computer processing demands are also stated as affecting style and content. This idea, as interpreted by other authors in more detail (Sparck Jones 1999; Tucker 1999; Orasan 2006), is further discussed under the heading *context factors* in Section 2.3. Rowley then goes on to discuss

several points of style which are relevant regardless of the factors mentioned above. She highlights brevity and clarity, and states that one of the main tasks in abstracting is "to convey the maximum quantity of information using the minimum number of words" (Rowley 1988: 26). She also advocates mirroring the source document author's style and ordering unless there are specific reasons to do otherwise.

The presence of a *topic* or *lead sentence* at the beginning of a summary is important because it summarises any essential information not given in the title, and helps the reader to decide whether the text is relevant to them. Abstractors are advised not to use long sentences, an average of 12 words being given as a suitable sentence length for a readable abstract. An abstract should usually be one paragraph in length and should be coherent, with all sentences except the first being complete and including verbs, prepositions and articles. The numbering and listing of points within a sentence is stated as being acceptable and widely used, especially in indicative abstracts. The abstractor is instructed to avoid using ambiguous words and terms with unclear meanings. Examples of lexical units which can aid brevity but can also be confusing and should therefore be used with caution are abbreviations, acronyms, trade names and subject jargon. Rowley points out that many abstracting organisations use a standard list of abbreviations that are suitable for inclusion in abstracts. As emphasised elsewhere by Rowley, conciseness is considered essential. This means that the abstractor should prefer the active voice over the passive, using the simple past and present tenses. They should also remove redundant and verbose clauses, phrases and words. As far as content goes, the instruction is simple: it must reflect the source document.

These points on style are similar to those described in the ANSI guidelines. Also similar to the ANSI publication, Rowley discusses what abstracts of different types of source documents should include and what types of abstract are suitable for these different types. Abstracts of research papers, reports and journal articles will typically include purpose and scope, methodology, results, conclusions and incidental findings. Indicative abstracts are considered to be the most suitable type of abstract for reviews and surveys, bibliographies, monographs and conference proceedings in their entirety. Informative abstracts are considered to be most suitable for patents, individual contributions to monographs and conference proceedings, and research papers, reports and journal articles. Rowley also briefly details guidelines used by abstracting organisations, which include instructions regarding presentation, style, length, language, citations, proofreading procedures and abbreviations, among other things.

4.2.3 Cremmins: *The Art of Abstracting*

Cremmins describes his book *The Art of Abstracting* (Cremmins 1996) as a guidebook for the writing, editing and revising of abstracts, based on his experience as an abstractor and his observations from within the field of abstracting. Throughout the book, Cremmins emphasises the interaction of an abstractor's reading, writing, editing and revising skills, and addresses these from the point of view of *analytical reading* for abstracting, which the abstractor should perform with abstracting rules and conventions in mind (see Section 2.4.3 for more details). One basic tenet is the role of the abstractor as information reductionist, condensing information throughout the whole abstracting process, which he sees as a cycle of reading, thinking, writing,

and editing or revising, advising the abstractor to "omit needless sentences and needless words and phrases within necessary sentences" (Cremmins 1996: 13).

Similar to the ANSI guidelines and Rowley (1988)'s work, Cremmins focuses on texts which tend to have overtly exploitable structures, such as experimental research and scientific documents containing distinct methods and conclusions. As well as suggestions for abstractors to help them think and write both clearly and thoroughly, Cremmins states a number of other pointers to ensure the abstract is of high quality. There are 10 general suggestions for the preparation of standard abstracts:

1. Prepare an abstract that access information services can reproduce with little or no change, copyright permitting.
2. State the purpose, methods, results, or findings, and conclusions or recommendations that are presented in the original document, either in that order or with initial emphasis on results and conclusions.
3. Make the abstract as informative as the nature of the document will permit, so that the readers may decide, quickly and accurately, whether they need to read the entire document.
4. Unless otherwise instructed, use fewer than 250 words for most papers and portions of monographs and fewer than 100 words for notes and short communications. For long reports and theses, do not exceed 500 words.
5. Avoid including background information or citing the works of others in the abstract, unless the studies are replications or evaluations of their works.
6. Do not include information in the abstract that is not included in the textual material being abstracted.

7. Verify that all quantitative or qualitative information used in the abstract agrees with the information contained in the full text of the document.
8. Use standard English and precise technical terms, and follow conventional grammar and punctuation rules.
9. Give expanded versions of lesser known abbreviations and acronyms, and verbalize symbols that may be unfamiliar to readers of the abstract.
10. Omit needless words, phrases, and sentences.

(Cremmins 1996: 14-15)

4.2.4 Borko and Bernier: *Abstracting Concepts and Methods*

Abstracting Concepts and Methods stemmed from discussions between two teachers of abstracting on what to teach their students and how to organise their classes. In their book, Borko and Bernier (1975) aim, amongst other things, to "provide a basis for a well-balanced course of instruction" (Borko and Bernier 1975: x). Their work is aimed at students and teachers of abstracting, and provides the reader with a practical view of not just abstracting techniques, but also abstracting services and abstracting as a profession, as does Rowley (1988). They argue that because abstracting is a particular type of literature with its own particular style and attributes, rather than a 'natural' form of writing, training and guidance are necessary to enable abstractors to produce a high-quality abstract. Feedback is considered important to allow the abstractor to hone their skills and constantly improve, meaning that in abstracting services, editors play a vital role. Borko and Bernier also present information about specific guidelines, instructions and standards which either come from previous research or are used by particular abstracting services or other organisations dealing with abstracts.

Similar to other work discussed in this section (Rowley 1988; Cremmins 1996), the authors advocate "Brevity without loss of novelty…" (Borko and Bernier 1975: 10) and see abstracting as a way of removing redundancy. In contrast however, Borko and Bernier reject the setting of a length restriction for abstracts, claiming that once the abstractor understands and employs all the techniques for achieving brevity, length will take care of itself. Another reason for their argument is that abstracts can vary widely in length and still be acceptable, depending on the source and on the task in hand. However, as a guide, the recommended length can be taken as approximately 10% of the source or 200 words. Standardisation is seen as important, as abstracts need to be uniformly accurate and free from error. Abbreviations, acronyms, symbols and citations should all be standardised to minimise mistakes, especially if an abstractor is issued with a list from an abstracting service. Abstractors are also encouraged to write clearly, using complete sentences, and, unless there is ambiguity, to use the author's own words to avoid changes in meaning between the source and its representation in the abstract. However, Borko and Bernier point out that abstractors always paraphrase the source, in order to abstract the document as opposed to extracting it.

Background knowledge such as history, introductions, and details of procedure, as well as what the author did not do, and what they plan to do next should be omitted from abstracts. This opinion concerning the omission of details of procedure differs from other guidelines (such as ANSI (1997)), which often advocate the inclusion of methodology, particularly for abstracts of research papers. The content of the abstract should be arranged to save the reader's time, by, for example, placing an important

conclusion at the beginning. The abstract should consist of one paragraph only, and should not contain labels or headings. After writing the abstract, the abstractor must check for errors, omission of important information, and adherence to appropriate policies and rules. If the abstractor is working within a larger organisation, as the authors tend to assume, their abstract will be subject to checking and editing by others after being written and checked by the abstractor.

Borko and Bernier note that all major abstracting and indexing services, as well as many smaller ones, issue their abstractors with guidelines which help to best achieve the end product for their particular customers. As well as guidelines, instructions and standards, four main types of training can also be provided: longer courses in educational institutions, such as degree programmes in universities; shorter courses or attendance at training institutes; on-the-job training, which includes tutorial training, feedback and review, and using published instruction manuals; and self-teaching via existing manuals, instructions and guidelines. In a similar manner to abstracting services, this thesis proposes guidelines for the computer-aided summarisation of news texts.

4.3 The three stages of summarisation and the computer-aided summarisation of news texts

Having presented several examples of available existing guidance for professional human abstractors, this section re-examines the three stages of human summarisation identified in Chapter 2 in the context of developing guidelines for summarisation. It further rationalises the human model taken as the basis for computer-aided

summarisation, by highlighting the separation of the different steps in summarisation and illustrating that different types of guidelines are necessary for some of the different stages. It also shows that within computer-aided summarisation there are already mechanisms in place to select important information from texts, but not really for the improvement of readability, and emphasises exactly how humans and machines can interact to produce high-quality summaries. It is worth pointing out here that there are no guidelines available for professional abstractors regarding the summarisation of news texts, as these texts are most often used in automatic summarisation and are not texts which professional abstractors typically summarise. Guidelines for humans involved in the automatic summarisation process usually focus on the annotation of important units, i.e., units which should (and/or should not) be included in a summary, and not on any issues of coherence, readability, or editing. They deal with the information in a source text and producing an extract from it, and not on an extract and producing an abstract from it. This makes the guidelines regarding the summary production stage of news text summarisation presented in Chapter 6 novel. Those presented in this chapter can also be considered as an original contribution to the field of computer-aided summarisation.

As discussed in Section 3.5.4, the three-stage model of human summarisation identified by Endres-Niggemeyer (1998) provides the theoretical grounding for the idea of human post-editing in the CAST system. The first two stages of document exploration and relevance assessment correspond to the automatic summarisation in CAST, which uses automatic methods to identify important sentences in the source and presents these to the user. The third stage, summary production, which involves cutting and pasting operations and reorganisation of the text, corresponds to a human

summariser's acceptance or rejection of the highlighted important information produced automatically by CAST in the first two stages, and their post-editing of the summary. Endres-Niggemeyer's model can also be used in a discussion of the computer-aided summarisation of news texts and the development of guidelines within this specific area.

4.3.1 Document exploration

In terms of *document exploration*, work has been done on the structure of news articles and how information is introduced, or where it is located in the document (van Dijk 1988). News articles are seen as having a top-down instalment structure, meaning that although the most important information is introduced first, there are 'degrees' of importance that are introduced in a cyclical manner. For example, in an article about a terrorist attack, "main participants and acts that are politically relevant come first, followed in each cycle by details of main participants, identity of secondary participants, components/conditions/consequences/manner of acts, Time and Location details etc." (van Dijk 1988: 48). Hasler (2003) uses van Dijk's observations to explain why three human summarisers annotated important information in 'similar' texts in the way that they did, selecting the first sentence of the article for inclusion in a summary in 70% of cases. Section 2.4.4 gives details about research regarding the structure of documents and abstracts in the scientific and medical domains. The fact that important information tends to appear in certain positions in documents is exploited in automatic and computer-aided summarisation by using extraction methods which give text appearing in certain positions a higher weighting than that appearing elsewhere (see, for example, Lin and Hovy (1997)).

The idea of a *lead summary*, which often performs better than other automatic summarisation methods for newswire texts, is a prime example of exploiting the document exploration step to produce summaries.

Regarding the guidelines developed for the annotation of important units in news texts (Section 4.4), document exploration relates to instructions such as including headings and sub-headings, and identifying the most important topic and selecting sentences referring to this. It also covers aspects which summarisers may not wish to include, particularly in the field of automatic (or computer-aided) summarisation, such as examples and direct speech, as well as units within selected sentences which are not considered important and can therefore be deleted from those sentences, such as text in brackets and between dashes.

4.3.2 Relevance assessment

Linked to the above step of *document exploration* is the second summarisation stage of *relevance assessment*, where the summariser assesses information in the document to see if it is relevant to the summary by identifying the document theme. Relevant information about the theme should be present in certain positions within the document, identified in the first stage. In the case of news texts, the 'headline' (or title) and lead paragraph, or at least the first sentence of the document, are usually good indicators of what the document is about. Relevant information is also indicated in documents through certain repetitions and rhetorical emphasis. As discussed in Section 3.4, automatic summarisation systems often exploit word or phrase frequency in order to identify 'important' sentences for extraction. In addition, cue or

indicating phrases can be seen as utilising rhetorical emphasis, whilst position/location and title overlap methods exploit the document's structure and the position of information in assessing relevance.

In terms of the annotation guidelines developed in this chapter, instructions to identify the main topic of the document and annotate sentences about this as important or suitable for inclusion in a summary fall under this heading. A news text often has a 'main topic' which is introduced either in the headline/title or probably more informatively in the first sentence(s) or paragraph. This is closely related to document exploration and again links in with the notion of a lead summary as previously discussed. There is often more than one topic present in news texts: sometimes secondary topics which are related to the main one but are not what the document is essentially about, and sometimes topics which can be seen as topics in their own right, but which are presented after the first main topic, or do not have as much space devoted to them, indicating that they are not as important as the topic presented first. It is important for the summariser to be able to identify this main topic, or what the document is essentially 'about', because without this step, the corresponding summary could contain all kinds of irrelevant or less relevant information.

Automatic summarisation methods and systems for extraction use a number of well-established and reasonably robust methods to solve the *aboutness* problem, or identify the main topic or theme of the document to be summarised. These are briefly discussed in Section 3.4. When humans summarise documents and when they annotate texts for summarisation in the fields of automatic or computer-aided

summarisation, they need guidelines to help them and to remind them that they need to identify the main topic. The guidelines presented in Section 4.4 contain elements relating to this step of summarisation.

4.3.3 Summary production

The existing guidelines and advice described in the first part of this chapter indicate the kind of advice given to professional abstractors in the field of human summarisation. Some aspects of these are concerned with the style of the abstract, and therefore come under Endres-Niggemeyer's heading of summary production. However, such advice is rare for human annotators or summarisers in the field of automatic summarisation. This is the least addressed stage of summarisation in terms of guidelines for human annotation in automatic summarisation, because humans are most often asked to annotate sentences within texts which they see as important and therefore worthy of including in a summary of the text they are annotating.

As this thesis attempts to bridge the gap between the fields of human and automatic summarisation by presenting a classification of human operations used to transform extracts into abstracts, it also offers guidelines as to how human summarisers or users of computer-aided summarisation systems can ensure that their summaries are readable and coherent. These guidelines regarding summary production (or style in terms of professional abstracting) must necessarily be presented after the classification of operations on which they are based. Therefore Section 4.4 does not address this stage of human summarisation; instead focusing on guidelines which

embody the first two stages only. Summary production guidelines are dealt with further in Chapter 6.

In automatic extraction the summary production stage is addressed, if it is addressed at all, by revision operations such as deleting dangling discourse connectives, resolving dangling pronouns and the use of pronominalisation to avoid repetition (see Section 3.4.7). However, the following two chapters of this thesis prove that these revision operations do not account for all of the changes a human summariser makes to an extract to improve its readability and coherence. Therefore it is necessary to explore how to establish this wider range of summary production stage improvements as an important part of the computer-aided summarisation process.

4.4 Guidelines for the annotation of news texts for summarisation: 2003 annotation task

This section presents a set of guidelines developed for the human summarisation (extraction) of news texts. These guidelines are based on the advice for professional abstractors discussed above (Section 4.2), as well as a preliminary investigation of news texts, and were used, with some amendments, to annotate the corpus described in Section 4.5. The guidelines described in this section are very different to those presented above, being developed for a different type of text, and to be used in automatic summarisation for the purposes of training and testing. Because of this, the guidelines are particularly suitable in the simulation of the first two stages of summarisation within a computer-aided framework, resulting in reliable extracts which are then post-edited. In computer-aided summarisation, the user of the system

takes a set of extracted sentences as the starting point for post-editing, and to gain reliable insights into this process, it is necessary to start with the same type of document, i.e., an extract rather than a source text. The remaining sections of this chapter address that need. By using human-produced extracts for the corpus, the best possible set of sentences in terms of informativeness is used, allowing the analysis to concentrate wholly on summary production rather than on any aspects of the previous two stages. If a corpus of automatic extracts was used as a starting point for the analysis, issues regarding document exploration and, particularly, relevance assessment would need to be focused upon in detail, meaning that improvements regarding information content as well as editing for coherence and readability would be necessary. This would detract from the main aim of this thesis, which is to identify ways of improving extracts via an investigation into human summary production operations.

4.4.1 General discussion

This section discusses guidelines issued to human summarisers to help them produce extracts of news texts. Guidelines are an essential tool in the attempt to ensure consistency when humans play a part in the summarisation process. It can be very difficult for a human to decide exactly what should be considered ‘important’ in a document, and what types of information should not be considered so. Without guidelines, summaries produced by different summarisers can be very different. Even when guidelines are used, this is sometimes the case, due to the subjectivity of the notion of ‘importance’ (see, for example, Hasler, Orasan and Mitkov (2003)). However, if guidelines are provided to summarisers, it is more likely that their

summaries will be consistent than if they are not. As mentioned above, the guidelines presented here do not deal with any aspect of summary production, as an extract, by definition, is a summary which is produced by taking text from the source verbatim. In addition, when annotating texts for summarisation, the annotator simply selects an important sentence using an annotation tool and therefore does not technically engage in what Endres-Niggemeyer (1998) terms summary production. Extracts created by an annotator form half of the corpus described in Section 4.5, the other half comprising human-produced abstracts based on these extracts.

The guidelines discussed in this section are based on those presented in Hasler, Orasan and Mitkov (2003) where they were developed to annotate the CAST corpus for summarisation. They were formulated after examining newswire and popular science articles, from *Reuters* (Rose, Stevenson and Whitehead 2002) and *New Scientist*, extracted from the *British National Corpus* (*BNC*) (Burnard 1995), respectively. These two types of text share many similar properties: essentially, both can be classed as *news texts* because of the way they present their content to the reader. Whilst they may differ in their purpose, audience, and subject matter, both types of text start with a 'headline' or title and a journalistic-style *lead paragraph* which serves to summarise and introduce the rest of the document. By using both newswire and popular science articles, it is possible to gain a better insight into different styles of texts which can be considered news texts due to their similarities, and to develop guidelines and carry out analyses which are not as restrictive as they perhaps would be otherwise. It was therefore deemed appropriate to use both types of text to develop the guidelines, and to term the resulting corpus a corpus of *news texts*. Although the guidelines used to annotate them are similar, the CAST corpus (Hasler,

Orasan and Mitkov 2003) is different from the corpus of (extract, abstract) pairs used in this thesis. The adapted version of the guidelines, used to annotate texts for the corpus analysis in this thesis, can be found in Appendix I. The discussion below contains examples from the CAST corpus, because those texts were used to develop the initial set of guidelines. The observations and assessment that follow are also based on that corpus, which functioned to test whether the guidelines were suitable for use in further annotation tasks, but only those parts relevant to the present research are addressed, in more depth where appropriate.

4.4.2 The 2003 annotation guidelines

Whilst the guidelines are designed to advise annotators how to mark important information in the documents easily and consistently, it is important to remember that there will always be cases which prove to be the exception to the rule. This is why the opportunity is provided for the annotators to give comments and explanations throughout the annotation process, and also why the instructions usually carry the caveat 'unless crucial to the main topic'. Annotators' comments were used in combination with a discussion in order to adapt the guidelines to make them more suitable for further annotation tasks such as the one completed to create the extracts in the corpus analysed in Chapter 5 and Chapter 6. A length restriction of 30% was imposed on the amount of sentences in the source text that could be marked; 15% to be marked as *essential* and 15% as *important*. This decision was based on a combination of the human and automatic summarisation literature and a preliminary investigation of the texts to be annotated for the corpus. The fact that two options for labelling important sentences are offered to annotators means that it is possible to

have two summaries of a document, a 15% summary which should contain the most important information (these sentences are labelled *essential*) as well as a 30% summary which contains more information. This distinction is useful when creating abstracts from these 30% extracts because there is already an indication of the most important information.

Marking important information

In addition to general strategies concerning 'best practice' for summarisation annotation, more specific guidelines were formulated based on the news texts to be annotated for the corpus. The annotators were instructed to identify the main topic of the text and mark sentences[26] which gave *essential* information about the topic, keeping as close to 15% of the full text as possible. A news text is usually 'about' one main topic, that is, it tends to concentrate on one main focus, although it can have other secondary topics related to the main one. This means that the information which is considered suitable for inclusion in a summary (i.e., marked as *essential*/*important*) should relate to this topic. A good indicator of this is usually the 'headline' and first sentence. Sub-headings generally summarise the text which follows and are consequently useful to mark as important. Unlike titles, they are not automatically included in the list of extracted sentences, so they need to be marked explicitly. The annotators were told to mark sub-headings as long as they are relevant to the main topic of the text.

[26] For details on the annotation scheme and annotation tool used, see Hasler, Orasan and Mitkov (2003).

Due to referring expressions there are sentences which rely on others for full understanding. An example of such a pair of sentences is:

a) *For film lovers* ***the Festival****'s the place to be in September.*
b) ***It*** *grows from strength to strength each year.*

In the annotation process, if b) is marked, the sentence containing the antecedent for *it*, should also be marked in some way. If a) is important enough, it will already be in the list of selected sentences, otherwise, it should be marked as a *referred* sentence, i.e., a sentence which does not contain important information as such, but is necessary for the full understanding of another marked sentence, for example, it contains the antecedent of a pronoun.

Tables and figures are not usually necessary in summaries of news texts, and therefore should not be marked,[27] along with sentences concerning sub-topics unless they directly influence the main topic, or present new, essential information on it, and do not repeat information. Examples, including constructions starting with *e.g.*, *for example*, *such as*, *like*, *for instance*, etc., are not generally suitable for inclusion in a summary as they serve to explicitly elaborate information which is already given, using up valuable space.

Direct speech should not be marked as important unless it presents vital and new information concerning the main topic of the text which is not presented elsewhere. Direct speech in news texts tends to provide opinions or statements to emphasise

[27] Tables and figures may be appropriate in summaries of other types or sub-domains of text or document where they best summarise the information present, for example, in a summary of a day's trading on various share indexes.

points already made in the article and does not usually warrant marking. It also has a different format to other information presented in the document, and may therefore take slightly longer for the reader to process than text which follows the same format as the rest of the summary. In a field where time is of the essence (summaries are often used as a time-saving device of one kind or another), this can really matter. However, it is important to distinguish between direct speech and other text in quotation marks which may be important, especially if any potential for extending these annotations for use in automatic summarisation, where the format of text can be used as a discriminating factor, is considered. Mitkov, Le Roux and Descles (1994)'s rejection rules include the instruction to eliminate text in quotation marks, which could result in the following important sentence from a text in the corpus annotated here being excluded from a summary:

> *But its "killer app" is reinventing how the software industry works.*

Sentences containing the same information as others which are marked should not be included, and each sentence should be considered carefully before selection. It is important to select the sentence with the most appropriate information, and not to include similar sentences, as this will increase redundancy and take up valuable space. The most appropriate sentence is not necessarily the longest or most descriptive one, but that which most succinctly expresses the essential information. To compare two sentences similar in information content, consider the following pair, where a) is preferable (in most cases) to b):

> a) *Inflow of export proceeds picking up: $300 million likely by February 15.*

b) *The inflow of stuck-up export proceeds has picked up pace and at least $300 million are expected before the dead-line of February 15, say banking sources.*

Removing unnecessary segments

Within the sentences which an annotator marks as important, there may be parts which are not relevant to the overall importance, or which repeat information present elsewhere. It is better to remove these parts of the sentences to minimise redundancy and maximise relevant information in the space available. Having marked the *essential* sentences in the text, the annotators were instructed to indicate segments (not single words) of these which were not vital to the understanding of the main topic as suitable for *removal*. This particular annotation overlaps to some extent with the summary production operation *deletion* (Section 5.4) observed in the corpus analysis in Chapter 5 and Chapter 6 in that it is concerned with removing parts of sentences to meet a specified compression rate. However, *removal* is not seen as a summary production operation because in both the CAST corpus and the one analysed in Chapter 5 and Chapter 6, sentences rather than smaller units are the unit of extraction from the source and there may be irrelevant or redundant parts of sentences that are not suitable for inclusion in an *extract*. *Removal* is therefore part of the relevance assessment stage of summarisation, similar to the marking of *essential* and *important* sentences. The *deletion* operation, on the other hand, is part of the summary production stage because it deals with a transformation which occurs during the production of an abstract, being used to reduce the text further or as part of another operation.

Within sentences already marked as *essential*, irrelevant subordinate clauses should be marked as *remove*. The relative clause introduced by *which* could be removed from the following sentence if it is not important within the context:

> *Customer interest is high for the whole product line~~, which underlines the strong fundamentals of the new period of growth~~.*

Text in brackets and text occurring between dashes should be removed unless central to the main topic. This text usually consists of elaborations and further explanations, which the writer has indicated, via the formatting, are not of such high importance. For example, bracketed text would be considered important if it is an abbreviation that will replace a noun phrase in a sentence appearing later which is then marked as important, as in:

> a) *A Poverty Reduction and Growth Facility* ***(PRGF)****…*
> b) *It is intended that* ***PRGF****-supported programmes…*

Adjuncts which specify dates, times and places should be marked for removal unless they are vital to the main topic, for example, if the date that something happened is a crucial aspect of the text. Likewise, examples (see above) should be removed, as should phrases such as *in addition to…*, *due to…* and *compared to…* which elaborate information and are therefore unnecessary in a summary which is designed to convey only the main points about the main topic. Reporting constructions like *a spokesman said*, *it was claimed* and *he explains* should be removed unless they need to be included for the set of marked sentences to make sense as a whole (as an extract).

Other annotation issues

Once this selection process of *essential* and *remove* was complete, the annotators were advised that if the total amount of marked text was substantially below 15% of the full text,[28] they should try to add more units which they considered essential to increase the percentage. Having completed the annotation for the *essential* classification, they had to repeat the process (using the same guidelines) for units of text they considered *important*, again keeping as close to 15% as possible. The annotators were also asked to comment on the annotation process noting any problems or indecisions. The annotators used a multi-purpose annotation tool, *PALinkA* (Orasan 2003a), to mark the important sentences in the full documents to produce the extracts.

4.4.3 Interesting observations of the 2003 annotation task

The observations discussed here stem from the annotation of the CAST corpus described more fully in Hasler, Orasan and Mitkov (2003), where the annotation guidelines described above were used by four annotators to annotate newswire from *Reuters* and popular science texts from *New Scientist* for summarisation. Chapter 7 uses a selection of texts from the CAST corpus to evaluate the classification and guidelines for summary production formulated in the following chapters. Not all the phenomena observed are reported here; due to space reasons only those of most relevance are presented. An analysis of the 2003 annotation was carried out to ascertain the suitability of the developed guidelines for further summarisation

[28] Marking a unit as *remove* subtracts that unit from the whole percentage of marked text, i.e., removing parts of sentences automatically lowers the amount of compression used so that the annotator has more space left to include other sentences.

annotation tasks. The guidelines were then adapted slightly and used to develop the corpus analysed in this thesis, described in Section 4.5.

Removed sentence segments

The most interesting part of the CAST annotation analysis for the present research is the discussion of removed sentence segments, as this can be compared with the deletions made by a human summariser when transforming an extract into an abstract. It also served to verify the instructions in the guidelines regarding what not to include and what to remove in a summary. Again, it should be pointed out that although *removal* and *deletion* do overlap, they are considered to fall into different stages of the summarisation process due to their function at the time that they are applied. *Removal* functions to create the most concise extract possible in the second stage of relevance assessment, whereas *delete* functions to improve an extract by playing a part in transforming it into an abstract during summary production. A comparison between *removed* segments found in the CAST corpus and instances of the *deletion* operation in the corpus developed in this thesis is given in Section 5.4.11. A manual analysis of the removed segments demonstrated patterns with regard to the types of constructions that are likely to contain irrelevant information within important sentences and from which it is possible to learn what kind of information is not desirable for inclusion in a summary. The types of information removed fall into several well-defined categories, each of which contain certain types of constructions. Many of the removed segments qualify previous information given in a text. It should be noted that, although instructed not to do so, annotators removed single words from marked sentences. On reflection, this is not such a bad idea due to

the fact that summaries are necessarily shortened versions of the full text and the compression rate was stated explicitly, and the guidelines have been changed accordingly.

The first category of removed sentence segments is that containing information referring to time and location. This information is characterised by constructions such as prepositional phrases (*in…*, *at…*, *on…*, *by…*) and adverb phrases (*during…*, *elsewhere…*) acting as temporal adjuncts. The second category includes information concerning the coverage of events via both direct and indirect speech. Whilst the actual speech or the gist of what was said may be important, who reported this was not usually so. Reporting verbs, such as *said*, *writes*, *revealed*, *pointed out*, and *alleged* either preceded or followed by a nominal group (usually person, organisation, type of report) are characteristic of this information.

General information irrelevant to the main topic of the text was mainly presented in subordinate clauses. This provides another group of removed sentence segments. These clauses contained different types of information, but were generally exemplified by starting with *where*, *when*, *with*, *as*, *after*, and *due to*. There were a number of coordinate clauses joined with the conjunctions *and* and *but* which were also removed. This occurred when the second coordinate clause elaborated information in the first and was therefore not needed as it increased redundancy, or when this clause was not pertinent to the main topic of the text.

Relative clauses introduced by *which*, *who* and *that* qualifying nominal phrases make up a fourth category. These served to give further information about an entity in the

text which is not needed in a summary. Examples typically started with *usually*, *such as*, *like*, *especially*, and comprise a fifth group of removed text. Another group is text in between or following dashes, which again qualified preceding information or entities. The seventh and final category of removed segments of sentences is more general and consists of other ways of elaborating textual units, such the use of the prepositions *to*, *from*, *by* and *for* to present more information about a change that had occurred. Words introducing contradictory information like *though*, *although*, *even though*, *versus*, *instead* started a number of the segments. Constructions following the *in a bid*+infinitive structure were also removed quite frequently.

Low interannotator agreement

Another interesting observation was low interannotator agreement on sentence selection. However, whilst annotators did not tend to agree on which sentences to select for inclusion in a summary, they did agree on the information that they included. The similarity was measured for texts annotated by both two and three annotators. The texts often contained similar information in more than one sentence, and the annotators did not necessarily mark the same sentences as important. This resulted in the low Kappa statistic (Siegel and Castellan 1988) value for interannotator agreement of *essential* and *important* sentences of 0.35. However, when the cosine distance (Salton and McGill 1983) was used instead to measure the similarity of the information rather than the actual selected sentences, the value was 0.73.

This is of interest to the operations discussed in Chapter 5 and Chapter 6 because it shows that even in extracts there are different ways of presenting the same or similar information and indicates that it is not necessarily the sentences which are the most important element of the text to mark, but the actual information that one would want to include in a summary. This is very closely related to the way human summarisers transform extracts into abstracts because they take the extracted sentences as a starting point and then apply operations which allow the same information to be expressed in a different way.[29] The human summariser's task is much less restricted than that of the human annotator, because they do not just have to choose the surface realisation in one sentence out of a possible several, but can provide their own alternative realisation if this is necessary. The human summariser is not restricted to using sentences or even words from the extract; they can combine the 'best' parts of the available sentences as well as provide their own input if they so desire.

4.4.4 An assessment of the 2003 guidelines

A discussion with the annotators of the CAST corpus generated several points about the suitability of the guidelines developed to facilitate their annotation. Although the guidelines seemed generally helpful and the comments positive, there are ways in which they could be improved to aid more consistent annotation in the future. The guidelines are useful in the way that they indicate what should and should not be marked, and give appropriate examples to illustrate instructions. They cover most of the different situations and alternatives that the annotators faced, are concise and

[29] It should be noted that this is not always the case. The human summariser also chose to keep original extract sentences and copy them straight into the abstract without changing them in any way at all.

clear, and strict in relation to the amount of sentences to be marked. However in some cases annotators found it difficult to adhere to the 15% *essential* and 15% *important* restriction, stating that it was not easy to distinguish between these types of sentence in a text, and also that more than 15% could have been considered *essential* or *important*.

An analysis of the corpus showed that there were a number of exceptions to the instructions given in the guidelines. This had, however, been pre-empted and was the reason for the inclusion of "unless vital to the main topic" in many of the instructions and of a *comments* option in the annotation tool. Dates and times were important in some contexts, for example, documents where time was an important factor due to a discussion of deadlines. Direct speech was also considered important at times, being marked by all four annotators, as were reporting clauses giving details of the speaker. This was especially true in cases of conflicting arguments from two different organisations or people, as in texts about conflict or war.

A more detailed investigation of direct speech in the CAST corpus showed that marked direct speech sentences had a different function to those which were unmarked. Marked speech had two main uses or functions. The first was to function as a standard non-speech sentence, i.e., as a statement of fact which, when its quotation marks and reporting clause (if present) are removed, does not necessarily need to be attributed to any speaker. The second was to convey opinions where opinions and their speaker are important to the overall understanding of the main topic, meaning that the fact that the speech is attributed to a particular person is important. Unmarked speech sentences elaborate and provide supporting quotations

to emphasise information already presented. They also tend to include opinions, but ones which are less relevant. See Hasler (2004b) for a fuller discussion of direct speech in the summarisation of news texts.

In light of these observations, the guidelines were changed accordingly. It was decided that instead of having concrete instructions about these phenomena, the guidelines should indicate that they should be treated with caution because their suitability for inclusion varies depending on the context. The annotator should be made aware of them but should be allowed to decide for themselves in the context of the full text to make sure that essential information is not discriminated against just because of its surface form: function is more important in these cases.

In contrast, sub-headings were not very relevant when they did occur, which was much more rarely than predicted. When they did occur in the documents, they tended to be subjective 'asides' on a particular situation and made no sense taken out of context, without the complete source text. As a consequence, this instruction was removed from the guidelines.

One obvious, but easily fixable, problem was a general misunderstanding of the term *figures*, from the instruction not to include tables and figures. Some of the annotators understood this to mean numerical figures, whereas the intended meaning was illustrations such as graphs. Instructions such as this should be explained more clearly to minimise confusion; this particular instruction has been changed in the guidelines. It is not relevant for the corpus annotated in this thesis as there were no tables or figures present in any of the texts, but the instruction was kept in the

guidelines because they could feasibly appear in other news texts and the guidelines can be reused in future. Many of the annotators' comments were concerned with the fact that they found it difficult to identify the main topics in fields they were not familiar with. For this reason it was decided to use a human summariser with experience in the field of summarisation[30] to produce both the extracts and abstracts for the corpus described in Section 4.5, as this would save time and result in more reliable abstracts. The guidelines discussed in this section, including the relevant amendments, were used to annotate the source texts to create the extracts analysed in the following two chapters.

4.5 Corpus Description

This section describes the corpus developed for the research in this thesis, comprising extracts and abstracts which are then analysed to formulate the classification and summary production guidelines presented in the next two chapters. The texts used in the corpus are described first, followed by a discussion of the annotation of extracts and the production of abstracts from them. This discussion deals with the corpus in two parts: extracts and abstracts, because they were produced in two distinct steps. The extracts and abstracts in the corpus were produced by one summariser (see footnote 30 above), and the annotation tool *PALinkA* (Orasan 2003a) was used to mark information in the source texts.

[30] The human summariser is the author of this thesis, who has studied human, computer-aided and automatic summarisation, and has experience of both creating summaries from scratch and using a computer-aided summarisation tool, although not in a professional abstracting environment. For practical reasons a professional abstractor could not be used, and the author is considered to have sufficient knowledge of summarisation to act as an effective human summariser in this situation.

4.5.1 Texts

In order to perform the analysis of extracts and abstracts presented in Chapter 5 and Chapter 6, a corpus of these texts had to be created. The corpus consists of 43 pairs of annotated extracts and their corresponding abstracts, 86 texts comprising 866 sentences and almost 22,000 words in total. Table 1 shows the statistics for the corpus. The source texts used for the annotation of the extracts and the subsequent production of the abstracts are texts from *New Scientist*, extracted from the *British National Corpus* (*BNC*) (Burnard 1995). Although they all deal with some aspect of science, the way they are written is very similar to news articles, and therefore the texts in the corpus are termed *news texts* (see Section 4.4.1 for a more detailed explanation). This is in keeping with the CAST corpus, containing both newswire texts from *Reuters* (Rose, Stevenson and Whitehead 2002) and popular science articles from *New Scientist*, which is used for the evaluation of human summary production operations in Chapter 7.

	Texts	**Sentences**	**Words**
Extracts	43	538	13,275
Abstracts	43	328	8,652
Total	86	866	21,927

Table 1: Corpus statistics

4.5.2 Annotation of extracts

Sentences from a full text considered suitable for inclusion in a summary are marked in one of three ways: *essential*, *important* or *referred*. *Essential* sentences are those containing the most important or relevant information relating to the main topic(s) of the text. *Important* sentences contain important or relevant information, but which is

not considered to be as crucial as that in sentences marked as essential. *Referred* sentences are those which do not contain enough important or relevant information to be marked as either essential or important, but contain a small piece of information necessary for the full understanding of an essential or important sentence. By annotating such information, it is possible to see exactly why a sentence which may not seem very relevant is included in the set of marked sentences. It is also useful for the human summariser in transforming extracts into abstracts because they can see immediately that not all of the sentence is relevant. This allows them to take the relevant information from the extract sentence and include it in a different abstract sentence, thereby saving valuable space and reducing redundant and irrelevant information.

In addition to containing information about the importance of the sentences, parts which can be *removed* from sentences marked as *essential* or *important* are also indicated. The idea behind marking these units is that annotators can remove as much redundant information as possible from important sentences, allowing more relevant information to be included in the summary before the compression rate is reached. In automatic summarisation, it is possible to extract clauses instead of full sentences. This can be better in terms of including the most relevant information in the least space, but problematic in terms of coherence because parts of sentences are presented as an extract. Giving the annotator the option to remove redundant parts of sentences means that the positive aspects of clause and sentence extraction can be exploited. It also means that during the production of the abstract from the extract, the human summariser does not have to perform as many deletion operations as they might otherwise have had to.

Another reason for retaining the option to mark *removed* segments is that the evaluation in Chapter 7 is carried out on texts from the 2003 annotation, and it is therefore necessary to keep the same annotations in all texts to ensure that the evaluation is fair and reliable. The annotation of this corpus was performed using the same set of guidelines employed in the 2003 annotation task for the CAST corpus (see Section 4.4.2) with several amendments regarding spatial and temporal adjuncts, speech and sub-headings, based on an assessment of the guidelines. Section 4.4.4 discusses the assessment and resulting amendments in more detail. In addition, the examples used to illustrate instructions were changed to ones more appropriate for the science-based news texts in this corpus. These adapted guidelines can be found in Appendix I.

4.5.3 Production of abstracts

The second part of the corpus comprises abstracts corresponding to the extracts produced by annotating the source documents for the information described above. A human summariser took the extracts rather than source texts as the starting point and produced abstracts by making them more concise, coherent and readable. This step is essential for the investigation in Chapter 5 and Chapter 6 because it simulates the actions of a human summariser using a computer-aided summarisation system to produce abstracts. No annotations are made on existing texts to produce the abstracts; the human summariser simply copies the information they need from the extract, and adds to, removes or rephrases this information as necessary to change it into an abstract. No guidelines regarding the production of abstracts from extracts of news

texts could be issued due to their non-existence. Guidelines are not necessary at this stage for producing abstracts, because the purpose of the corpus investigation is to identify precisely what a human summariser does intuitively to reach a coherent and readable abstract which takes an extract and not a source text as the starting point. Guidelines for creating abstracts from extracts (summary production guidelines) are a novel contribution to the field of computer-aided summarisation (CAS) and are developed as a result of the classification based on the corpus analysis described in Chapter 5 and Chapter 6.

The only instruction given was that the abstracts should be 20% of the full source text. The CAST environment (see Section 3.5.4) was used to create the abstracts because it measures the compression rate as the summary is produced, allowing the human summariser to adhere to the 20% length restriction. It was decided to reduce the length restriction to 20% as it was noticed during the annotation of extracts in the CAST corpus that 30% seemed quite long for a summary of the news texts. Coupled with this is the fact that humans have much more creative means of shortening texts other than simply removing either whole sentences or parts of sentences from them, as is the case with shortening extracts. These means are discussed in Chapter 6.

Referring back to context factors (see Section 2.3), creating an extract from a source text must take *input*, *purpose* and *output* factors into account, whereas producing an abstract from an extract mainly relates to *output* factors only, as these are most concerned with the surface realisation of the final summary. However, during summary production, the purpose factor *relation to source* is particularly important as this is concerned with whether a summary is an extract or an abstract, and the fact

that it is an abstract means that its surface realisation will necessarily differ from the source, and, in this case, the extract from which it was produced.

4.6 Conclusions

The main purpose of this chapter was to present the corpus of extracts and abstracts analysed for human summary production operations in the following two chapters, describing the texts used and the annotation performed (Section 4.5). To achieve this, it was necessary to present a set of annotation guidelines which facilitate consistency when human annotators mark important information in texts, allowing suitable extracts to be created (Section 4.4). These guidelines were developed for the annotation of important information in texts to build the CAST corpus in 2003 (Hasler, Orasan and Mitkov 2003), and an analysis of them proved their suitability for a similar task in this research, providing certain amendments were made to them (Section 4.4.4).

To justify the development and use of guidelines to annotate a corpus of news texts for important information, as well as to contextualise them in the framework of the research undertaken in this thesis, Endres-Niggemeyer (1998)'s three stages of summarisation were revisited and more specifically related to the computer-aided summarisation of news texts and the guidelines necessary to achieve this task (Section 4.3). The development of the guidelines was further justified by a review of existing guidance for professional summarisers (Section 4.2), which showed that there are no guidelines available for the human extraction of news texts. The annotation guidelines are crucial for the research in later chapters, as they are

necessary to produce high quality extracts that are the basis for the corpus analysis, being used as the starting point from which to produce abstracts. Relevant findings from an analysis of the 2003 annotated corpus were also discussed and related to the work in the following chapters of this thesis (Section 4.4.3).

The annotation of extracts (Section 4.5.2) corresponds to the first two stages of the human summarisation process, document exploration and relevance assessment, and therefore to the automatic processes in computer-aided summarisation which result in an extract being presented to the user of a system. The production of abstracts by editing these extracts to make them more concise, readable and coherent (Section 4.5.3) corresponds to the third human summarisation stage of summary production. In terms of computer-aided summarisation, it is equivalent to the user of a system post-editing the automatically-produced extract presented to them. Because the focus of this thesis is to investigate how computer-aided summarisation can be improved by analysing and applying human summary production operations, it is imperative to distinguish extracts and abstracts in the corpus. The purpose of producing abstracts is to simulate the third step of human summarisation within computer-aided summarisation to allow the analysis and subsequent classification of human summary production operations.

The next chapter exploits the corpus developed in this chapter and introduces the classification of human summary production operations based on the corpus analysis, focusing on the most basic operations identified.

Chapter 5. From extracts to abstracts: atomic summary production operations

5.1 Overview

Chapter 2 discussed the three general stages of the human summarisation process (Endres-Niggemeyer 1998), amongst them the third stage of summary production, which involves the summariser cutting and pasting from the source text in order to create a summary. However, when humans cut and paste this material, they do not always use the text straight from the source without any refinement, but perform different operations on it to make it more concise and coherent and to ensure that it contains only the most relevant information in the most appropriate form. Chapter 4 related the three stages to the computer-aided summarisation of news texts and highlighted the lack of guidelines for human summarisers involved in such a task. This justified the importance of an investigation into the summary production operations applied to extracts to produce abstracts, which would allow guidelines to be developed to facilitate consistency in the task. This chapter addresses the next step when human summarisation is taken as a basic model for computer-aided summarisation, by proposing a classification of human summary production operations.

Because this thesis is concerned with the application of human-style operations to extracts to transform them into abstracts and improve their readability and coherence,

the focus of the chapter is on the operations a human summariser uses to get from an extract to an abstract. This means that the product of the first two summarisation stages (*document exploration* and *relevance assessment*), i.e., the extract, is taken as the starting point, and the final, edited summary (the abstract) is taken as the end product, and the analysis is of what happens in between.[31] By analysing the operations a human summariser uses, an insight can be gained into how to produce high quality summaries in the field of computer-aided summarisation. It should be pointed out that this thesis is not concerned with attempting to imitate *human understanding* for any kind of automated summarisation, but with showing that the coherence and readability of summaries produced with the help of a computer-aided tool and then edited by a human summariser can be improved by analysing and subsequently applying what humans do.

Section 5.2 describes existing research on the summary production stage of summarisation, looking at work based on operations present in human abstracts (Jing and McKeown 1999; Jing and McKeown 2000; Chuah 2001a; Chuah 2001b) or operations which human summarisers apply to reach an edited abstract (Cremmins 1996; Endres-Niggemeyer 1998), to contextualise the analysis. Section 5.3 introduces the corpus analysis and classification of human summary production operations. *Atomic* summary production operations are detailed in Sections 5.4 and 5.5, along with examples and a discussion. The chapter finishes with conclusions.

---

[31] It has already been pointed out in Section 4.4 that for the purposes of this investigation, it is better to use human-produced extracts rather than automatic ones in order to allow the focus to be solely on the *summary production* stage. However, it is only for this corpus analysis that the extracts need to be produced by a human annotator; in future, and indeed in the evaluation in Chapter 7, the extracts can be either human-produced or automatically produced. As this research is conducted within the field of computer-aided summarisation, it is more likely that the extracts will in future be automatically produced.

5.2 Operations to transform source text into summary text

This section reviews existing work which deals with the specific operations summarisers employ to transform material extracted from the source document into material suitable for inclusion in a summary. The review covers the fields of both human and automatic summarisation. All of the work examined analyses human-produced texts and discusses the operations found there, regardless of whether it is done to enhance knowledge and produce advice for professional abstractors (Cremmins 1996), or with a view to implementing an algorithm for fully automatic summarisation (Jing and McKeown 1999; Jing and McKeown 2000). However, in contrast to the original contribution of this thesis, the existing work described here focuses on source to summary operations rather than extract to abstract operations.

5.2.1 Jing and McKeown

Jing and McKeown (1999) performed a manual analysis of more than 120 summary sentences in 15 human-written summaries of newspaper articles and identified six major *cut and paste operations* used by abstractors to transform the source text into a suitable abstract: *sentence reduction*, *sentence combination*, *syntactic transformation*, *lexical paraphrasing*, *generalization/specialization* and *sentence reordering*. From the analysis of the 15 summaries, two rules are generalised: 1) humans are more likely to cut phrases than single words from a source text for the summary, and 2) humans are more likely to combine nearby source sentences into a single summary sentence than source sentences which are further apart. These rules were mainly assumed to aid the automatic decomposition of summary sentences, but

they are useful in a manual context too. Jing and McKeown found that the operations are not often used on their own, but are combined by the summariser for maximum summary effect.

They describe *sentence reduction* as a common technique, where humans remove less important information from a sentence they have selected from the source. The removed elements can be of any granularity: word, phrase or clause, and more than one element at once can be removed from a sentence. *Sentence combination* occurs when the summariser takes important information from more than one source sentence to make one single summary sentence. *Syntactic transformation* can occur in both of the previous operations, examples being the reversal of word order and a subject moved to a different position in the sentence. The fourth operation is *lexical paraphrasing*, where the human summariser replaces certain words or phrases in the source sentence with a paraphrase, for example, to avoid repetition of a noun phrase. *Generalization/specialization* is the fifth operation, occurring when humans replace words, phrases or clauses with less and more detailed descriptions respectively. The final operation is *sentence reordering*, which involves changing the order of the source sentences in the summary. Similar operations were observed in the corpus analysed in this thesis.

Jing and McKeown used this information as a basis for the development of their summary sentence decomposition program. Their results show that 78% of sentences in 300 human-written summaries of newspaper articles could be matched with sentences in the source document and are therefore based on cut and paste operations. The corpus analysis presented in this chapter indicates an even stronger

preference for editing original sentences rather than writing summary sentences from scratch. Of Jing and McKeown's 78% of sentences, over half match back to more than one source sentence, meaning that sentence combination is very useful.

5.2.2 Chuah

In two related papers, Chuah (2001a; 2001b) performs a linguistic analysis of sentences in 57 scientific abstracts and their corresponding source text sentences to explore the kinds of operations used to reduce information for presentation in the summary. Chuah (2001a) discusses what can be deleted and compressed in abstracting. Linguistic units which are commonly deleted from source text sentences to form abstract sentences include illocution markers containing first person pronouns, connectives, parenthetical text, apposed text and repetitions. Chuah's illocution markers are similar to what other researchers term *cue* or *indicating phrases* in summarisation (see Section 3.4.2), and these are usually taken as good indicators of relevant information in both human and automatic summarisation. However, the reasons for their deletion in Chuah's corpus may be because they signal rather than encode relevant information. As in Jing and McKeown (1999; 2000)'s work (see above), deletion can be at any level of granularity. Compression is very briefly presented as an alternative to deletion and the focus is on the simplification of complex linguistic units. Chuah draws the conclusion that whilst deletion is often a first step in the condensation of information for inclusion in a summary, even deletion on its own can "significantly abridge a text without critical loss in core content." (Chuah 2001a: 344).

Chuah (2001b) deals with the aggregation of important information from different sentences in the source into one summary sentence, again for scientific articles. It was found, similar to Jing and McKeown's assumption that humans will usually combine nearby sentences, that most sentences were aggregated from the same section of scientific articles. Chuah splits the aggregations into three categories: *conflation* (no explicit textual aggregation marker), *connective* and *semi-colon*. 75% of two-sentence aggregations were formed by conflation, which can be achieved by *splicing and joining*, i.e., cutting sections from different sentences and joining them together, or *merging* semantically equivalent text units. The use of connectives depends on the semantic relation between the text elements to be aggregated, and includes conjunctions and adverbials. In a quarter of cases where connectives or semi-colons are used for aggregation, the two are used together. Although texts from different domains were analysed, Chuah's findings generally support those of Jing and McKeown (1999) who looked at newspaper articles, suggesting that similar operations may be used across different domains to transform source text material into appropriate summary text.

5.2.3 Cremmins

The transformations discussed above can be linked to what Cremmins (1996) terms *revision*. Although the concept is not exactly the same, the operations can occur during the revision of abstracts. According to Cremmins, these operations or revisions can appear locally and globally in human summaries, both during cutting and pasting for summary production and afterwards in his fourth stage of summarisation (see Section 2.4.3), where the rough abstract is edited. Global

revisions take place across sentences and therefore include those operations such as *sentence compression* (or *aggregation*) and *sentence reduction* (or *deletion*) described above. He gives the examples of a sentence being aggregated so it becomes a modifier of a noun phrase in another sentence, thus combining sentences and shortening the overall summary text. Local revisions operate only within a sentence and therefore cannot include sentence aggregation/compression. Reference adjustment, lexical substitutions and the dropping of redundant, vague and superfluous terms are all considered local revision operations. Revision also involves correcting typographic, spelling, grammatical and other errors as well as rephrasing and rearranging text to maximise coherence and conciseness, making the summary easier to read. Again, more than one revision operation can affect the same unit of text.

5.2.4 Endres-Niggemeyer

Endres-Niggemeyer (1998) describes some individual summarisation strategies used by professional summarisers. As mentioned in Section 2.4.2, the most interesting strategies in terms of this thesis are those which directly affect the surface form of the final summary after the important information has been identified, i.e., those falling into the third stage of summarisation, summary production. The strategies in Endres-Niggemeyer's sub-category of *information presentation* under the broader heading of *professional skills* are of most relevance here, as they are concerned with the presentation of information in an abstract. Endres-Niggemeyer lists 102 separate information presentation strategies, divided into 16 groups. The strategies listed

below are those most pertinent to the analysis in this thesis, and come under the heading *Formulating*:

1. *Acquiring ready-made formulations* (e.g. *acronym*: Use well-known abbreviations, *con-title*: Integrate the document title, or part of it, in the abstract, *pattern*: Use standard patterns of formulation, *ready-made*: Use ready-made passages from the original)
2. *Integrating text modules* (e.g. *connect*: Connect individual statements to compose texts, *reorganize*: Reorganize text passages to make them fit in a new context)
3. *Elaborating style* (e.g. *emphasize*: Highlight important information by rhetorical reinforcement, *no-rep*: Avoid repetitions, *style*: Write good style)
4. *Writing typical abstract style* (e.g. *active*: Use active sentences where possible, *direct*: Be as direct as possible in your expression, *no-clumsy*: Avoid clumsy constructions or units, *nominal*: Nominalize expressions for use in lists, *precise*: Express yourself as precisely as possible, *present*: Use the present tense where possible, *short-sentence*: Build short sentences, *third-person*: Use the third person in your statements)

(Adapted from Endres-Niggemeyer 1998: 284-286)

5.2.5 Comments on previous work

The above review highlights the sparseness of research similar to that addressed in this thesis and the limitations of that research. On the one hand, the operations observed by some researchers are given as general categories and although they offer examples, there is no explicit classification describing sub-operations or typical

forms which operations usually take. The function[32] of the operations or the units to which they are applied is also not addressed in depth. On the other hand, when a more detailed description of operations is offered, far fewer general operations are covered. As well as this, the applications of operations can be so specific and summariser-dependent that it is difficult to assess how useful they would be if translated into a set of guidelines. In addition, existing work is concerned with operations used when transforming the *source text* into a summary and the type of text investigated is usually different to that analysed here. In terms of automatic summarisation, the operations identified in the corpus are related to the various *surface rejection rules* such as eliminating text in brackets, subordinate clauses and quotation marks as described in Section 3.4.6. The *problems and possible solutions* discussed in Section 3.4.7, for example, the presence of dangling anaphors and the necessity of either ignoring sentences containing anaphors or adding the sentence containing their antecedents, are also of relevance. These issues are addressed during the classification. Also related to the above discussion of work on the summary production stage are some aspects of the guidelines for professional summarisers, particularly regarding style, detailed in Section 4.2.

5.3 Corpus analysis: from extracts to abstracts

This section introduces the corpus analysis carried out in order to identify human summary production operations used to transform extracts into abstracts. The corpus consists of 43 (extract, abstract) pairs, with 866 sentences in total. In the remainder

[32] In this thesis, the *function* of operations and the units of text to which they are applied is frequently referred to. It should be noted that this use of *function* is the most general sense of the word rather than any specific linguistic meaning such as *syntactic function* unless otherwise stated.

of this chapter, and in the next, a detailed classification of the operations observed is presented. Five general classes of operation are identified: *deletion*, *insertion*, *replacement*, *reordering* and *merging*, each comprising a number of sub-operations. This classification is defined in Section 5.3.1. The analysis and classification is vital for computer-aided summarisation because it enables the development of guidelines based on the transformations a human summariser utilises, which can then be used to help human summarisers consistently edit the output of a computer-aided system such as CAST. The operations identified as being applied by humans to transform relevant information into acceptable summary text are important when human summarisation is considered as a model for computer-aided summarisation in the current automatic summarisation climate, because they enable the progression from an extract to an abstract.

The classification presented in this chapter and in Chapter 6 focuses on a *qualitative* rather than a quantitative analysis. The classes identified are those which are deemed most useful to the summary production stage of summarisation. Although some transformations were not classified based on their very low frequency (for example, one or two occurrences), frequency alone should not be taken as the defining factor in the classification.

Because the operations classified below are concerned with the summary production stage of summarisation, the most relevant context factors (see Section 2.3) are output factors, although these depend to a large extent on purpose factors (Tucker 1999). In terms of differences from the source, which in the case of this analysis, are extracts rather than the original full source text, there are three output factors which are of

most importance. The form factor *style* deals with the linguistic realisations in the summary, important in this analysis because the transformation of extracts to abstracts means that in virtually all pairs in the corpus there are some differences between the linguistic realisation in the extract and that in the abstract.[33] The form factor *scale* is also relevant because it controls the compression of the summary, in this case 20% of the original full source text, and 66.6% of the extract which was taken as the starting point, or source, for this investigation. The output factor *subject matter* is also important because abstracts can include background information which is not present in the extract. This was not very widespread in the corpus because the human summariser tried not to introduce any completely new information, assuming that, as an extract was taken as the starting point, all the relevant information would already be included. However, clauses or phrases are occasionally inserted from scratch, suggesting that a change (or the addition) of subject matter is a possibility. These units are also inserted to emphasise information already present in the abstract. In terms of the influence of purpose factors on the corpus analysis, the overriding one is Orasan (2006)'s *relation to source*, i.e., the fact that the final summary in this case needed to be an abstract. The output factors affecting the abstract are based mainly on this purpose factor.

5.3.1 Classification of human summary production operations

The five operation classes identified as a result of the corpus analysis, *deletion*, *insertion*, *replacement*, *reordering*, and *merging*, are divided into two types: *atomic*

[33] There was only one instance of the (extract, abstract) pair being identical in the corpus. This was an extremely short text where the human summariser viewed the extract as a suitable abstract, which suggests that summarisers only make changes to the source text when necessary.

operations and *complex* operations. This distinction is discussed further below, but the basic idea is that atomic operations cannot be broken down into further general classes of operation, whereas complex operations can. That is, complex operations consist of atomic operations. The atomic operations identified in the corpus are *deletion* and *insertion*. As well as being operations which are applied in their own right to extracts, they make up the complex operations *replacement*, *reordering* and *merging*.

The operation classes defined here each encapsulate a range of sub-operations, which are determined by certain surface forms acting as *triggers* by which the operation can be identified. The reason for classifying the operations in terms of recognisable triggers is the possible future implementation of these operations, or aspects of them, in a computer-aided summarisation system. In addition, time is often of the essence in summarisation, meaning that easily recognisable forms which the summariser can quickly identify are useful. The classification may not seem the most logical from a grammar point of view, in that the classes could have been organised differently, but this classification is determined within the area of computer-aided summarisation and so operations are ordered by recognisable surface forms, or triggers, where possible, even though these may not always fit exactly into well-defined grammatical classes. The function of the operations and the text they involve is vital, as is evident from the subsequent discussion. However, it must be remembered that it is impractical to work with functions alone without any means of reliably identifying the units which typically embody them.

Atomic operations

Two classes of operations are considered atomic: *deletion* and *insertion*. *Deletion* is defined in this context as *the process of removing a unit*[34] *from a certain place in the extract so that it does not appear in the same place in the abstract*. It includes the deletion of subordinate clauses, adverb phrases, parts of noun phrases and punctuation, amongst other things.

Insertion is defined in this context as *the process of adding a unit which is not present in the extract into the abstract*. It has fewer sub-operations than deletion, some of these being the insertion of connectives and modifiers. Atomic operations are discussed in Sections 5.4 and 5.5, in the remainder of this chapter.

Complex operations

Three classes of operations are considered complex: *replacement*, *reordering* and *merging*. *Replacement* is defined in this context as *the deletion of one unit and the insertion of a different unit in the same place in the text*. It comprises the atomic operations deletion and insertion and includes sub-operations such as pronominalisation, restructuring of noun phrases, passivisation and verbal changes.

Reordering is defined in this context as *the deletion of a unit from one place in the extract and its insertion in a different place in the abstract*. This operation differs from the others because there are no recognisable surface triggers by which to

[34] For the purpose of the discussion of summary production operations, the term *unit* includes words, phrases, clauses, sentences and punctuation.

identify it. Instead, the function is taken as the factor that splits reordering into the sub-functions of emphasising information and increasing coherence, both of which include the atomic operations of deletion and insertion.

Merging is defined in this context as *taking information from different units in the extract and presenting it as one unit in the abstract*. Like the other complex operations, it comprises the atomic operations deletion and insertion, but it can be further classified as often including elements of replacement and reordering. There are two main sub-operations of merging identified: restructuring and punctuation/ connectives. Complex operations are discussed in Chapter 6.

5.3.2 General observations

Before commencing a detailed discussion of the classification of operations, it is necessary to make some general remarks about the operations identified during the corpus analysis. It is important at this point to remember that the analysis was conducted on extracts and abstracts which are not the same length, and that the human summariser worked with the instruction to produce 20% abstracts from 30% extracts (see Section 4.5.3). This meant that the abstracts had to be further reduced by one third, to 20% of the source text length. Whatever the class of operation, or any of its recognisable surface forms which constitute sub-operations in the classification, they all work towards shortening the text. Whilst the human summariser has the opportunity to make an abstract more coherent and readable than its extract, with more scope for changing the surface realisations of concepts and

ideas, it still needs to be shortened. This is a major reason for the application of the operations, in addition to the desire to rephrase, reorder and combine information.

As well as shortening the text, the operations are generally used to avoid repetition and strengthen coherence and readability. The text type of the documents in the corpus also affects the sub-operations identified, with examples of all the general classes of operations being used to make the abstract sound more like a conventional news article. Also related to the text type, but more to do with the specific domain (popular science), is the observation that there are fewer operations applied to more specialised or technical texts. One reason for this is that the human summariser was not as confident in changing texts with a higher level of technicality because they were not a subject expert and did not want to risk changing the meaning of the text when creating the abstract.

Similar to the work reported in Section 5.2, the corpus analysis proved that operations are often combined as this is most effective in creating an abstract. The corpus provides many examples of 'real abstraction', where different operations combine to produce a unit in the abstract, often taking information or ideas from more than one sentence in the extract and presenting them together in one sentence in the abstract, with a different surface realisation. This provides evidence that, given the option, humans really do perform *abstraction* when editing extracts, and do not only make small-scale revisions which are viewed by some as not resulting in significant enough transformations to produce an abstract (see Mani (2001)). This is especially important given that computer-aided summarisation developed out of the

field of automatic summarisation, where the editing of summaries is either non-existent or limited to small-scale revision operations.

5.3.3 Atomic human summary production operations

The remaining sections of this chapter present in detail the classification of the atomic human summary production operations mentioned in Section 5.3.1 above. These atomic operations are considered to be the two most basic classes of operations identified in the corpus. Section 5.4 discusses *deletion* and its sub-operations, and Section 5.5 deals with *insertion*. The sub-operations identified within the deletion and insertion operations are contradictory at times. There are several examples of the insertion and deletion of the same types of surface form. As mentioned above, the function of the unit is important, but in these cases it becomes absolutely crucial if an explanation is to be attempted as to why very similar units are both deleted and inserted. This is particularly necessary when the atomic operations performed individually are examined; when they are used in combination within a complex operation it seems less strange. In the classification presented below, the examples aim to show only the sub-operation under consideration at any one time unless otherwise stated. This makes some examples very different from others containing the same extract text. In particular, because atomic operations are often applied to the same unit as complex operations, the examples below sometimes form only part of a wider picture. The example of an (extract, abstract) pair in Section 6.7 attempts to show all operations applied to units at once. A selection of (extract, abstract) pairs can be found in Appendix II.

5.4 Deletion

The first class of operations to be discussed is *deletion*. This is one of the easiest and most basic operations to categorise, as it does not involve any reformulation of the extracted text, although it does frequently occur in combination with other operations which do. In the context of the corpus analysis presented in this chapter, *deletion* is defined as *the process of removing a unit from a certain place in the extract so that it does not appear in the same place in the abstract*.

A distinction is made between *permanent* and *non-permanent* deletion operations, where *permanent* deletions are those in which the information does not appear again in the abstract at all and *non-permanent* deletions are those whose information does appear in the abstract although not necessarily in the same form. These non-permanent deletions only happen as part of complex operations, for example, the deletion of a sentence from the middle of the extract and its insertion at the end of the abstract. For the sake of simplicity, the discussion of the deletion operations in this section is restricted to the permanent deletion of units. Cases of deletion as part of a complex operation (non-permanent deletion) are considered under the appropriate headings in Chapter 6. The reason for using the deletion operation alone, as an atomic operation, is relatively simple: to shorten the text. As mentioned above, the human summariser produced abstracts which were shorter than the length of their corresponding extracts. However, when this operation is used in combination with others to form a complex operation, its function is also to ensure grammaticality and coherence.

The class of *deletion* operations corresponds to Jing and McKeown (1999)'s and Cremmins (1996)'s *sentence reduction* and Chuah (2001a)'s *deletion*, although Chuah focuses on illocution markers rather than other types of deletion. It also plays a part in Cremmins' local revision operations, in terms of the dropping of redundant, vague and superfluous terms. It is difficult to find equivalent strategies observed by Endres-Niggemeyer (1998) because the strategies are very specific and are not necessarily related to recognisable forms. There are no related strategies under her *information presentation* heading. There are several strategies which could be related to deletion in her *information acquisition* category, but this group relates to selecting information from the source based on its relevance to the summary and so is not considered appropriate for discussion here. Similar to the existing work described in Section 5.2, deletion can occur at any level of granularity: word, phrase, clause, sentence.

In terms of revision in automatic extracting, Nanba and Okumura (2000) propose the deletion of dangling anaphors and connectives, and extraneous adverbial particles, whilst Mani, Gates and Bloedorn (1999) use sentence *compaction* which involves deleting parentheticals, sentence-initial prepositional phrases and certain adverbial phrases. Although they are not strictly deletion, but rather the prevention of selecting units for extraction which could be problematic in terms of coherence or wasting compression, rejection (and selection) rules proposed for automatic summarisation systems are still relevant here. Johnson et al. (1993)'s rejection rules for anaphoric references and connectives, Brandow, Mitze and Rau (1995)'s exclusion of sentences containing anaphors, and Mitkov, Le Roux and Descles (1994)'s elimination rules for examples, bracketed text, subordinate clauses and text within quotation marks are

all related to the *deletion* operation identified in this corpus. With reference to the problems and solutions and the selection and rejection rules in automatic extraction, the classification below proves that the deletion of units such as subordinate clauses, prepositional phrases, adverb phrases, bracketed text, and direct speech or text in quotation marks cannot be done indiscriminately, solely based on the surface realisation. The forms which trigger the *deletion* operation have certain functions and it is the function rather than the form which ultimately determines when it can be safely deleted. However, as mentioned in Section 5.3.1, it is necessary to identify recognisable forms because this work is carried out in the field of computer-aided summarisation.

It should be pointed out here, to avoid unnecessary repetition in the following discussion, that there are *always* cases where the sub-operations described below are not applied. There are often more cases of non-deletion than deletion. This can be explained by the fact that part of the transformation process involved further shortening of the text to a given compression, meaning that some material had to be deleted. However, it is not feasible to delete all possible units, in other words to apply the sub-operation in all cases, partly because some units are necessary for the understanding of the text, and partly because to remove all of them would result in an abstract well below the specified compression rate. In this analysis, compression is important because of the context of computer-aided summarisation where automatically produced extracts of a pre-specified length are presented to the user for editing.

Ten sub-operations of the deletion operation were identified: complete sentences, subordinate clauses, prepositional phrases, adverb phrases, reporting clauses and speech, noun phrases, determiners, the verb *be*, specially formatted text, and punctuation. Each of these sub-operations is dealt with in turn below, and examples are given to illustrate them. Because the examples are necessarily taken out of context, it may seem that some information is lost during deletion operations. However, this is not the case. The information appears in other sentences in the abstract, which is why the sub-operations can be safely applied. ~~Strikethrough~~ indicates deleted units in the examples.

5.4.1 Complete sentences

The most obvious case of deletion is where a complete sentence which appears in the extract is not included in the abstract. In the corpus description (Section 4.5), the corpus annotation was discussed, stating that human annotators marked sentences as *essential* and *important*. The majority of deleted sentences in the corpus are those annotated as important rather than essential. This is hardly surprising as essential sentences are considered to contain more relevant information than important sentences, therefore it seems sensible not to delete this information when there is a better option. In total, 107 complete sentences were permanently deleted: 100 important sentences (including 4 *question* sentences) and 7 essential sentences. The DELETE: SENTENCE operation functions mainly to remove information which the original annotator deemed suitable for inclusion but which was not considered to be the *most* important information from the source. Important sentences often emphasise or support information given in essential sentences, providing more detail. In the

example below, an *important* sentence is deleted because it adds information which is not as relevant as that in other retained sentences:

> *A team led by Dr David Jones, of Bangor's renowned School of Ocean Sciences, designed a revolutionary new feed.* ~~*After their initial breakthrough, the team still had much to do before the process became commercially viable.*~~ *Dr Jones and his team are now extending their work to other species, and to the post-larval stage…* (sci02done-an)

In cases of the deletion of essential sentences, the human summariser's personal opinion about which information is more important to keep has more influence. This highlights the challenges that the subjective nature of 'importance' poses to the task of summarisation. The example below shows a deleted *essential* sentence. In this case, the annotator considered the sentence to be of equal importance to the one preceding it in the extract but not in the abstract, perhaps because it provides additional detail for the first sentence and seems to function as an example in the type of information it adds:

> *According to one theory, held by some psychiatrists, patients may find doctors increasingly unsympathetic to their plight.* ~~*They may be neurotics or malingerers or both.*~~ (sci09done-ljh)

A sub-type of DELETE: SENTENCE is *question* sentences. There are 4 instances of permanently deleted *question* sentences, all of which were labelled *important*, in the corpus. There is also one which was non-permanently deleted, where the *answer* sentence is retained in the abstract and rephrased in a concise way to include the

necessary information for the question sentence. The following example shows an instance of permanent question sentence deletion:

> *Britain certainly has various nuclear weapons there. ~~Can we be sure they will never be used?~~ Units of the Fleet, sailing south from Gibraltar on Monday, 29 March, 1982, were carrying nuclear weapons.* (new-sci-B7L-70-ljh)

In this case, the question sentence conveys speculation, not something that is a definite fact (compare it to the surrounding sentences). It is not answered in the form of a corresponding *answer* sentence in the extract, further strengthening the argument for speculation. As it is, speculation about Britain's use of nuclear weapons is partly what the text is about, but the question is used to emphasise rather than add new relevant information; this question can be inferred from the rest of the extract without having to be explicitly stated.

5.4.2 Subordinate clauses

In summarisation, concepts introduced in subordinate clauses are often not considered to be as important as those in main clauses, therefore making them more suitable for deletion in order to save space. These units are easily recognisable most often in this corpus because of the presence of a subordinator. A subordinate clause is one which is embedded in another clause, or is a constituent of another clause, or of a phrase in the case of a relative clause (Quirk et al. 1985: 44). Quirk et al. (1985) list five formal indicators of subordination, which are useful in this discussion as

they relate to sub-operation triggers. Only those found most frequently during the corpus analysis are given here:

1. The clause is initiated by a subordinating conjunction
2. The clause is initiated by a *wh*-element
3. The verb element of the clause is either nonfinite or absent

(Adapted from Quirk et al. 1985: 997)

The fact that particular information is presented in a form signalling subordination means that, as well as being relatively easy to recognise, it is easier for the summariser to remove these parts of sentences without having to make any further changes to the remaining clause than if the main clause was deleted. However, the major reason for the deletion of subordinate clauses, as with other types of permanently deleted unit, is to reduce the text by removing non-essential information. The deletions were evenly split between relative clauses modifying noun phrases (NPs) (introduced by *which*, *who*, *that*) and clauses introduced by subordinators (*because*, *for*, *as*, *if*, *with*, *while*). It should be noted that in the case of relative clauses, only those whose NP is retained are considered to be cases of DELETE: SUB_CLAUSE; if the whole NP including the relative clause is deleted, it is considered as NP deletion (see below). There were only two examples of a non-finite subordinate clause signalled solely by the lack of a finite verb, and in one of these cases only part of the clause was deleted. Other deleted non-finite clauses were also introduced by a subordinator, and as this is easier to recognise, they are considered in terms of their subordinators. Some examples of the permanent DELETE: SUB_CLAUSE operation applied to clauses which give too much detail, or whose content can be inferred from elsewhere in the abstract, are:

Three papers published recently in Science move us a little closer to understanding the basis of the disease~~, which turns out to be highly complex~~. (sci04done-an)

~~While many politicians have only recently seen their green light,~~ many diesel engine manufacturers have looked further ahead… (sci06done-an)

This research, to enable identification of a released GEM – and its whereabouts – is essential ~~because of the risks involved in letting GEMs out of the laboratory~~. (sci11done-ljh)

There are 15 definite cases of permanent subordinate clause deletion in the corpus. It was extremely difficult to classify a number of possible cases because in the process of producing the abstracts, chunks of text which cannot be classified as a clause, sentence or other discrete unit, are deleted. This again emphasises the need to reduce the text, and as this kind of 'chaotic' deletion aids the complex operations of *replacement*, *reordering* and *merging*, further strengthens the argument that simple revision operations are not enough to produce a human-style abstract. In this analysis, only clearly defined subordinate clauses are considered under the heading of deleted subordinate clauses.

Interestingly, a small number of *coordinate clauses* using the coordinating conjunctions (or coordinators) *and* and *but* are also removed, but too few to merit a sub-operation. This occurs when the second coordinate clause elaborates information in the first and is therefore not needed as it increases redundancy or gives too much detail. Four cases of main clause deletion in a complex sentence were also observed,

but again they were too few to class as a sub-operation. This was initially surprising, but further examination again emphasised the importance of the function of the text or sub-operation applied. These main clauses are deleted when two or more sentences are merged and the information in them is not needed as it is taken from elsewhere. There is an example of a main clause (and a few examples of subordinate clauses) being deleted, along with other units (marked with [square brackets] in the example below), to form a 'headline' from the first sentence of extracts. These are considered permanent deletion because they are not used in combination with *insertion* to function as another transformation operation. This is also a prime example of context factors in use, illustrating how Tucker (1999)'s form output factors *style* and *structure* affect the abstract (see Section 2.3.3) due to genre conventions:

> ~~*Water sports on reservoirs are threatened*~~ [*as*] *more evidence emerges of* [*the*] *blue-green algae~~, which feed on pollution~~* [*and were*] *first seen in Britain this summer.* (sci08done-ljh)

5.4.3 Prepositional phrases

The third sub-operation of deletion in the corpus is DELETE: PREP_PHRASE. A prepositional phrase (PP) comprises a preposition and a prepositional complement, usually a noun phrase, and can carry the syntactic function postmodifier, adverbial or complement (Quirk et al. 1985: 657). The functions of most relevance are postmodifiers and adverbials because these can be deleted without necessarily changing the meaning of the sentence. Prepositional phrases are deleted from the extracts as they are transformed into abstracts because they add small pieces of

information often present elsewhere, or which can be inferred from the rest of the text by the reader. There are 25 instances of prepositional phrase deletion, 11 of which are postmodifiers referring to either a physical object immediately preceding them or a previously-mentioned entity. Keeping them in the abstract repeats information and, as in other cases of deleted units, wastes compression. The preposition appearing immediately after a noun and being a constituent of a noun phrase is the trigger by which the full PP is recognised, allowing it to be identified as a potential candidate for deletion. Examples of the DELETE: PREP_PHRASE operation from the corpus are:

> *The four volumes ~~of this work~~ are soon to be accompanied by a further five…* (new-sci-B7L-63-ljh)

> *DEFRA told WWT samplers to moisten a sterile swab ~~on a stick~~ with saline…* (h02-ljh)

There are 14 examples of deleted prepositional phrases functioning as adverbials, which can be split into adjuncts referring to time and space (11 cases) and conjuncts used to explicitly aid coherence (3 cases). Similar to postmodifiers, temporal and spatial adjuncts are often deleted because they mention again an entity previously mentioned in the text. PPs should only be deleted when they repeat information given elsewhere or when the information they contain can be reliably inferred from other parts of the abstract. The following examples contain deleted prepositional phrases functioning as spatial adjuncts:

H5N1 was most likely carried to the UK by migratory ducks, which could have spread the virus ~~to wintering grounds all over the country~~. (h02-ljh)

~~In a report this week,~~ Ian Fairlie and David Sumner, two independent radiation scientists [*from the UK*]*, say that the death toll will in fact lie somewhere between 30,000 and 60,000.* (h03-ljh)

Prepositional phrases whose syntactic function is a conjunct are deleted from abstract sentences much less frequently than other PPs. This deletion is interesting because they are used to explicitly indicate coherence, and it is widely accepted that human produced abstracts are more coherent than either computer or human produced extracts. In the task of transforming an extract into an abstract, these conjuncts could be expected to be retained. However, the automatic summarisation literature refers to *dangling connectives* and *discourse ruptures* (see Section 3.4.7), which are equivalent to these kinds of conjunct, as well as to certain adverbs. A frequently cited example is the problem caused by the extraction of only one of the discourse connectives *on the one hand* or *on the other hand* (and their alternative realisations) and the necessity of deleting the remaining one to improve coherence. Although there was only one case of each of these in the corpus, and although they both appeared within a sentence of each other, both conjuncts were permanently deleted. The function of this operation applied to these particular units is to rephrase the sentence as part of the *replacement* and *merging* operations; the corresponding abstract sentences are somewhat different to their extract counterparts. But because these conjuncts are removed and do not appear in any other form or position in the abstracts, despite the fact that the deletion of other parts of the sentences are non-permanent, the deletion of the conjuncts is permanent:

> ~~*On one hand*~~ *McElroy is preparing for private companies to take over the reigns of the weather craft;* ~~*on the other,*~~ *he is inviting other countries to become involved in what could be a link only between governments.* (new-sci-B7L-1-ljh)

Conjuncts are also sometimes added to abstract sentences (see Section 5.5), and not all conjuncts which signal coherence are deleted. They are only deleted when there is too much rephrasing of the sentence to keep them, or when the human summariser needs to find 'neat' units which they can easily remove to further shorten the text. As with other units discussed in this section, prepositional phrases are not deleted if they are central to the meaning of the noun phrase, where time and place are important, or they are necessary to avoid ambiguity.

5.4.4 Adverb phrases

Although the deletion of adverb phrases (including single-word adverbs) is not as widespread as PP deletion, they are similar in that they can be recognised fairly easily and are therefore worth mentioning in the context of computer-aided summarisation. Similar to subordinate clauses and prepositional phrases, they contain a recognisable trigger, in this case an adverb. An adverb is a word "which expresses any relation of place, time, circumstance, causality, manner, or degree… a word that modifies or qualifies an adjective, verb, or other adverb" (*The Oxford English Dictionary Second Edition* 1989, Volume I: 118). In terms of this analysis, for simplicity and due to the nature of the transformation process, the heading adverb phrase includes adverbs, units introduced by adverbs and phrases which have

adverbs as the head. There are only 9 adverb phrases deleted in total, and deletions occur when they are used to modify information, adding further detail which is unnecessary for the summary and to emphasise information. These phrases can be deleted without affecting the coherence of the resulting abstract, and therefore can be safely removed from extracts during the production of abstracts. Again, the reason behind the deletion of adverb phrases is the fact that the information in them is not as crucial as the information contained in other textual units, meaning that keeping them in the abstract will use up valuable space. Examples of the sub-operation DELETE: ADVERB from the corpus are:

> *Potential and actual savings which have been identified include: shorter lead times, ~~presently as long as six months;~~…* (new-sci-B7L-54-ljh)

> *In a report this week, Ian Fairlie and David Sumner, two independent radiation scientists from the UK , say that the death toll will in fact lie ~~somewhere~~ between 30,000 and 60,000.* (h03-ljh)

There are many instances of adverb phrases, and particularly single-word adverbs, which are retained in the abstract because they convey information essential for avoiding ambiguity, or because they are obligatory; the clause or sentence does not make full sense if they are removed. As with other deletion sub-operations, the compression rate plays a part in identifying units which can be 'safely' deleted when producing an abstract from an extract.

5.4.5 Reporting clauses and speech

This type of deletion is concerned with the removal of verbs and other elements which describe or report speech or findings. It was decided to extend DELETE: REPORT to include findings or the reporting of information, rather than just speech, because the structures of reporting in both cases are very similar. In cases of indirect speech or reporting, the conjunction *that* (or occasionally a preposition such as *to* or *on*) following the reporting verb is also deleted unless it is needed to ensure the grammatical correctness in instances of merging; for simplicity, these units including *that* are termed *reporting clauses*. There are 8 cases of the deletion of reporting clauses attached to indirect speech and reporting. As discussed by Hasler, Orasan and Mitkov (2003) and in Section 4.4, whilst the actual (reported) speech or the gist of what was said may be important, who reported it is not necessarily so. The indirect reporting clauses deleted in this corpus are deleted because they introduce irrelevant speakers, or those which have been previously mentioned, thereby using up valuable compression. They also add information which can be inferred from elsewhere in the text. One example highlights the influence of news text style on the production of abstracts, with the reporting clause being removed to leave a headline-style unit (the first example below). Some examples of DELETE: REPORT are:

> ~~*Tim Radford reports on*~~ *the prospects and options for farmers in the next year* (sci14done-ljh)

> ~~*Fairbanks found that*~~ *when sea level first began to rise as the ice sheets melted, 17,000 years ago, it did so at a rate of about 4mm per year.* (sci16done-ljh)

> ~~*I suspect that*~~ *the set would be the ideal book for a physicist to be cast away with on a desert island.* (new-sci-B7L-54-ljh)

Reporting clauses similar to those described above are not suitable for deletion when it is important that the reader knows who to attribute the speech or claim to. As mentioned in Section 4.4.4, this is particularly important where there are conflicting opinions regarding the main topic of the summary. The following example contains two reporting clauses which it was necessary to retain so that the reader could reliably understand that the writer of the source text and other scientists did not necessarily agree with what Zuccarelli was saying. The retained reporting clauses appear in **bold**:

> ***According to Zuccarelli****, the ear generates a reference signal…* ***He claims that*** *his dummy head does likewise.* (new-sci-B7K-37)

5.4.6 Noun phrases

As well as the deletion of relative clauses and prepositional phrases modifying nouns, other modifiers such as adjectives and nouns, or even whole noun phrases, are deleted. For the sake of simplicity, this sub-operation is termed DELETE: NP, although it is also concerned with the deletion of *parts of* noun phrases, such as modifiers. The deletion of modifiers, which add more detail to noun phrases, serves to shorten the text and reduce redundancy in the abstract, i.e., it makes the abstract more concise. Parts of noun phrases which are not needed to disambiguate the head because this has been done earlier in the text can be safely deleted, as can information which can be inferred from earlier in the document. Sometimes, whole noun phrases appear to

further describe and add detail to another which is present in the text, or are repeated close to another mention of the same entity and therefore are repetitive. These can also be deleted. In addition, they are deleted as the first step of depersonalising a sentence, which is then further transformed using the *replacement* (and sometimes *merging*) operation. Cases such as these are discussed in more detail in Chapter 6, and can be linked to DELETE: REPORT, in the sense that it also depersonalises sentences. This is perhaps an indication of a wider pattern in summaries, where mentions of people or organisations are deleted unless they are closely related to the main topic. Due to the rareness of pronouns (2 cases) being permanently deleted, they are also included in DELETE: NP.

In total, 43 instances of this sub-operation were observed, 11 adjectives and 32 noun phrases or parts of noun phrases. However, these units are so widespread in the corpus that many more are not deleted than are, because they are necessary for avoiding ambiguity, or provide new relevant information. In the first example below, *Clarks* can be safely deleted because prior to this mention, it is clear that the only program discussed so far in the text is the Clarks program, meaning that this part of the NP is not needed to disambiguate the program in this sentence from any other possible program. The fact that only one word is deleted in some cases again highlights the need to shorten the text by any means possible to arrive at the final abstract. Examples from the corpus are:

> *Pattern flattening is done by additional mathematics on a specially written part of the ~~Clarks~~ program.* (new-sci-B7L-54-ljh)

It was ~~a polymer~~ so unlike the polymers around at the time that no one could envisage a use for it. (new-sci-B7L-74-ljh)

It hit ~~the coast at the town of~~ Innisfail at 0700 ~~Eastern Standard Time (AEST) on~~ Monday. (e02-ljh)

The reason that such repetitions appear in the extracts is that the same sentences in the source may be much further away from each other and therefore it is more appropriate to clarify noun phrases more often to avoid confusion over longer distances. However, summaries tend to be about one (or two) main topics, meaning that sentences from the source containing references to this topic are brought together as one text. The references appear much closer together in the summary, so there is less ambiguity. This means that the parts of noun phrases which appeared in the extract because they were necessary for disambiguation in the source can be safely deleted during the production of the abstract.

5.4.7 Determiners

A *determiner* is a limiting expression which modifies a noun or noun phrase, for example *the*, *an*, *every*.[35] The sub-operation DELETE: DETERMINER is concerned with the deletion of determiners, usually definite articles, and has the main function of shortening the text without impeding readability. There are 11 cases of deleted determiners, 9 definite and 2 indefinite. The removal of short, single-word units such as these highlights the pressing nature of text reduction as the summariser seems willing to use any means possible to achieve it. It has the additional function of

[35] This definition is adapted from *The Oxford English Dictionary Second Edition* (1989, Volume IV: 551).

transforming the first sentence of the abstract into a conventional 'headline' when applied there and is also used when a noun is mentioned for the first time due to the deletion of other units which previously introduced it.[36] This sub-operation is dependent on the individual text being summarised at the time in a different way to the others discussed so far, in that it is used as a 'last resort' to reach the compression rate and some texts are easier than others to reduce by a further third during the abstracting process without needing to do this. Some examples from one sentence in the corpus are:

> ~~*The*~~ *scientists' work confirmed that the undoubted improvement in parachuting ability was a fortuitous result of the development of* ~~*the*~~ *fringes for* ~~*these*~~ *other purposes.* (sci12done-ljh)

5.4.8 The verb *be*

The sub-operation DELETE: BE is concerned with the deletion of the verb *be* from constructions such as *are removed* and *is coloured*. This sub-operation is not very common (3 cases in the corpus) but, similar to *determiners* discussed above, the deletion of such small units again emphasises the importance of shortening a text during summarisation and depends on the individual text. As with DELETE: DETERMINER, this type of deletion also functions to transform the first sentence of a news text into a 'headline', which is of interest in this particular corpus due to its text type. In fact, the two together could be taken as a simple formula by which human summarisers working with news texts can easily create a 'headline' for their

[36] This use of DELETE: DETERMINER is equivalent to Mani, Gates and Bloedorn (1999)'s *indefinitization*.

summary based on the first sentence of a document if one is not already included. Other deletion operations may also be applied. For example:

> *Britain ~~is~~ among the front runners as tomorrow's supercomputers take shape.* (sci05done-an)

5.4.9 Specially formatted text

The sub-operation DELETE: FORMAT covers the deletion of text which appears in a certain format, such as within or following certain punctuation. Text within brackets, text following a colon or a semi-colon, and text in between or following dashes is included in this sub-operation. However, permanent deletions are relatively few considering the number of instances, with only 10 cases observed. It was predicted that there would be very few cases of specially formatted text in the extracts because of the guidelines used to produce them. The fact that not only was this type of unit present in the extract when the annotator was instructed not to include it, but that it was also retained in the abstract, indicates that rejection rules used in automatic summarisation (see Section 3.4.6), for the news texts in this corpus at least, are not necessarily appropriate. Despite this, specially formatted text is still an easily recognisable unit which can be safely deleted in certain circumstances.

The main function of bracketed text is to introduce acronyms or expansions of acronyms which are used later in the text, or information that the writer considered less important than the surrounding text such as additional detail. There are only 5 permanent deletions of bracketed text. Where acronyms are not retained in the

abstract, the text and the brackets surrounding them can be safely deleted. An example of bracketed text which is deleted from the corpus is:

> *Over the same distance, people use two and a half times more energy ~~(per kilogram of their bulk)~~ than the average camel.* (sci15done-ljh)

Text following a colon, semi-colon or dash often provides more detail about the text appearing before the punctuation, which usually 'summarises' the more detailed description or clarification which follows. However, there are only 5 permanent deletions, again meaning that not all instances can be safely deleted. These units are deleted when the text following the punctuation adds much more detail or clarification than is necessary in a summary. They are also unsuitable for deletion when the text does not need to be reduced further and the whole sentence is coherent and relevant. An example from the corpus is:

> *Modern science is critically dependent on high-performance computing~~; studies of the world's changing climate, structural engineering, and medical imaging simply could not have progressed to their present state without access to the sort of computing power that can only be provided by parallel machines~~.* (sci05done-an)

5.4.10 Punctuation

Although it may not seem as important as other deletions, and does not concern the text of the extract or abstract as such, DELETE: PUNCTUATION is worth including as a separate sub-operation. There are 21 cases where commas are deleted from sentences in the corpus, as well as less frequent instances (10 in total) of semi-colons, colons,

full stops, quotation marks and ellipses. Reasons for this are that the punctuation in the extracted texts is not always grammatically accurate and the necessity of reducing the text as much as possible. In CAST, punctuation is counted in the compression,[37] explaining why very small units such as commas may be deleted. Users of computer-aided summarisation systems need to be made aware of how compression is calculated. Deleting punctuation can be a useful way of reducing the text to as near the stated compression rate as possible, but it may not apply in systems other than CAST.

It was predicted that the atomic operations involving punctuation, especially commas, would be very much subject to individual style preferences; however, a small-scale experiment showed that this is not always the case. Six human judges were shown pairs of sentences (or groups of sentences) with different punctuation and asked which they preferred. In 50% of cases, there was full agreement between all six judges. 25% of the time there was full disagreement (three selected one sentence and three the other), and there was some disagreement (one judge selected a different sentence to the other five) in 25% of cases. The DELETE: PUNCTUATION sub-operation does not deal with punctuation which is deleted and replaced in sentences during *merging*; this is discussed in Section 6.4.

5.4.11 Comparison with the 2003 annotation

It is useful at this point to compare deleted units from this corpus with removed sentence segments from the CAST corpus presented in Section 4.5, which was used

[37] It would also be possible to calculate compression based on words or sentences in a computer-aided summarisation system, as opposed to characters which are used in CAST.

to validate the annotation guidelines for the corpus in this thesis. It has been mentioned earlier (Sections 4.4.2 and 4.4.3) that although the *deletion* operation identified in the classification in this chapter and the *removal* of segments during the annotation of the 2003 corpus belong to different stages of the summarisation process, they both include the deletion of parts of sentences to adhere to a certain compression rate. Despite the differences in the task and the texts, most of the categories identified for removal in the CAST corpus have equivalents in the deletion sub-operations presented above. However, the boundaries are different because of the focus on the function of the *removed* text in the 2003 corpus and the form, or triggers, of the *deleted* units here. A comparison shows that similar types of information were deleted regardless of whether the aim was to create an extract of a source text or an abstract from an extract.

The 2003 annotation resulted in 7 categories of removed segments, and because the aim of the annotation was not geared towards computer-aided summarisation as such, the categories are defined in terms of their function rather than recognisable forms. Segments containing information referring to time and location, realised by prepositional phrases and adverb phrases functioning as adjuncts, correspond to the DELETE: PREP_PHRASE and DELETE: ADVERB sub-operations. The second category contained reporting clauses, similar to the DELETE: REPORT operation here. Subordinate clauses were another category of removed segments, which correspond to DELETE: SUB_CLAUSE, although they were less restrictive about what kinds of information were conveyed by them. The category of relative clauses qualifying nominal phrases is also covered by DELETE: SUB_CLAUSE. Examples are given as a category based on the 2003 annotations, but these are introduced by adverbs and

therefore if they appear, are covered by the DELETE: ADVERB sub-operation in this corpus. The sixth category of removed segments was text in between or following dashes. In the current corpus analysis, such units do not merit a sub-operation of their own as they are too infrequent. However, they are part of the DELETE: FORMAT sub-operation (also relatively infrequent) which deals with specially formatted text. The final category created from the 2003 annotation was very general: other ways of elaborating information, using adverbs functioning as connectives and prepositions. These do not correspond to one discrete sub-operation in the present analysis, but the removed segments would fall into the following classes: DELETE: ADVERB, DELETE: NP and DELETE: PREP_PHRASE.

In the 2003 annotation task, exceptions were identified which informed the extent of possible deletions in this corpus and resulted in amendments to the original guidelines. The first of these concerned speech and reporting clauses. As there was very little direct speech in the corpus analysed in this thesis, that is of no relevance here, but the fact that it is necessary to keep reporting clauses when the text is concerned with conflict, arguments and different viewpoints is reflected in both corpora. The second main exception was regarding dates and times, which were annotated in some cases in the CAST corpus. As the discussion of the DELETE: PREP_PHRASE and DELETE: ADVERB sub-operations above shows, it is not always acceptable to delete such information. In fact, there are exceptions in all of the *deletion* sub-operations, which highlights the necessity of giving the summariser the option not to delete a unit (the *unless vital to the main topic* clause added to most instructions in the guidelines) if they really do not think it is appropriate to do so. The analyses of both corpora demonstrate that these categories cannot *always* be

removed/deleted: they are context-dependent. They also show that even when sentences which have already been considered relevant (or important) need to be further reduced, similar operations are used to achieve this.

Although the human-produced extracts analysed in this thesis already have parts removed from sentences due to the way they were annotated, this will not be the case with automatically produced extracts. However, the similarities between the *deletion* operations in this corpus, and the *removed* segments in the CAST corpus mean that *deletion* can successfully be applied to automatically produced extracts to remove unnecessary information as well as improve their readability and coherence. Applying deletion operations to extracts during summary production will have the same effect as marking parts of them for removal during relevance assessment. In addition, the fact that these operations are so similar explains the low frequency of some of the deletion operations in the corpus investigation: units to which deletion operations are applied had already been removed during the extraction stage.

5.5 Insertion

The second class of operations observed in the corpus of extract and abstract pairs is *insertion*. As with the deletion class, this is also relatively easy to recognise and classify as a broad class, as it does not involve any reformulation of the extracted text and insertions can be reliably identified when comparing an extract and its corresponding abstract. However, as is obvious from the discussion of deletion, these atomic operations are not trivial to classify into sub-operations. In the context of the corpus analysis presented in this chapter, *insertion* is defined as *the process of*

adding a unit which is not present in the extract into the abstract. This includes units in the extract which are deleted and then added in another place or replaced by another unit in the abstract. Similar to the permanent/non-permanent distinction made in the class of deletion operations, a distinction needs to be made between insertions which are made *from scratch*, and those which are the *addition of a deleted unit* into the abstract in a different place or form. This section deals only with those units which are inserted from scratch, the other type being considered in Chapter 6 as they fall under the headings of *reordering* and *merging*. Insertion from scratch is a particularly interesting class because of the overall nature of summarisation as a text reduction process. In the discussions of human and automatic summarisation (Chapter 2 and Chapter 3), and of guidelines (Chapter 4), one of the aspects of summaries that is agreed on is that they should be shorter than the source text they summarise. Insertion from scratch is viewed as being solely concerned with readability and coherence, as no consideration is overtly given to space considerations or compression rate.

Jing and McKeown (1999; 2000) and Chuah (2001a), do not discuss the insertion of text into abstracts, presumably because they are concerned with summarisation as text reduction. Chuah (2001b) discusses insertion only in terms of aggregation, and only of connectives and semi-colons, which is covered in Section 6.4 (merging) in this classification. Similarly, Cremmins (1996) does not discuss insertion as a category of revisions, although some of his revision operations will include this operation, particularly the rephrasing and rearranging of text to maximise coherence and conciseness, and sentence aggregation. In terms of the work of Endres-Niggemeyer (1998), the strategies *connect* and *emphasize* under the headings of

Integrating text modules and *Elaborating style*, respectively, relate to the insertion from scratch operation detailed here. Other strategies could also include elements of insertion, but not necessarily insertion form scratch. As with other existing work, Endres-Niggemeyer does not explicitly address insertion as an operation. In terms of insertion as addressed in automatic summarisation, Nanba and Okumura (2000)'s shallow coherence smoothing includes adding conjunctions to 'rescue' dangling connectives as well as adding missing adverbial particles to improve coherence. Brandow, Mitze and Rau (1995) include the first sentence of a paragraph if the second or third is extracted in an attempt to minimise problems with gaps. In the corpus analysed here, connectives are inserted, although to improve the overall coherence of the abstract rather than to 'rescue' dangling connectives in the extract. There is no evidence of the insertion of extract sentences in the same way that Brandow, Mitze and Rau (1995) insert first paragraph sentences from the source if necessary, which suggests that gaps caused by the selection of non-consecutive sentences from the extracts are not an issue in this corpus.

Four sub-operations of insertion are identified in the corpus: connectives, formulaic units, modifiers and punctuation. Insertion occurs at word, phrase and clause level, but there are no instances of complete sentences being inserted from scratch. This strongly suggests that, whilst smaller units are inserted to improve coherence and highlight information, human summarisers would not waste compression to such an extent by inserting a whole sentence from scratch into an abstract produced with the help of a computer-aided summarisation tool. Each of the insertion sub-operations is dealt with in turn below, with examples where appropriate. **Bold text** indicates the inserted units in the examples.

5.5.1 Connectives

A connective is defined as an "item whose function is to link linguistic units, such as conjunctions and certain adverbs" (Crystal 2003: 460). Connectives are inserted into abstracts to make them more coherent by explicitly indicating relationships between units. This helps to improve the flow of the text as it is read, because the reader does not have to decide whether to infer such a relationship, and also helps to avoid ambiguity of interpretation. Whilst this sub-operation may appear broad because it includes more than one word class, it is suitable for the present analysis because all the occurrences function to connect textual units in one way or another and they have a relatively limited set of recognisable triggers. The sub-operation INSERT: CONNECTIVE includes conjunctions (coordinators and subordinators) and adverbs functioning as conjuncts; there are 26 instances in total. Connectives are also inserted as part of the *merging* operation (see Section 6.4.2).

Sentence-initial connectives are easiest to recognise in terms of computer-aided summarisation, as they are inserted immediately after a full stop. Therefore, sentence-initial and mid-sentence connectives are differentiated. Sentence-initial connectives are more frequently inserted, constituting 16 of the 26 insertions. *However* was the most common connective to be inserted (11 occurrences) with the function of emphasising a contradiction between one sentence and information in previous sentences. It is useful in texts as short as summaries to indicate contradictory information to the reader because a summary is usually about one topic or presents the same kind of information about the topic, and explicitly signalling

possibly unexpected information can help the reader to process it more easily. Other sentence-initial connectives inserted are *but*, *and*, and *additionally*. The following examples illustrate the insertion of *however* to strengthen the coherence of the abstract:

> *Potential and actual savings which have been identified include: shorter lead times, presently as long as six months; producing better, more up-to date styles; and shoes that fit better.* ***However*** *CAD had, and still has, its doubters at Clarks.* (new-sci-B7L-54-ljh)

> *He claims that his dummy head does likewise.* ***However,*** *it is unclear whether this theory is serious or a misunderstanding.* (new-sci-B7K-37)

The connective most frequently inserted mid-sentence is *also* (4 cases), functioning to draw the reader's attention to additional information which has not been previously mentioned and strengthening the argument created in the rest of the text. Other mid-sentence insertions included *actually*, *then*, *because*, *still* and *however*. The following example illustrates the insertion of *also*:

> *Governments of developing countries should certainly be making greater provision for agricultural research. But they* ***also*** *need to be reassured that in all countries results from research tend to be indirect and to take time before they come to fruition.* (new-sci-B7K-58)

The fact that the insertion of these connectives occurred in the corpus whilst their deletion was rare and quite random (and hence did not merit classification as a sub-

operation) again proves that coherence and readability really do matter in abstracts. The insertion of units into an abstract when the summariser is attempting to reduce the text to a specified compression rate suggests that the incorporation of the necessary operations and sub-operations dealing with coherence and readability into the computer-aided summarisation process is a must.

5.5.2 Formulaic units

In terms of summarisation, INSERT: FORMULAIC is the most interesting sub-operation because it most obviously illustrates a disregard for the basic concept of summarisation as text reduction. As well as being related to the formulation of headline-style units created using deletion, it is related to the sub-operation INSERT: CONNECTIVE because it functions to emphasise subsequent information and its relationship with other units. However, unlike connectives, the unit inserted in this case is larger, comprising at least three words including a verb. These units can be considered almost 'set' or 'formulaic' patterns or linguistic items, especially in news texts due to their 'reporting' nature. This again highlights the effect of context factors such as *style* in summarisation, and is linked to the standard sentence patterns identified by Endres-Niggemeyer (1998) as used in the summary production stage. Such large insertions show that explicit coherence and emphasis is an important issue because of the risk it poses to wasting (or at least using up) compression.

It could be argued that this sub-operation of insertion also fits into the *merging* and *replacement* operations (see Sections 6.4 and 6.2, respectively) because certain information is inferred from elsewhere in the text and presented in a different way.

However, it is dealt with under this heading because in these cases it is extremely difficult to match these units with any material explicitly stated in the extract and so is taken as being inserted from scratch.[38] These types of possible multiple interpretations emphasise the difficulty of formulating a completely clear-cut classification. Because INSERT: FORMULAIC occurs with other operations such as *reordering* or *replacement*, showing the insertion in the original extract sentences would result in ungrammatical sentences and so the examples given for this sub-operation are taken from abstracts rather than extracts. There were only 5 examples in the corpus, two of which combine to form a pattern, but it is worth mentioning because it was so unexpected. The first example also infers information from elsewhere in the extract to derive the formulaic pattern, highlighting the way the operations interact and illustrating that it is not easy to classify them. All the examples are shown here, and patterns are given in [square brackets]:

> *A chemist recognised its similarity to the material used to sheathe submarine telephone cables -* ***tests showed that*** *polythene was superior,* ***making large-scale production worthwhile.*** (new-sci-B7L-74-ljh) [X showed that Y, making Z worthwhile]

> ***He sees*** *the need to raise public awareness and demystify science and technology* ***as*** *a key point…* (new-sci-B7L-75-ljh) [X sees Y as Z]

> ***And we should bear in mind that*** *however we may feel about overseas aid, the requirement for agricultural research is a matter of tens of millions of dollars, not billions.* (new-sci-B7K-58)

[38] The fourth example in this section is also dealt with under *merging* in Section 6.4.1; see below.

> ***The research should give insights into*** *the way in which the atmosphere interacts with the oceans, which remains a last great unknown in Earth science.* (sci13done-ljh)

The last example above is not really 'formulaic', but it does carry the same function as the other examples: to draw attention to what follows. This example is also used in Section 6.4.1 (merging), where it is interpreted in terms of being inferred from other units in the extract rather than as insertion from scratch.

5.5.3 Modifiers

In addition to being deleted during the production of abstracts, modifiers are also inserted. The reasons behind this are similar to those which explain why some modifiers but not others are deleted. By introducing a modifier of a noun or noun phase, any ambiguity is dealt with in a concise unit rather than requiring a fuller explanation, for example in the form of an additional sentence. There are only 5 cases of the sub-operation INSERT: MODIFIER, because the necessary information is usually available in other sentences and is inserted as part of the *replacement*, *reordering* and *merging* operations. In the cases discussed here, however, modifiers are inserted from scratch to aid the understanding of the abstract. A sub-operation for these few instances is considered acceptable because the broader operation *insertion* is such an interesting case for summarisation. Because there are so few instances of this operation, NPs, PPs and adjectives are considered together. The example below shows the insertion of a noun phrase which was necessary because the extract does not mention that *Men of Science* is a television programme, although this could be inferred from the use of the verb *shown*. The human summariser felt it necessary to

clarify this so that there was no uncertainty for the reader as to what was actually being reviewed. The PP in the second example clarifies the reason that Sri Lanka may end up poorer.

> ***The TV series*** *Men of Science is now being shown in a few other areas.* (new-sci-B7L-69-ljh)
>
> *Sri Lanka seems likely to end up the poorer* ***from this programme****, with the poorest suffering most.* (new-sci-B7K-19)

5.5.4 Punctuation

Similar to modifiers, as well as being deleted during abstract production, punctuation (13 cases: usually commas, but also hyphens, colons and full stops) is also inserted in otherwise unchanged sentences to make the text easier to read. As punctuation is counted in the compression rate in CAST, the fact that the human summariser chose to insert anything which would use up valuable compression, especially punctuation which does not provide any information, means that they really felt it to be necessary. This sub-operation does not deal with punctuation which is deleted and replaced in sentences during *merging*; this is discussed in Section 6.4.

5.6 Conclusions

The purpose of this chapter was to present the first part of a classification of human summary production operations based on an analysis of the corpus described in Chapter 4. Section 5.2 offered a brief overview of related studies in order to contextualise the classification. Two types of operation classes identified in the

corpus, *atomic* and *complex*, were introduced in Section 5.3 and a detailed classification, with sub-operations and examples, of the atomic summary production operations *deletion* and *insertion* was presented in Sections 5.4 and 5.5, respectively.

Operations similar to those observed by other researchers (see Section 5.2) were identified in the corpus and classified accordingly. However, the classification presented here is much more detailed than existing work, listing both general operations and sub-operations which can be applied to specific units within extracts recognised by certain *triggers*, or surface forms. In addition, the operations identified in this corpus are split into *atomic* and *complex* operations, atomic operations constituting the complex ones. This chapter addressed only atomic operations to pave the way for the discussion of complex operations in the next chapter. This atomic-complex dichotomy means that it is not always easy to assign classes because, as well as being operations in their own right, the sub-operations are often combined to achieve a complex operation. Atomic operations are not applied in every possible case because this is impractical in terms of both compression and understanding, and would result in an abstract which is impossible to understand fully and is substantially below the specified compression rate.

Similar to Chuah (2001a), *deletion* (Section 5.4) was found to be a very useful operation as it reduced the text without losing relevant information, and could be used alone as well as being a first step in more complex operations. Whilst the focus in this chapter was on what can safely be removed without impeding (or even can be deleted to improve) the coherence and readability of the text, the fact that the abstract had to be one third shorter than the extract it was produced from meant that the

extract text had to be reduced without deleting the most relevant information. In contrast to the work reviewed in Section 5.2, *insertion* was addressed as an operation in its own right because it is concerned explicitly with coherence and readability. It is an interesting operation due to the fact that units are inserted from scratch in a text which, by definition, needs to be reduced, and therefore highlights the importance of coherence in abstracting.

In terms of *deletion*, it was found that the operations observed in the corpus investigated here do not always agree with some of the surface rejection rules and revisions used in the field of automatic summarisation, such as the rejection of specially formatted text and the need to 'rescue' dangling connectives. However, the analysis was consistent, for the main part, with the analysis of the 2003 annotation task (Hasler, Orasan and Mitkov 2003) discussed in Chapter 4, and it was possible to identify similarities between the types of unit deleted even though the task in hand was different. This suggests that these deletion sub-operations are important throughout the summarisation process and that similar types of deletion operations are applied in transforming extracts into abstracts as they are in creating an extract of a source text. This makes them particularly suitable for computer-aided summarisation.

The next chapter addresses the *complex* operations classified as a result of the corpus analysis and consisting of the atomic operations described in this chapter. It also discusses possible future implementations of both atomic and complex operations in a computer-aided summarisation system, and presents an example of an analysis of

an (extract, abstract) pair from the corpus and the summary production guidelines based on the classification.

Chapter 6. From extracts to abstracts: complex summary production operations and summary production guidelines

6.1 Overview

The previous chapter introduced a classification of human summary production operations based on an analysis of the corpus presented in Chapter 4, focusing on the most basic operations identified in the corpus: the atomic operations deletion and insertion. This chapter completes the classification by investigating the second type of operations identified: *complex* operations. They are discussed separately because they comprise the atomic operations, which are applied to achieve different ends to when they are applied alone. The complex operations identified in the corpus are *replacement*, *reordering* and *merging*, and are considered to be the operations which really capture the essence of abstracting as material from the extract is rephrased and rearranged to form a distinct text: the abstract. These operations are discussed in Sections 6.2, 6.3 and 6.4, respectively. After the classification of complex operations, an analysis of the sub-operations in terms of possible implementations is given in Section 6.5, proposing ways in which specific sub-operations could be automatically processed to aid summary production. This chapter also develops a set of guidelines for the summary production stage of summarisation based on the classification, to facilitate consistency of editing. The guidelines can be found in Section 6.6. In addition, an example of an (extract, abstract) pair from the corpus which illustrates

both atomic and complex operations using a variety of sub-operations is presented in Section 6.7. The purpose of this is to emphasise the extent to which human summary production operations are applied to extracts, and proves that it is often not just a case of applying simple revision operations. The chapter finishes with conclusions.

As with atomic operations, the complex operations discussed in this chapter subsume more specific sub-operations, or in the case of *reordering*, sub-functions. However, due to their more complex nature, it is more difficult to classify them using recognisable surface forms (triggers), especially in the case of reordering and merging. Replacement, reordering and merging best highlight the differences between extracts and abstracts, and therefore between summaries which can be produced automatically and manually. Similar to the classification in Chapter 5, it is impractical to apply complex operations in every possible instance.

6.2 Replacement

The first class of complex summary production operations to be discussed is *replacement*. This is the 'easiest' of the complex operations to deal with, because similar to atomic operations, the summariser works with units which usually (but not always) have recognisable triggers in the extracts and/or abstracts. However, because it is often used in combination with *reordering* and *merging*, it is not always easy to identify. The smaller the unit replaced, the easier it is to classify as an instance of the replacement operation. *Replacement* is defined in terms of deletion and insertion: *the deletion of one unit and the insertion of a different unit in the same place in the text*. The main reasons for using the replacement operation include shortening the text,

avoiding or minimising repetition, and changing units so that they appear more like those around them to improve readability. Replacement only occurs when the human summariser considers that units should be worded differently, and not in all possible cases.

This class of operations is broader than those discussed by Jing and McKeown (1999), covering their *lexical paraphrasing*, *generalization/specialization* and *syntactic transformation* operations. It also covers various local revision operations described by Cremmins (1996), such as reference adjustment, lexical substitutions and rephrasing of the text to make it more concise and coherent. Correcting errors would also count as replacement, as defined in this classification. The replacement operation is an element of Chuah (2001b)'s *aggregation*, which includes the merging of semantically equivalent units and their expression using synonyms, hypernyms, partial repetitions and metonyms. In terms of Endres-Niggemeyer (1998)'s strategies, those listed under *Writing typical abstract style*, for example, *nominal*, *present*, *active* would all involve the abstractor replacing units from the source with more suitable units for the abstract if those source units were not already in the forms stated. Strategies from other groups, such as *pattern* and *no-rep* would also involve replacement (see Section 5.2.4).

Mani, Gates and Bloedorn (1999)'s sentence *smoothing* deals with *reference adjustment*, for example, expanding pronouns, and *coordination reduction* which involves simplifying coordinate constituents. These are only applied intra-sententially, whereas this classification accounts for cases of inter-sentential replacement due to its combination with reordering and merging. Nanba and

Okumura (2000) advocate the replacement of dangling anaphors with their antecedents if these appear in the sentence previous to that containing the anaphor. Additionally, pronominalisation and the addition of demonstratives are given as possible ways of overcoming the problem of repetition. (Paice 1981)'s inclusion of sentences preceding extracted sentences which contain anaphoric references is related to one particular aspect of replacement (*referred* sentences). Pollock and Zamora (1975)'s rules concerning spelling and abbreviations also fall into this class.

Eight sub-operations of the replacement operation were identified in the corpus: pronominalisation, lexical substitution, restructuring of noun phrases, nominalisation, referred sentences, verb phrases, passivisation, and abbreviations. Each of these sub-operations is dealt with in turn below, and examples are given to illustrate them. As with the examples of other operations, only the sub-operation under consideration at any one time is illustrated, unless otherwise stated. Some examples may seem ambiguous, but this is because they are necessarily taken out of context to highlight a point. In their full context, there is no ambiguity unless it was present in the source. ~~Strikethrough~~ is used to show deleted units and **BOLD CAPS** to show insertions, which together constitute replacement.

6.2.1 Pronominalisation

The first sub-operation of replacement addressed is REPLACE: PRONOMINALISE, which involves deleting a noun phrase in an extract and inserting a pronoun in its place in the abstract. This sub-operation functions both to avoid repetition and to shorten the text. In total, there are 44 cases of pronominalisation of noun phrases or parts of noun

phrases (determiners). Seven of these concern the replacement of a determiner with a demonstrative or possessive pronoun. In 34 cases, whole nouns or noun phrases are replaced with pronouns. Four of these involve demonstrative pronouns replacing NPs. Replacement by demonstratives, particularly of determiners, does not necessarily shorten the text, but functions to increase the coherence of the abstract by strengthening links between units and the information contained in them. In the case of whole NPs replaced by demonstrative pronouns, the text can be shortened considerably, as the first example shows:

> *The Countryside Commission has just announced an imaginative scheme for 12 "community forests" on marginal urban fringe farmland. But ~~the community forests - plus the Commission's other plan for 100,000 acres of new forest in the Midlands, which will also soak up redundant land - together with new land released for development,~~* **THIS** *will barely scratch the surface of the problem.* (sci14done-ljh)

Other examples of the sub-operation REPLACE: PRONOMINALISE are given below. These highlight the function of using pronominalisation to avoid repetition, usually of a noun phrase which plays a large part in describing the main topic of the text. Whilst this means that the reader of the abstract has to resolve the pronoun and assign the correct antecedent to it, it gives the text a much more readable style than it would have if the same complete NP was repeated throughout. Pronominalisation also usually functions to shorten the text, as virtually all pronouns are shorter than the NPs they replace.

If the Nimslo 3D system, like all previous 3D snapshot and movie systems, fails to catch on, then investors who have put up around $30 million will catch a cold. Even if ~~3D snapshooting~~ **IT** *does catch on there may still be no gravy train for ~~Nimslo's~~* **ITS** *backers.* (new-sci-B7L-55-ljh)

Perrin subsequently took charge of the high-pressure research… On 19 December, 1935, ~~Perrin~~ **HE** *produced several grams of polythene.* (new-sci-B7L-74-ljh)

In a report this week, Ian Fairlie and David Sumner, two independent radiation scientists from the UK, say that the death toll will in fact lie somewhere between 30,000 and 60,000. ~~Fairlie and Sumner's~~ **THEIR** *accusations are backed by other experts.* (h03-ljh)

Three cases of pronominalisation could also be considered as part of the *merging* operation between two sentences. After the replacement of the full stop with a comma (usually), the noun phrase which starts the second merged sentence is replaced by the relative pronoun *which*. The remainder of the second sentence is then restructured accordingly to ensure grammaticality (see Section 6.2.6). It is difficult to decide whether *which*-replacement should be considered as replacement or merging. However, as pronominalisation is considered to be a sub-operation within the replacement class of operations, and *which* is a relative pronoun, it is taken to be a case of REPLACE: PRONOMINALISE. Whilst the function of replacing an NP with *which* in these cases may be to merge units, it remains that the sub-operation used to achieve this is replacement, and although cases of punctuation and connectives for merging are considered to be a sub-operation of merging, *which*-replacement

involves much more because it includes some transformation of a noun phrase and needs to be resolved during reading. For example:

> *The Mahaweli Programme is a $2000 million scheme to dam Sri Lanka's 300 kilometre Mahaweli river~~. The programme~~* ***WHICH*** *aims to make the country self-sufficient in food and to cut unemployment and energy imports.* (new-sci-B7K-19)

In contrast to pronominalisation, there are also instances of pronoun expansion, although not enough to constitute a sub-operation of their own. During pronoun expansion, a pronoun is replaced by the noun phrase it refers to, and this type of replacement most often occurs in order to avoid repetition when pronouns are used in very close proximity to one another. In one case, the pronoun was expanded because the NP it referred to had not been previously introduced and therefore the abstract would not have made complete sense if the pronoun was not replaced. This aspect of pronoun expansion is related to **referred sentences**, discussed in Section 6.2.5.

6.2.2 Lexical substitution

This sub-operation is related to pronominalisation (see above), but it is kept separate due to the widespread use of both, and the different forms the resulting unit takes. Lexical substitution encompasses more than NPs alone, providing another reason for two separate sub-operations. REPLACE: LEXEME involves the replacement of one lexical item by another which refers to the same entity, concept or action in the text, or something related to that entity, concept or action by means of synonymy, hyponymy, hypernymy or metonymy. The replaced unit in the abstract can be more

or less precise than its corresponding unit in the extract, as the examples below show. In this corpus, REPLACE: LEXEME is applied to noun phrases, verbs, adjectives and adverbs, but the most obvious cases are those involving NPs. In total, 71 occurrences of this sub-operation were observed, and its main function is to avoid repetition so that the abstract reads better. A secondary function is to shorten the text during abstract production, but this does not always happen. The first example below shows the replacement of a noun with a much longer, more descriptive and specific NP which enables the summariser to convey as much information as possible in that NP without having to resort to inserting another sentence to describe it. This does function to reduce the text overall, although it may not appear to do so when only this sentence is considered. In addition, the nature of the noun in the extract assumes that the reader knows who or what *Brandt* is, although this may not be the case, which means that lexical substitution in this case also clarifies the entity referred to:

> *Yet, the answer to the drought and related problems lies not with the worthy, necessary and spatchcocked aid to Ethiopia but with a positive response to ~~Brandt~~* **THE BRANDT COMMISSION'S SECOND REPORT, COMMON CRISIS, NORTH AND SOUTH, CO-OPERATION FOR WORLD RECOVERY**. (new-sci-B7K-58)

The following example illustrates the replacement of the name of a person working for or representing an organisation by the name of the organisation itself. The reason for this is that in the context of this particular text, it is a report by the World Health

Organisation (WHO)[39] which is under scrutiny from others, and it is therefore simpler to refer to the organisation than to its spokesperson. The reader is much more likely to recognise *the WHO* than *Zhanat Carr*, so replacement in this case simplifies the text and at the same time shortens it. The replacement of *says* with *admits* is an example of the output factor *style* on the abstracting process, replacing a relatively neutral word with one typical of news reporting to suggest that the WHO was in the wrong.

> ~~*Zhanat Carr, a radiation scientist with the WHO in Geneva,*~~ **THE WHO** ~~*says*~~ **ADMITS** *the 5000 deaths were omitted because the report was a "political communication tool".* (h03-ljh)

Other cases of this sub-operation serve to shorten the text by generalising the NP, as in the first example below, and to avoid repetition, illustrated by the second example. The avoidance of repetition is the simplest reason for using REPLACE: LEXEME.

> *By 1984* ~~*those who sift through your files in police stations, hospitals and local authority housing departments*~~ **AUTHORITIES** *could be* ~~*talking with*~~ **USING** *artificial intelligence.* (new-sci-B7L-41-ljh)

> *The historical development of quantum theory by Jagdish Mehra and Helmut Rechenburg, Springer, vols 1-4… If* ~~*Jagdish Mehra and Helmut Rechenburg*~~ **THE AUTHORS** *succeed in maintaining the standard set so far…* (new-sciB7L-63-ljh)

---

[39] Interestingly, this abbreviation is also discussed in Section 6.2.8, where it is argued that such common abbreviations can be 'safely' left in abstracts as the reader will know exactly which organisation is referred to. However, the need to expand it here, in a PhD thesis, illustrates the fact that the text type and its 'main topic' are important when dealing with lexical substitution.

6.2.3 Restructuring of noun phrases

The sub-operation REPLACE: RESTRUCTURE_NP involves the restructuring of noun phrases during the production of abstracts in order to shorten the unit by transforming postmodifiers into premodifiers which usually contain fewer words. Twelve cases of this sub-operation were identified. The most common case involves prepositional phrases as postmodifiers, but there are two cases each of relative clauses and adverb phrases, which also involve lexical substitution. Prepositions are deleted from PPs postmodifying nouns and the remaining part of the PP is relocated in front of the original NP head. Similarly, with relative clauses postmodifying nouns, the relative pronoun following the NP head is deleted and the rest of the relative clause is replaced before the head. In both cases, other transformations may be necessary to ensure grammaticality. A MODIFIER+HEAD structure is preferred to HEAD+MODIFIER in abstracts because it makes the text more 'snappy' by shortening it. This sub-operation can be linked to the deletion of relative clauses and prepositional phrases, as well as to the deletion of parts of noun phrases discussed above (Section 5.4). Some examples from the corpus are:

> *Coming at a time of deep economic recession, the drought has affected ~~the whole economy of Australia~~* **AUSTRALIA'S WHOLE ECONOMY.** (new-sci-B7K-42)

> *Photopia denies that the evident haste is the result of poor ~~sales in the US~~* **US SALES.** (new-sci-B7L-5-ljh)

> *Potential and actual savings which have been identified include: shorter lead times, presently as long as six months; producing better, more up-to date styles; and ~~shoes that fit better~~* **BETTER-FITTING SHOES.** (new-sci-B7L-5-ljh)

Several of the sub-operations described in Chapter 5 and Chapter 6 relate to noun phrases and prepositional phrases. The observation of the sub-operation REPLACE: RESTRUCTURE_NP suggests that if these units cannot be deleted or pronominalised or substituted by a more appropriate lexical item, for example, they can at least be shortened in an attempt to pack the necessary information into the smallest possible space. This again illustrates the importance of text reduction throughout the summarisation process. There are numerous cases when this sub-operation is not applied, because of the high frequency of noun phrases in the corpus and the fact that it is not necessary or desirable to restructure every single one.

6.2.4 Nominalisation

REPLACE: NOMINALISE is another relatively infrequent but interesting and useful sub-operation, which occurs 7 times in the corpus. The use of this sub-operation is in keeping with the *nominalize* strategy used by Endres-Niggemeyer (1998)'s expert abstractors, and also fits into the revision operations discussed by Cremmins (1996) (see Section 5.2). The function of REPLACE: NOMINALISE is to improve readability and to shorten the text by expressing the information in fewer words, and it occurs in combination with *merging* because it can replace a whole clause or sentence but does not form a grammatical sentence on its own. The two examples below illustrate this sub-operation along with the new sentence merged using the nominalisation:

All this is hardly Culver's fault. [*The same difficulties are to be found in all other parts of evolutionary ecology.*] → ***THESE GENERAL DIFFICULTIES OF EVOLUTIONARY ECOLOGY*** *are hardly Culver's fault.* (new-sci-B7L-63-ljh)

The algaes were very unpredictable. [*One week the algae could be clear of toxins and the next week very toxic but no one yet knows why.*] → ***THE TOXICITY OF THE ALGAES*** *were very unpredictable but no one yet knows why.* (sci08done-ljh)[40]

6.2.5 Referred sentences

Referred sentences are marked in the corpus because they are not *essential* or *important* themselves, but they contain some information which makes them necessary for the full understanding of an *essential*/*important* sentence (see Section 4.5.2). These sentences are difficult to classify, even in terms of general operations, because whilst they may appear to have been permanently deleted, it must be remembered that they were included in the extract for a reason. In the extracts, the only reason for including referred sentences is to introduce this information necessary for the understanding of other sentences. In the corresponding abstract, this information can be added (or *inserted*) in the sentence which needs it, present in both the extract and the abstract. The replacement operation which achieves this uses both of the atomic operations deletion and insertion. Therefore the sentence cannot truly be said to have been permanently deleted, because a small part of it, perhaps just one

[40] This sentence is not grammatically correct: *were* should be *was*, however, the human summariser did not correct it during summary production, and so it remains in any examples containing the sentence.

word, is also present in the abstract. Consequently, cases such as these are considered to belong to the replacement and merging operations, where the information needed to resolve a pronoun, for example, is used to replace that pronoun in an abstract sentence, thus merging information from two distinct units. There are four instances of referred sentences in the corpus, and also several cases of *essential* (3) and *important* (5) sentences which behave in the same way. Because referred sentences are a novelty of this particular corpus, it is not possible to specify a sub-operation which would always be suitable for abstracting tasks using other texts. In another corpus, this could be dealt with via pronoun expansion. A similar idea to the human annotation of referred sentences is proposed in the field of automatic summarisation by Paice (1981), who advocates including additional source sentences in a summary to combat the problem of dangling anaphors. An example of REPLACE: REFERRED, where information from a referred sentence replaces a pronoun in a different sentence in the abstract, is:

> ~~*You could be forgiven for not having heard of the Treaty of Tlatelolco.*~~ *The non-aligned nations have just dragged* ~~*it*~~ **THE TREATY OF TLATELOLCO** *on the stage by backing an Argentine call for the withdrawal of all nuclear weapons from the Falklands.* (new-sci-B7L-70-ljh)

6.2.6 Verb phrases

REPLACE: VP is concerned with changes to verb phrases during the production of abstracts. This sub-operation deals with verb changes made necessary by other sub-operations, for example, certain deletions, to ensure grammaticality. It also addresses the simplification of units by transforming verb phrases to make the text easier

and/or quicker to read. At this stage, it becomes increasingly difficult to distinguish sub-operations due to the fact that several are applied to the same unit in the text.

Most of the guidelines for professional abstractors discussed in Chapter 4 make some reference to the use of a preferred tense: the present. However, extract sentences are not always in the present tense and the human summariser did not always change them accordingly. In cases where the human summariser did change the tense of verbs in sentences, they changed it to the simple present or past, which served to shorten the text and to simplify the verb phrase, making the abstract quicker and easier to read. Bearing in mind the importance of further reducing the text during summary production, in some cases a tense is changed to another mainly because it takes up less space. In these cases, lexical substitution also sometimes occurred in combination with the verb change.

The general rule which can be gleaned from the human summariser's application of REPLACE: VP is to prefer the simple present or past tense where possible, but to *always* transform verb phrases in order to ensure grammaticality. This is particularly important because when other sub-operations of replacement, as well as merging, reordering, deletion and sometimes insertion occur, it is necessary to change the verb accordingly. The fact that verbs need to be changed to cooperate with other sub-operations is the main reason for verbal transformations in the corpus. The pronominalisation in the first example below necessitates a change in order to have number agreement with the pronoun, whilst the second example shows a tense change to the simple past to save space and simplify the structure. This happens

because although the source text is a narrative recalling the writer's school days and the tense used there adds to the atmosphere, it is not necessary in a short abstract.

> *The web of confusion is now so tangled that not even Nimslo can say exactly how much investment ~~the company~~* **they** *~~has~~* **HAVE** *attracted.* (new-sci-B7L-5-ljh)

> *He ~~would silence~~* **SILENCED** *noise by yelling threats of appalling punishment.* (new-sci-B7L-72-ljh)

The following example illustrates how verb tenses need to be changed according to the deletion of the reporting construction at the beginning of the sentence, in this case, from present to past. Other deletions are also shown, in order to make the sentence grammatical:

> *~~DEFRA told WWT~~ samplers ~~to moisten~~* **MOISTENED** *a sterile swab on a stick with saline, ~~take~~* **TOOK** *a faecal sample from the bird, then put the swab back in its dry plastic tube~~. The tubes were kept~~***,** **KEPT** **it** *at refrigerator temperature and ~~taken~~* **TOOK it** *to ~~the testing~~ laboratories the next day.* (h02-ljh)

6.2.7 Passivisation

In contrast to the existing guidelines on abstract style (see Section 4.2), the human abstractor did not always prefer the active voice during this process of producing abstracts from extracts. In 7 cases, the summariser passivised an active sentence. A passive construction is one in which "the logical object of the action is made the grammatical subject of the assertion" (*The Oxford English Dictionary Second*

Edition 1989, Volume XI: 312), where the logical subject can either be absent, or present in a prepositional phrase. REPLACE: PASSIVISE is not a very frequent sub-operation, but it is worth noting because it contradicts the advice available to professional abstractors, and also strategies used by the 'expert' abstractors observed by Endres-Niggemeyer (1998). One of the reasons for the use of this sub-operation is to depersonalise sentences, as in the example below, where *we* cannot be resolved easily. Another reason is to remove subjects of sentences which are irrelevant or are not referred to later in the text. This also reduces the length of the abstract text overall.

> *It was a polymer so unlike the polymers known at the time that no one could envisage a use for it. And ~~we couldn't make it~~* **IT COULDN'T BE MADE** *consistently.* (new-sci-B7L-74-ljh)

6.2.8 Abbreviations

An obvious way to utilise space well when transforming an extract into an abstract is to abbreviate words which have standard abbreviations, acronyms or initialisms,[41] and symbols. Symbols are included as abbreviations because they shorten the text using a well-recognised form as a substitution for the word or phrase it represents. There are 16 cases of the operation REPLACE: ABBREVIATE in the corpus, mainly involving the % symbol and numbers. The human summariser replaced *per cent* with its symbol % 7 times, and written numbers with their corresponding figures 5 times. However, not all instances of numbers were replaced, even within the same text. A

[41] Crystal (2003: 120) states that the difference between acronyms and initialisms is that acronyms are pronounced as single words whereas initialisms are spoken as individual letters.

reason for this could be that further reduction of the text was not necessary after a certain point in the abstract due to the application of other operations. There are also 4 cases of words being replaced by standard abbreviations: *kilometres* by *km* and *per hour* by *ph*.

Abbreviations referring to organisations or projects are left in abstracts rather than replaced with their expanded forms. This is acceptable in these texts because the initialisms and acronyms are relatively common and easy for the reader to work out (or expand) from the given context, for example, *WHO*, *ME*, *DEFRA*, *EC*, *UK*. In this corpus, initialisms and acronyms are only deleted or replaced by their corresponding full form if they are not mentioned again in the abstract. This suggests that it would also be appropriate to replace the full forms of common abbreviations or acronyms with their short form as part of the text reduction process. However, it should be noted that the presence of too many abbreviations in a text can impede readability. This problem does not occur in this corpus because the texts are usually short and so only contain one or two acronyms and/or initialisms. In addition, the human summariser uses other replacement operations (REPLACE: PRONOMINALISE, REPLACE: LEXEME) to combat such repetition. Professional abstractors are usually given a list of abbreviations, acronyms, formulae etc. which are acceptable for inclusion in summaries by the organisation they are working for. Although this type of resource was not available to the human summariser producing abstracts for this corpus, they were consistent in the abbreviation of everything except numbers, suggesting that such units can be dealt with consistently even without guidelines. The application of this sub-operation reflects Cremmins (1996)'s advice to professional

summarisers to expand *lesser known* abbreviations and acronyms and to verbalise unfamiliar symbols (Section 4.2.3).

6.3 Reordering

The second complex human summarisation operation found in the corpus is *reordering*. *Reordering* occurs when a sentence or part of a sentence is moved from its original position in the extract into a new position in the abstract and is defined in this context as: *the deletion of a unit from one place in the extract and its insertion in a different place in the abstract*. This operation can also involve aspects of *replacement*. It may seem to overlap to some extent with *merging* in that units are moved around in the text. However, during reordering, units can be reordered within a sentence and units can be relocated without being incorporated into another unit. In the merging operation, different units can be incorporated into one unit without being reordered, for example, if units in between them are deleted. The main reasons for reordering units are to emphasise information and to improve the coherence and readability of the abstract. This operation relates to Jing and McKeown (1999)'s *sentence reordering*, although it also takes into account other units as well as sentences, and Endres-Niggemeyer (1998)'s *reorganize* strategy. It could also play a part in Cremmins (1996)'s rearranging which aims to make the text more concise and coherent. Although not explicitly related to transformations made to texts when summarising, Borko and Bernier (1975) advocate rearranging the text to save the reader's time, for example, by placing a conclusion at the beginning of the abstract. The automatic extraction literature does not discuss revision operations which explicitly deal with the reordering of sentences.

Unlike the other operations described in Chapter 5 and Chapter 6, reordering does not usually have any recognisable triggers allowing it to be identified and associated with certain units. This is because it is the content of the sentence which is important in this operation, and the forms representing the different content of all reordered units are too numerous and diverse to classify based on their surface realisation. In this section, therefore, reordering is discussed in terms of the function it achieves rather than any typical forms associated with it. Two main functions of reordering were identified during the corpus analysis: emphasising information and improving readability and coherence. Each function is dealt with in turn below, and whilst examples are discussed, they cannot be fully illustrated here due to space restrictions. However, the extract and abstract pairs containing most of the examples can be found in Appendix II.

6.3.1 Emphasising information

Reordering information so that it is positioned earlier in the abstract than it is in the extract emphasises the importance of that information. This is based on the idea that the most important information in a news text will be presented before other information (see the description of a *lead summary* in Section 3.4.1). An example from the corpus, in a text about face transplants (h01-ljh),[42] is the reordering of sentences about tissue-rejection from S7 and S8 position (the last two sentences) in the extract to S3 and S4 in the abstract, which contained 6 sentences. The reason for this reordering is that sentences about tissue-rejection are considered to be more

[42] This (extract, abstract) pair can be found in Appendix II (p.303-304), as well as in Section 6.7.

important than a sentence about scarring, which appears before them in the extract. In the same text, a sentence about the predicted rise in the number of face transplants is moved from S2 position in the extract to S6 (the last sentence) in the abstract. Quirk et al. (1985: 88-89) discuss *grammatical highlighting*, which performs the same function as REORDER: EMPHASISE: to emphasise certain information by relocating it in the text. Further aspects of grammatical highlighting are discussed under the class of *merging* operations below (Section 6.4.1).

6.3.2 Improving coherence and readability

The second function of reordering observed in the corpus is to try to improve the coherence of the text and make the abstract easier to read, by positioning sentences, or parts of sentences, about the same thing together. This means that the 'topic' of the abstract does not alternate, making it easier for the reader to process. The sub-function REORDER: COHERENCE includes repositioning sentences which seem to introduce or summarise the information in the extract to the beginning or end of the abstract, so it does not appear in the middle of more detailed text. For example, in a pair of texts which summarise a review of a television programme about science (new-sci-B7L-69-ljh), the clause *This was a solidly practical view of science* was moved from the middle of the extract to last sentence position in the abstract, because it concludes or summarises the text well.

A text about school memories (new-sci-B7L-72-ljh) provides an example of sentences being reordered so that sentences containing similar information appear together in the abstract. *It seemed to me that the teachers at school were, if not*

actually off their trolleys, a trifle on the demented side is reordered to earlier in the abstract so it appears immediately after other sentences about teachers at the school. In the extract, this sentence is separated from the teacher sentences by other sentences recalling a comic which helped the writer through his school days. As only one of the comic sentences is included in the abstract, it is more coherent to place this after all the teacher sentences.

6.4 Merging

The third and final class of complex operations identified in the corpus and discussed here is *merging*. *Merging* is defined in this context as *taking information from different units in the extract and presenting it as one unit in the abstract*. Like the other complex operations, it comprises the atomic operations deletion and insertion, but it can be further described as often including elements of *replacement* and *reordering*. Merging can take place both within and across sentences, and merged units can be of any granularity: word, phrase, clause, sentence. Based on the analysis, it is believed that this is the operation which best captures the essence of abstracting. As mentioned above (Section 6.3), it can be difficult to distinguish merging from reordering; one way to delimit the operations is to say that merging focuses on creating one unit out of information from more than one unit, which may or may not be moved to a different position within the text. Merged complete consecutive sentences, for example, will not involve reordering, although taking a concept from one sentence and inserting it into a non-consecutive sentence will do. Reordering is primarily concerned with the relocation of units, not necessarily units which are

being merged. A basic example is the deletion of a complete sentence from one place in the extract and its insertion without modification into another place in the abstract.

The main reason for the application of this operation is to make the text more concise, by fitting as much relevant information as possible into the compression rate. It allows the most appropriate parts of the most appropriate units to be combined, and as it often occurs with replacement, these can then be expressed in the most appropriate way. Similar to reordering, merging also functions to emphasise certain information, by combining it with information in other units.

This operation is similar to Jing and McKeown (1999)'s *sentence combination*, Chuah (2001b)'s *aggregation*, Cremmins (1996)'s global revision operation of *compression* and Endres-Niggemeyer (1998)'s *connect* strategy which comes under the heading *integrating text modules*. It is also related to Mani, Gates and Bloedorn (1999)'s sentence *aggregation* in the field of automatic summarisation, although that is concerned only with combining constituents from different sentences rather than from within the same unit, which is an option here.

There are two sub-operations of merging identified in the corpus: restructuring and punctuation/connectives. Of these, punctuation/connectives is the easiest to recognise, and can be identified most easily when consecutive sentences are merged. There are numerous variations in triggers for restructuring especially as other operations such as replacement and reordering are so often applied to the same unit, which is the reason for grouping them into one large sub-operation. Because not just existing surface forms are abstracted and merged, but the concepts, entities and

actions represented by them with other possible realisations, information can take almost any possible (appropriate) form. Both of these sub-operations are dealt with in turn below, and examples are given to illustrate them. In the examples, both extract and abstract sentences are presented to enable the reader to see exactly what changes are made; the changes are shown in [square brackets], with **bold** used to highlight the central change within the square brackets in the abstract sentence and ~~strikethrough~~ to show deleted units. Where insertions were also made, these are shown in **bold** without square brackets, and units subject to obvious replacement operations are indicated in {curly brackets}.

6.4.1 Restructuring of clauses and sentences

The sub-operation MERGE: RESTRUCTURE is concerned with the restructuring of units in terms of their position and form in the abstract compared with that in the extract. This sub-operation is widespread in the corpus, but it is very difficult to classify further because it so often involves most, if not all, of the other operations identified. The main functions of MERGE: RESTRUCTURE are to reduce the text, to make the abstract more concise and to improve its style or readability. It also emphasises important information, similar to reordering, as units containing certain information are relocated within the abstract. Quirk et al. (1985: 88-89) discuss *grammatical highlighting*, which performs the same function as MERGE: RESTRUCTURE in terms of emphasising information. The communicative and stylistic nature of grammatical highlighting is mentioned, which relates to the idea of post-editing a computer-aided extract as it is concerned with moving information in order to convey meaning in the best possible way. This also relates to the context factors presented in Chapter 2,

which discuss the potential communicative use of the summary as crucial to the final product as well as listing *style* as a factor affecting summarisation.

MERGE: RESTRUCTURE often involves sentences or parts of sentences (not always necessarily a clause) being transformed into a subordinate clause introduced by a relative pronoun or a non-finite verb and inserted into a previous sentence. Other realisations are sentences or parts of sentences appearing as prepositional phrases modifying the object of a previous sentence, adverb phrases and adjectives. Units are also transformed into coordinate clauses. There was one instance of an interrogative sentence being replaced by a declarative where the question and answer sentences were merged together. However, there are no apparent patterns as to when this happens, the human summariser seems to merge units intuitively to achieve maximum conciseness and coherence, depending on the abstract being produced. The combination of operations coupled with the difficulty of assigning a sub-class is typical of human summary production operations which embody abstracting, and highlights the complexity of a classification of these operations. Several examples are given of different cases of MERGE: RESTRUCTURE, along with an explanation of what exactly happened to the units involved and why. The examples given below are restricted to particularly interesting operations which best illustrate MERGE: RESTRUCTURE due to its prevalence in the corpus.

The first example shows an extract sentence which is transformed into a modifier of the object of another sentence, introduced by the preposition *into* and containing a relative clause of its own. This is due to the insertion of material inferred from, but not explicitly stated, elsewhere in the extract: *The research should give insights*, is

inferred from an earlier mention of *research* coupled with information about partners collaborating on a *project*, which enables the summariser to deduce that *research* and *project* refer to the same thing. This example is also discussed in Section 5.5.2, where it is shown that it can also be considered as *insertion from scratch*. This again highlights the interactive nature of the operations in the production of an abstract, as well as the difficulties in classifying them. The example also illustrates an attempt to 'summarise' the summary; the human summariser creates a final sentence which concludes the abstract.

> [*The way in which the atmosphere interacts with the oceans remains a last great unknown in Earth science*].
>
> ↓
>
> ***The research should give insights*** [***into*** *the way in which the atmosphere interacts with the oceans,* ***which*** *remains a last great unknown in Earth science*]. (sci13done-ljh)

The second example illustrates a case of three sentences being merged using restructuring as well as deletion, reordering and replacement to save space by taking only the most relevant parts of the extract sentences and putting them together in a grammatical and coherent way. The first extract sentence is transformed into a prepositional phrase postmodifying the object of the second extract sentence, whilst the third extract sentence becomes a subordinate clause. Because this example is more difficult to present, elements are labelled and an intermediate step is shown so that the reader can see more easily how merging takes place. The elements marked are not necessarily grammatical classes at the same level, but parts of the sentence which are involved in the operation in the intermediate representation.

[*Chernobyl reactor number 4 ~~was ripped apart by~~ an explosion on 26 April 1986.*}]$_1$ *Last September,*$_2$ {*the IAEA and the WHO*}$_3$ *released a report*$_4$[{*. Its headline conclusion*}$_5$ *that radiation from the accident would kill a total of 4000 people*$_6$ *~~was widely reported~~.*]$_7$

↓

2, {3}, 4, {1}, {5}, 6

↓

Last September,$_2$ {*the IAEA/WHO*}$_3$ *released a report*$_4$ [{***on*** *the explosion of Chernobyl reactor number 4 on 26 April 1986*},$_1$ {***concluding***}$_5$ *that radiation from the accident would kill a total of 4000 people.*]$_6$ (h03-ljh)

The following example illustrates a small piece of information introduced into the first sentence of an abstract from later in the extract. Even though the merged unit is small, consisting of the two word prepositional phrase *at Clarks* inserted to modify a noun phrase, it is useful because space does not have to be spent introducing *Clarks* later in the abstract. It can also be omitted from subsequent mentions of *the program*, *the system* etc., so saving even more space. Another reason that the PP is inserted is because the next mention of *Clarks* in the extract is the definite noun phrase *the Clarks program* which presupposes that the noun *Clarks* has already been introduced. The abstract sentence example contains information merged from four sentences, including a *referred* sentence, and involves the third extract sentence being transformed into a subordinate clause. Deletion and replacement operations are also applied:

~~*It took four years to develop*~~ *a CAD system for shoe making*~~*. The software, which was*~~ [{*written*} *by CAD Centre in Cambridge, gives pictures and patterns of finished shoes on a high-quality monitor screen*~~*.*~~] [*It uses the Centre's Polysurf program which can define free-form surfaces.*] … [*Pattern flattening is done by additional mathematics on a specially written part of* [***the Clarks***] *program.*]

A CAD system for shoemaking [[**at Clarks**]**,** [{**developed**} *by CAD Centre in Cambridge, gives pictures and patterns of finished shoes on a high-quality monitor screen* [***using*** *the Centre's Polysurf program which can define free-form surfaces.*]]] (new-sci-B7L-54-ljh)

The next example is one of the best illustrations of how human summarisers can identify a relevant piece of information, reduce it to one word based on their background or world knowledge and locate it in the most appropriate place in an abstract. It is still considered to be summary production because the information is taken from the extract and presented in a more appropriate way; it just happens to use world knowledge to infer meaning and greatly reduce the text at the same time. This example shows just how well restructuring (and replacement) works to reduce the text and make it concise and easier to read, with two sentences which total 37 words being replaced by one adjective in an earlier sentence:

In October 1980 {*Zuccarelli*} *filed a European patent application, covering nine countries including Britain*[*. … The cost of pushing a European patent through in nine countries is around $10000. The cost of application alone is around $2000 and Zuccarelli has already paid an extra $500 for a further stage of official examination*].

In October 1980, {*he*} *filed an* [{***expensive***}] *European patent application, covering nine countries including Britain…* (new-sci-B7K-37)

The last example of the sub-operation MERGE: RESTRUCTURE again reduces the abstract text substantially, this time by transforming two complete sentences in the extract into a prepositional phrase acting as an adjunct in the first sentence of the abstract. It again highlights the usefulness of a human summariser's ability to perform complex replacement and restructuring sub-operations which work together to merge sentences and information into a concise and coherent abstract. This example also contradicts the instructions given to human summarisers and annotators (see Section 4.4) not to include examples in their summaries because they give too much detail or are unnecessary in a summary. In this case, examples of the two extremes of the climatic events are necessary to demonstrate the severity of the situation.

The southern hemisphere has experienced extreme climatic events in the past year[*. … Widespread flooding, loss of life, property damage and economic disruption has occurred across one-third of Ecuador. Since October last year, South Africa has been in the grip of serious drought, with agricultural production expected to decline by at least 700 million Rand*].

The southern hemisphere has experienced extreme climatic events in the past year[*,* ***from*** *widespread flooding* ***to*** *serious drought*]. (new-sci-B7K-44)

6.4.2 Punctuation/connectives

As mentioned in Section 5.5, punctuation is replaced by other punctuation during merging. Although this could be classed as *replacement*, because strictly, one unit (punctuation) is replaced by another (punctuation) via the deletion of one unit and the insertion of another, this particular sub-operation only occurs during the merging of units. Connectives are treated in a similar manner, although these can be used either to replace punctuation or in addition to punctuation which has been replaced. The MERGE: PUNCTUATION-CONNECTIVE sub-operation is necessary because without it, sentences would not be merged but remain separate. The use of commas during merging also helps with readability, and connectives function to ensure grammaticality and improve the flow of the text.

The replacement of punctuation discussed here only refers to punctuation replaced during the merging of complete sentences and units no smaller than a clause across sentences. This is because when sentences are merged, the punctuation at the end of the first merged sentence is *always* replaced, whereas during the merging of smaller units within the same sentence or of concepts rather than specific linguistic realisations, punctuation is not always an issue. In total, there are 69 cases of punctuation being replaced by punctuation or a connective, 56 of these being punctuation-punctuation, and 13 punctuation-connective. The most common punctuation used in merging is a comma (43 cases), but hyphens are also used. The connective *and* is preferred when a connective replaces punctuation, as well as when a connective is added after punctuation (20 cases in total, 12 *and*), although *but*, *as*, *so* and *although* also feature in these cases. An example of two consecutive sentences merged with the connective *and* from the corpus is:

A pointer is touched against each point of intersection on the grid and the digitizer records the position[**.**] *The computer displays a wire-frame model…*

↓

A pointer is touched against each point of intersection on the grid ~~*and*~~***,*** *the digitizer records the position* [***and***] *the computer displays a wire-frame model.* (new-sci-B7L-54-ljh)

6.5 Implementability of operations

As this research is carried out in the field of computer-aided summarisation, it is appropriate to discuss any possible future implementations of the operations and sub-operations identified during the corpus analysis in a computer-aided summarisation system such as CAST (see Section 3.5.4). There are several possibilities for future implementation which arise from the discussion of sub-operations above. Due to the subjective nature of summarisation, and due to the fact that this thesis addresses issues of style and summary production, it is not possible to have a fully automatic implementation of many of the sub-operations identified in the corpus. There are always exceptions to the rule and often the context and function of the unit is important in deciding when an operation can be applied. This means that the most suitable implementation plan involves possible applications of sub-operations being indicated to the human summariser, using automatic processing methods (see Section 6.5.1). However, this can be inappropriate if, for example, there are so many instances of potential application that it is confusing for the user to indicate so much information on the screen (see Orasan and Hasler (2006) for a related discussion of user feedback on CAST). When this is the case, the summariser needs to access

information about transformations in a set of summary production guidelines (see Section 6.5.2).

6.5.1 Sub-operations which can be facilitated by implementation

Implementation to facilitate certain sub-operations can generally do so by highlighting units to which sub-operations could be applied in the computer-aided extract presented to the user. This may not be practical in some cases because of the high number of instances of particular units and it would be confusing to highlight them all, especially if all the sub-operations are to be taken into account. In these cases, a dialogue box could appear to notify the user that a particular unit may be suitable for transformation via a certain sub-operation when they select it for inclusion in their summary. Sub-operations suitable for some type of implementation are: DELETE: SENTENCE, DELETE: REPORT, DELETE: FORMAT, DELETE: NP, DELETE: *BE*, INSERT: MODIFIER, REPLACE: PRONOMINALISE, REPLACE: LEXEME, REPLACE: VP, REPLACE: ABBREVIATION and REORDER: COHERENCE.

In terms of specific implementations for particular sub-operations, there are several options. To facilitate DELETE: SENTENCE, during automatic extraction the weights of complete sentences can be required to meet a certain threshold in order to be considered *essential* and *important*. This could allow *important* sentences in the computer-aided extract to be indicated to the summariser in one of the ways described above, showing that they are particularly suitable for deletion. In the case of DELETE: REPORT, reporting constructions can be identified fairly easily automatically because they usually follow the same pattern of

NP+V(reporting)+CONJ(*that*). A list of genre-dependent reporting verbs can be provided to aid the identification, similar to the way indicating phrases (Paice 1981) are used in automatic summarisation. They can be indicated to the user during extraction, who would then determine whether they should be deleted or retained based on the context. Similarly, DELETE: FORMAT can be identified using a list of punctuation and highlighted in the extract to the user, or indicated via a dialogue box. It would also be feasible to create a list of frequent words for news texts which can be abbreviated and thus to abbreviate them automatically before the extract is presented to the user, thereby automating the application of the sub-operation REPLACE: ABBREVIATION.

Although noun phrases are extremely frequent in the corpus, it is possible to automatically compute coreference chains for entities in a text so that it is easier for the summariser to decide which NPs, or parts of NPs such as modifiers, can be safely deleted without leaving any ambiguity about the entity to which it refers. This would facilitate the identification of units most suited to the application of the sub-operation DELETE: NP. One visualisation option for coreferential chains would be to highlight the chains within the extract which is presented to the user. Alternatively, when the user selects a sentence containing an entity which is part of a coreferential chain, the chain could be displayed at the side of the main window. This is also a possibility for the sub-operation INSERT: MODIFIER. By highlighting or presenting chains, the summariser can see when it is necessary to insert a modifier to avoid ambiguity or to reduce the text because inserting a modifier earlier in the abstract for one NP would save space later.

Also related to the way that computing coreferential chains could facilitate sub-operations are REPLACE: PRONOMINALISE and REPLACE: LEXEME. In the case of REPLACE: PRONOMINALISE, NPs appearing within a certain distance of the first mention could be automatically replaced by pronouns during extraction. Although not all subsequent mentions of an entity in the abstracts in the corpus are pronominalised, almost all instances of pronominalisation occur either in the same sentence as their antecedent[43] or in the following sentence. This means that it is feasible to automatically replace subsequent mentions of a given NP within two sentences of the initial mention with a pronoun referring to it. This proposal is further justified by the fact that most anaphora resolution systems which do not identify discourse segment boundaries have a search scope of the current and two or three preceding sentences (Mitkov 2003). An alternative would be to allow the user to see the coreferential chain for NPs mentioned in sentences they select for their abstract and allow them to determine which instances to pronominalise. This alternative approach could also be taken for REPLACE: LEXEME, although if a resource such as WordNet is included in the system it may be possible to automatically suggest alternative lexical realisations. One problem with operations involving NPs is the frequency of noun phrases in the extracts, which could make implementation impractical and confusing if such a high number of coreferential chains need to be presented to the summariser.

The verb *be* in the first sentence of extracts can be automatically identified during extraction and indicated to the summariser via highlighting in the extract to allow

[43] For simplicity, and due to the widespread practice of lexical substitution (see Section 6.2.2), in this analysis the antecedent of a pronoun is considered to be its most recent mention, even if there is a coreferential chain stretching further back.

them to decide whether or not they wish to transform the first sentence into a 'headline' for the abstract using the sub-operation DELETE: *BE*. Another sub-operation dealing with verbs, REPLACE: VP depends so much on other operations and does not allow the inference of any definite 'rules', that an attempt to implement aspects such as preferring the simple present or past tense may seem inappropriate. However, tenses and other aspects of verb phrases are *always* changed according to other sub-operations applied so that they are grammatical within their new context. A 'grammar checker' which would underline ungrammatical units could be employed in a computer-aided summarisation system and used during post-editing to aid the user in their task.

In terms of REORDER: COHERENCE, if a method is available to identify 'topics' of sentences reliably, then it would be possible to automate this sub-operation by presenting sentences about the same topic together, starting with the first sentence extracted. Sentences classed as being about different topics would be presented towards the end of the extract. However, it is not clear that this is the most suitable course of action, given that if the compression rate is substantially lowered, it could result in something important being missed out of the extract before it is presented to the user simply because it does not contain exactly the same topic as the first sentence. As an example, in the text about school memories mentioned in Section 6.3.2, this could mean that sentences about the comic are omitted, which would not be desirable. This aspect would need to be investigated further before a decision is made.

Although a number of sub-operations identified during the corpus analysis have been discussed in this section as suitable for possible future implementation, a word of caution is necessary. Computational linguistics methods used to (semi-)automate such sub-operations, such as coreference resolution, are not perfect, i.e., they will not achieve 100% accuracy, and this could hinder the computer-aided summarisation process by introducing inaccuracies which would not be present if a human summariser performed all aspects of the operations manually. An investigation would be needed to fully assess the usefulness of automating the sub-operations or parts of them, to determine whether it actually helps the user, but this is beyond the scope of the present research.

6.5.2 Sub-operations unsuitable for implementation

In cases where even the indication of material which is potentially suitable for transformation may be less helpful than leaving the summariser to decide about the application of sub-operations, the suitability of these particular units can be indicated via a set of summary production guidelines. Guidelines should be presented to the summariser in order to help them transform the computer-aided abstract consistently and should contain information about all possible operations, not just those which cannot be facilitated by any kind of automatic processing. The summary production guidelines presented in Section 6.6 include guidance on using all the operations and sub-operations discussed in the classification.

Reasons why certain sub-operations are unsuitable for any kind of implementation are either a very high frequency of triggers or a very low number of applications in

the corpus. In such cases, it would either be confusing to attempt to indicate possible triggers where they occur very frequently or of little use to indicate the possibility of applying a sub-operation because it occurs very infrequently. Sub-operations from which definite patterns regarding their application cannot be derived are also unsuitable for implementation because of their unpredictable occurrence and seemingly intuitive application. When sub-operations are often subject simultaneously to other operations, automation can be unsuitable if they are difficult to distinguish. Sub-operations which need to be detailed in guidelines rather than dealt with automatically or semi-automatically are: DELETE: DETERMINER, DELETE: SUB_CLAUSE, DELETE: PREP_PHRASE, DELETE: ADVERB, DELETE: PUNCTUATION, INSERT: CONNECTIVE, INSERT: FORMULAIC, INSERT: PUNCTUATION, REPLACE: PASSIVISE, REPLACE: NOMINALISE, REPLACE: NP, REORDER: EMPHASISE, MERGE: PUNCTUATION-CONNECTIVE, and MERGE: RESTRUCTURE.

6.6 Guidelines for summary production in the computer-aided summarisation of news texts

As discussed in Chapter 4, professional summarisers often have guidance, in the form of instructions about what summaries should and should not include, both content- and style-wise, to ensure that their work is of a consistently high standard. Human summarisers interacting with automatically-produced summaries in the field of computer-aided summarisation are no different. Appropriate guidelines are necessary to facilitate consistency and quality. Section 4.4 presented a set of guidelines for the production of extracts from news texts. This section describes guidelines based on the corpus analysis and resulting classification discussed in this

chapter and the last. They have been developed for users of computer-aided summarisation systems such as CAST to help them consistently apply transformations which create a coherent and readable abstract from the computer-aided extract presented to them. These guidelines represent a condensed version of the classification, and their purpose is to instruct and advise during summary production rather than to analyse and describe. They can be found in their entirety in Appendix III. The guidelines, and therefore the classification, are assessed in Chapter 7 via an evaluation of abstracts which were produced from extracts by a human summariser using the summary production guidelines. The guidelines are split into two main sections: general strategy and information, and specific summary production operations.

6.6.1 Guidelines: section 1

Section 1 addresses general issues of the summary production (or post-editing) process, such as ensuring that the extract is read through and fully understood before any editing commences, and the fact that summarisers always have the option to refer to the source text in any cases where clarification is required. It is important that the summariser identifies the main topic of the extract on the first reading, and in the second reading, attempts to identify units which can be subject to transformation operations during summary production. It also states that an abstract should be created in one intensive period in order to avoid problems of inaccuracy and the need for the summariser to refamiliarise themselves with the extract. Once completed, the abstract should be read thoroughly to check for any problems or inconsistencies, and for any other changes that the summariser wishes to make.

It is pointed out to the summariser that the operations described in the guidelines should only be applied if they are sure that the meaning of the text will not be changed in any way, and only if they are appropriate in the context of a particular extract. There will *always* be exceptions to the rule or problematic cases in any task involving human language, and it is important to remember this. Therefore the summariser must use their linguistic expertise, the available context, and the guidelines to create the best possible abstract. It is mentioned that the operations described in the guidelines are not always applied in all cases and the guidelines should be taken as indicators of operations which are appropriate and *can* be applied to an abstract rather than a strict set of instructions about what *must* be done in every instance of certain phenomena in an extract. The basic idea is that if the summariser finds they need to perform transformation operations on a computer-aided extract, then they should perform those operations detailed in the guidelines because they have been observed in the creation abstracts from extracts by a human summariser.

6.6.2 Guidelines: section 2

Section 2 of the guidelines deals with the specific summary production operations classified in Chapter 5 and the earlier part of this chapter. Each class of operations is dealt with in turn, starting with the atomic operations *deletion* and *insertion*, followed by the complex operations *replacement*, *reordering* and *merging*. Instructions for the sub-operations within each class are presented, along with examples to enable the summariser to see an application of each sub-operation identified in the classification. Examples are not given here because several

examples of most of the sub-operations are presented earlier in this chapter and in Chapter 5.

Deletion

Deletion is important because even though the abstract is created from an extract, and even if there is no specified length restriction or compression rate, the text can be further reduced to fit the most relevant information into the smallest space. Each unit should be considered carefully before deletion, because it is important to retain the unit which represents the relevant information in the most appropriate form. Post-editors are instructed to delete **complete sentences** which are less important than other complete sentences in the extract and are not directly related to the main topic. Complete sentences introducing examples, additional detail and explanations, or presenting redundant information which is better presented elsewhere in the extract, are also suitable for deletion, as are those offering speculation or opinion unless this is a significant part of the main topic. In the field of computer-aided summarisation, it may also be necessary to delete complete sentences due to the failure of the automatic summarisation methods employed in the system to eliminate irrelevant and redundant sentences. Experiments using the CAST system also provide evidence that human summarisers initially select more 'important' information than they need for a summary, and then reduce this to the desired length (Orasan and Hasler forthcoming). **Subordinate clauses** containing non-essential information such as examples, explanations and temporal and spatial information can be deleted if necessary, as can those containing information which can be inferred from other

units in the text. However, subordinate clauses which are necessary to avoid ambiguity should be retained.

Prepositional phrases (PPs) which add too much detail and introduce redundant information by referring to something previously mentioned in the extract are identified as suitable units for deletion. They include temporal and spatial adjuncts and postmodifiers of noun phrases. However, those prepositional phrases which disambiguate noun phrases should not be deleted because this will impede the meaning and understanding of the abstract. **Adverb phrases** behaving in a similar manner to PPs suitable for deletion can also be safely deleted. **Modifiers of noun phrases** (NPs) such as adjectives and other parts of NPs which are not necessary to avoid ambiguity are identified as appropriate candidates for deletion, as are **noun phrases** which repeat previously-mentioned entities. The **reporting clauses** accompanying indirect speech, or indirect reporting of any kind (thoughts, claims, findings, etc.), can be deleted from sentences providing that the main point of the extract is not to offer contrasting opinions on a particular topic. In extracts which are not about giving different opinions, the reported information is important but who reported it is not.

The deletion of **determiners**, the verb **be**, and **punctuation** are listed as ways to reduce the text further by small amounts, which can be especially useful if the summariser is working towards a compression rate and has applied other sub-operations to the extract but is still just outside the length restriction. In addition, the deletion of determiners and *be* help to create a 'headline', which can be desirable in abstracts of news texts. Punctuation is sometimes incorrect in extracts, which is

another reason for its deletion. **Specially formatted text** such as bracketed text or text following a semi-colon is also given in the guidelines as a possible candidate for deletion. However, post-editors are warned that care must be taken to assess the information content of this text, because it cannot be safely deleted as often as some of the other units mentioned: the content is much more important than the fact that it appears in a certain format.

Insertion

Insertion is a useful operation because extracts can be incoherent due to the fact that they comprise sentences or units taken straight from the source text without any modification. Inserting units during summary production can increase coherence and improve the readability of the summary. Post-editors are instructed to insert **connectives** to make the text flow better by explicitly signalling coherence relations. Coordinators, subordinators and adverbs functioning as conjuncts can all be inserted. *However* in sentence-initial position is particularly appropriate, but connectives can also be inserted mid-sentence. Standard patterns or units typical of the style of news texts (**formulaic units**) can also be inserted to improve readability in a standard style, as well as functioning to draw attention to the information contained in that unit. Summarisers are advised that **modifiers** which disambiguate nouns and clarify meaning should be inserted because this saves space by making it unnecessary to introduce the information later in the abstract in a larger unit, such as a sentence. **Punctuation** is also given as a potential candidate for insertion, depending on its accuracy in the extract and summariser's style preferences regarding readability and flow of text.

Replacement

Replacement is an important aspect of summary production as it allows the summariser to reformulate units to make the abstract more concise rather than just shortening it. The summariser is instructed to **pronominalise** noun phrases in order to reduce the abstract and to avoid repetition. This is useful because extracts are usually 'about' one thing, references to which would be repetitive if pronominalisation is not applied. **Lexical substitution** should be applied to units for the same reason. Different lexical realisations are not restricted to NPs alone, and so lexical substitution is possibly more useful, or at least further-reaching, than pronominalisation. Post-editors are also instructed to **restructure noun phrases** specifically by transforming postmodifiers into premodifiers, which saves space by conveying the same information in a smaller unit. Similarly, sentences or clauses should be **nominalised** where possible because of the extent to which they can reduce the text and the way that they present information more succinctly. However, this has implications for the surrounding units as a nominalisation never constitutes a sentence or clause alone, but must be merged into a larger unit which will also require transformation to ensure grammaticality.

Pronoun expansion is mentioned explicitly in the guidelines as an option to transform text for the abstract. Pronoun expansion is only necessary on the first mention of the pronoun in the abstract, because if the converse to pronominalisation occurs throughout, extra material is introduced into the abstract, lengthening rather than shortening the summary. Words and phrases which have **standard**

abbreviations or symbols to represent them should be abbreviated, and standard acronyms and initialisms introduced in the abstract should be retained, to save space as well as making the abstract easier to read. If acronyms and initialisms which are usually kept in their abbreviated form are expanded, this will make the abstract less concise and possibly affect the ease of processing of the unit because it is unexpected.

Summarisers are advised to apply a variety of changes to the extract under the heading of **verb phrases**, which includes preferring the simple present or simple past tense where appropriate in the abstract. This makes the text easier to read as well as usually shorter, depending on the verb phrase taken as the starting point. In addition to individual uses of verb phrase editing, verb phrases need to be altered when other operations applied to a unit containing them or adjacent to them require it. For example, the summariser must ensure the verb agrees in number with the subject of the sentence. **Passivisation** is also presented as an editing option, particularly to depersonalise sentences and to delete irrelevant subjects of sentences, or subjects which are not referred to again in the abstract.

Reordering

Reordering is used to **emphasise information** by placing sentences containing more important information than others earlier in the abstract and relegating less important information to nearer the end of the abstract. Important units within a sentence should also be reordered to the beginning of the sentence. **Coherence and readability** should be improved by reordering the abstract so that units about the

same topic, or the same aspects of the main topic, appear together rather than being separated by other information. The summariser is also instructed to place sentences which introduce or conclude the main topic of the extract at the beginning or end of the abstract, respectively.

Merging

Merging is an extremely important aspect of summary production but it is difficult to formalise into instructions within guidelines because the possible methods of merging are many and varied. Summarisers are instructed, therefore, to **restructure clauses and sentences** so that they reflect the importance of the information conveyed by them. This could take the form of transforming a sentence into a relative clause modifying the object of the previous sentence, or a prepositional phrase or an adjective modifying an NP in a different sentence. Restructuring sentences and clauses also saves space and avoids repetition, making the abstract more concise. **Punctuation** and **connectives** are essential parts of merging. When two or more sentences are merged, the punctuation at the ends of the sentences needs to be dealt with. Post-editors should replace full stops with commas or hyphens or a connective, or they should replace them with commas or hyphens and then insert a connective to make the abstract read better and to ensure that the sentence is grammatical.

6.7 Analysis of a summary: an example

The previous sections of this chapter have completed the classification of human summary production operations identified during the corpus analysis and discussed

issues relating to them such as possible future (semi-)automation. The set of guidelines formulated from the classification has also been presented. This section ties all the aspects of the classification together by presenting an (extract, abstract) pair from the corpus analysed in terms of summary production operations. The aim of this is to illustrate the extent to which operations can be applied within an extract to transform it into an abstract. This particular pair of texts shows examples of sub-operations from all of the general operations: *deletion*, *insertion*, *replacement*, *reordering* and *merging*. There may be cases in the pair where the transformation applied to a unit does not seem to fit precisely into a particular sub-operation from the classification, or where it could be interpreted in more than one way. This is because the classification is based on the whole corpus and not just this pair of texts, and most texts have idiosyncracies which can be difficult to classify.

The text taken from the corpus to illustrate the classification of human summary operations is about the world's second face transplant being performed in China (h01-ljh). The extract and abstract are presented first, without annotations for operations applied, to enable the reader to clearly see the starting point and the final summary. For ease of reference, sentences in the extract and abstract are numbered, the extract sentences being prefixed with *E*, and the abstract sentences with *A*. The discussion takes the sentence ordering in the abstract as its basis, and presents annotated sentences to illustrate the operations applied to the extract in order to create the abstract. The key for the operations used in the discussion, which allows more than one operation applied to a particular unit to be illustrated at once, is as follows: ~~Deleted~~, (INSERTED), **replaced**, {reordered}, [merged]. Insertions are not shown in the annotated extract sentences, nor are deletions shown in the annotated

abstract sentences, for ease of reading. Punctuation involved in merging operations is not discussed, as it can be seen easily in the examples and would further complicate the explanation of the other sub-operations.

World's second face transplant performed in China.$_{\mathbf{E1}}$ *Experts predict the number of these operations will rise rapidly as centres around the world gear up to perform the procedure.*$_{\mathbf{E2}}$ *Thirty-year-old Li Guoxing received a new upper lip, cheek and nose from a brain-dead donor to repair injuries sustained after an attack by a black bear.*$_{\mathbf{E3}}$ *He was reported to be in a stable condition and taking liquid food following the 13-hour surgery on Thursday at Xijing hospital in Xian.*$_{\mathbf{E4}}$ *The surgery scar will not be obvious but there is a difference in the donor's and recipient's skin colour, so that will be noticeable.*$_{\mathbf{E5}}$ *Guoxing is reportedly happy with his new face, which will be improved by further treatment over time.*$_{\mathbf{E6}}$ *It will be two months until they are sure that Guoxing has not rejected the new tissue.*$_{\mathbf{E7}}$ *Rejection of the transplanted facial tissue could have life-threatening consequences, and the immunosuppressant drugs used to keep this from happening can make a person more prone to certain cancers.*$_{\mathbf{E8}}$ (h01-ljh: extract)

In the world's second face transplant, Li Guoxing, 30, has received a new upper lip, cheek and nose from a brain-dead donor following a bear attack.$_{\mathbf{A1}}$ *He is in a stable condition after the 13 hour surgery at Xijing hospital, Xian, China, and is happy with his new face, which will be improved by further treatment.*$_{\mathbf{A2}}$ *Doctors need 2 months to be sure that the new tissue is not rejected.*$_{\mathbf{A3}}$ *Rejection can be life-threatening and the immunosuppressants used to stop it can increase the risk of cancer.*$_{\mathbf{A4}}$ *The scar will not be obvious, but the donor's and recipient's skin colour is different.*$_{\mathbf{A5}}$ *The number of face transplants is predicted to rise rapidly, as centres world-wide prepare for the procedure.*$_{\mathbf{A6}}$ (h01-ljh: abstract)

Producing sentence A1

[*World's second face transplant* ~~*performed in China.*~~]$_{\mathbf{E1}}$ … [{***Thirty***}***-year-old*** {*Li Guoxing*} ***received*** *a new upper lip, cheek and nose from a brain-dead donor* ~~*to repair injuries sustained*~~ ***after an attack*** ~~***by a black***~~ ***bear***.]$_{\mathbf{E3}}$

[[(*IN THE*) *world's second face transplant*(,)] [{*Li Guoxing*}(,) {*30*}(,) ***has received*** *a new upper lip, cheek and nose from a brain-dead donor* ***following a bear attack***.]]$_{\mathbf{A1}}$

In order to arrive at sentence A1, sentence E1 is merged with E3, which is reordered to emphasise important information (REORDER: EMPHASISE), via MERGE: RESTRUCTURE. E1 becomes a prepositional phrase acting as an adjunct, conveying important information in relation to the reordered E3 and demonstrates an instance of DELETE: PREP_PHRASE (*in China*). The verb *performed* does not fall into any of the atomic deletion sub-operations, but it is still part of MERGE: RESTRUCTURE because the sentence structure is changed by the merging of the two sentences. Also applied to the extract to form A1 are REPLACE: ABBREVIATE to shorten *thirty-year-old* to *30* and REORDER: COHERENCE so that *Li Guoxing* appears before *30*. After this, REPLACE: VP is applied, and DELETE: SUB_CLAUSE permanently deletes *to repair injuries sustained* which adds unnecessary information and uses up compression. REPLACE: LEXEME replaces *after* with *following*, which is also an instance of MERGE: RESTRUCTURE as it introduces a subordinate clause to the sentence. Finally, REPLACE: RESTRUCTURE_NP transforms the noun phrase *an attack by a black bear*, where the NP is postmodified by a prepositional phrase, into a smaller unit (*a bear attack*),

because the type of bear is not important in this summary. INSERT: PUNCTUATION is applied in two places to make the resulting abstract sentence easier to read.

Producing sentence A2

[~~*World's second face transplant performed in*~~ {*China*}.]$_{\mathbf{E1}}$ … [*He* ***was reported to be*** *in a stable condition* ~~*and taking liquid food*~~ ***following*** *the 13-hour surgery* ~~*on Thursday*~~ *at Xijing hospital* ~~*in*~~ *Xian.*]$_{\mathbf{E4}}$ … [{~~*Guoxing*~~ *is* ~~*reportedly*~~ *happy with his new face, which will be improved by further treatment* ~~*over time.*~~}]$_{\mathbf{E6}}$

↓

[[*He* ***is*** *in a stable condition* ***after*** *the 13 hour surgery at Xijing hospital*(,) *Xian*(,) [{*China*}]](, *AND*) [{*is happy with his new face, which will be improved by further treatment.*}]]$_{\mathbf{A2}}$

Sentence A2 is produced by firstly applying the sub-operation REPLACE: VP to simplify and shorten the verb phrase *was reported to be* to *is*, and DELETE: SUB_CLAUSE to delete information deemed unnecessary by the summariser to sentence E4. MERGE: RESTRUCTURE is then used to merge the subordinate clause *following the 13 hour surgery…* with the main clause using REPLACE: LEXEME to change *following* to *after*, which also avoids repetition with the preceding sentence. The prepositional phrase *on Thursday* is deleted with DELETE: PREP_PHRASE, and the preposition *in* is deleted as the sentence still makes sense without it, although this operation was not frequent enough in the corpus to merit a sub-operation. *China* is reordered from sentence E1 using REORDER: COHERENCE so that it appears with other location information. This is also an instance of merging as it is deleted from one sentence and inserted in another. E6 is merged with E4 using MERGE: PUNCTUATION-CONNECTIVE, so that information about Guoxing's condition or state appears in the

same sentence. *Guoxing* is deleted by DELETE: NP because it can be recovered from the beginning of sentence A2. DELETE: ADVERB is applied to remove *reportedly* and DELETE: PREP_PHRASE to remove *over time* because these are not crucial to the understanding of the abstract. INSERT: PUNCTUATION in two places makes the sentence more readable by separating the 'list' describing the hospital location.

Producing sentence A3

{***It will be two*** *months* ***until*** {***they***} ***are*** *sure that* ~~***Guoxing***~~ ***has not rejected the new tissue.***}$_{\mathbf{E7}}$

[[{{(***DOCTORS***)}] [***NEED 2*** *months* ***to be*** *sure that* ***the new tissue is not rejected.***}]]$_{\mathbf{A3}}$

A3 corresponds to E7, reordered by REORDER: EMPHASISE because tissue rejection is more important than scarring. A3 is formed by applying pronoun expansion to *they*, using world knowledge that in a text about medical issues, *doctors* would be the people who assess tissue rejection. This can be classified as an instance of MERGE: RESTRUCTURE because information is inferred from a different place, or as an insertion of an NP because this information does not appear elsewhere in the extract. It is also an instance of REPLACE: LEXEME because, even though a pronoun is involved, pronominalisation does not occur and one lexical realisation is replaced by another. It is then reordered to the beginning of the sentence using REORDER: COHERENCE. The conjunction *until* is deleted, although there were not enough cases in the corpus to merit a sub-operation.

REPLACE: LEXEME changes the verb *be* to *need*, and REPLACE: VP introduces the present tense. REPLACE: VP also affects the verb *be* in the subordinate clause, where the clause constituting the object is passivised using REPLACE: PASSIVE because it is shorter than the subordinate clause from the extract. In addition, *Guoxing* can be 'recovered' from elsewhere in the text because the reader knows that the summary is about Guoxing's face transplant, so DELETE: NP is applied to *Guoxing*.

Producing sentence A4

{*Rejection* ~~*of the transplanted facial tissue*~~ ***could have life-threatening consequences~~,~~*** *and the* ***immunosuppressant drugs*** *used to* ***keep this from happening*** *can* ***make a person more prone to certain cancers.***}$_{\mathbf{E8}}$

{*Rejection* ***can be life-threatening*** *and the* ***immunosuppressants*** *used to* ***stop it*** *can* ***increase the risk of cancer.***}$_{\mathbf{A4}}$

A4 corresponds to E8, also reordered by REORDER: EMPHASISE for the same reason that E7 is, and utilises DELETE: PREP_PHRASE to remove redundant information and minimise repetition about the transplanted tissue. *Life-threatening consequences* is replaced by *life-threatening* using an operation which can be classed as REPLACE: RESTRUCTURE_NP in order to reduce the text, which means that the verb in the sentence has to be changed accordingly using REPLACE: VP. REPLACE: VP is also used to simplify the tense and REPLACE: LEXEME changes *have* to *be*. The rest of the sub-operations applied in this sentence also function to shorten the text. REPLACE:

LEXEME results in the noun *immunosuppressants*, and further REPLACE: LEXEME sub-operations are applied, creating *stop it* from *keep this from happening*, as well as *increase the risk of cancer* from *make a person more prone to certain cancers*. DELETE: NP also deletes the modifier *certain* to generalise the noun. DELETE: PUNCTUATION is used to remove a comma from the abstract sentence because it is unnecessary.

Producing sentence A5

{*The ~~surgery~~ scar will not be obvious but* ***there is a difference in the donor's and recipient's skin colour~~, so that will be noticeable~~.***}$_{\mathbf{E5}}$

{*The scar will not be obvious*(,) *but* ***the donor's and recipient's skin colour is different.***}$_{\mathbf{A5}}$

Sentence A5 corresponds to the reordered E5, which was relocated using REORDER: EMPHASISE because its content was considered less important than the sentences about tissue rejection. DELETE: NP is applied to remove *surgery*, because the reader can easily infer that the scar is related to the surgery. The second main clause is shortened by simplifying the structure with REPLACE: VP by deleting *there is*, and placing the focus on the people involved. DELETE: SUB_CLAUSE is applied to remove *so that will be noticeable*, removing redundant information which the reader can infer from the previous clause. DELETE: PUNCTUATION and INSERT: PUNCTUATION are also involved, both applied to commas.

Producing sentence A6

{~~*Experts*~~ ***predict the number of these operations will*** *rise rapidly as centres* ***around the world gear up to perform*** *the procedure.*}$_{\mathbf{E2}}$

{***The number of face transplants is predicted to*** *rise rapidly*(,) *as centres* ***world-wide prepare for*** *the procedure.*}$_{\mathbf{A6}}$

The final abstract sentence, A6, corresponds to sentence E2, reordered by the sub-operation REORDER: COHERENCE as it allows other related sentences to be presented together. It is not exactly related to the discussion of Li Guoxing in the rest of the sentences, but it provides a conclusion to the abstract. REPLACE: PASSIVISE is applied to the main clause in E2 and the agent *experts* is deleted, after which REPLACE: LEXEME is used to emphasise the type of operation (*face transplants*) as it was not mentioned explicitly in at least the three previous abstract sentences. REPLACE: LEXEME is also used to replace *around the world* with *world-wide* in order to shorten the NP and again to shorten the verb phrase which is changed to *prepare for* in the abstract. INSERT: PUNCTUATION inserts a comma to enhance readability.

It should be noted that there are other possible interpretations available for certain aspects of the analysis of the operations applied to the extract above, but due to space restrictions they cannot be discussed here. The subjectivity involved in the interpretation of natural language means that alternative analyses are always available. The analysis presented above is the one which fits best when the other

examples given in the classification and the texts themselves are taken into account. The discussion of the (extract, abstract) pair emphasises the complex nature of abstracting due to the extensive interaction of summary production operations.

6.8 Conclusions

The purpose of this chapter was to present the second part of the classification of human summary production operations introduced in Chapter 5: *complex* operations, and to discuss issues arising from the classification as a whole. Sections 6.2, 6.3 and 6.4 presented the classification of the complex operations *replacement*, *reordering* and *merging*, respectively, including their various sub-operations and examples. Section 6.5 discussed issues of implementation of all the sub-operations classified, and Section 6.6 addressed a set of summary production guidelines based on the classification. Section 6.7 offered an example of an analysed (extract, abstract) pair from the corpus.

The complex operations observed in this corpus are similar to those identified by other researchers (see Section 5.2). However, as with atomic operations, the classification of complex operations presented here is much more detailed than existing work, discussing *triggers* by which some of the sub-operations can be identified. However, because they comprise atomic operations, and more than one sub-operation is often applied to the same unit when transforming an extract into an abstract, it was not always possible to present triggers for every sub-operation. *Reordering* and *merging* were particularly difficult operations for which to identify typical surface forms, although two distinct functions were identified by which to

classify the reordering operation. Merging is a much more intuitive operation and is very difficult to sub-classify because not only are different units and their surface realisations merged into one, but also concepts or ideas which can be inferred from the extract as a whole.

In keeping with findings by Jing and McKeown (1999) that their *sentence combination* operation is very useful, the corpus analysis undertaken in this thesis showed that *merging* is a crucial operation, although it is not restricted to combining information from different sentences. Merging best demonstrates the true nature of abstracting due to the way in which it is applied and the text which results from it. The analysed (extract, abstract) pair in Section 6.7 highlights the complexities of producing an abstract from an extract which can arise when more than one operation is applied to the same unit. Similar to atomic operations, complex operations are not applied in every possible case due to potential impracticalities concerning understanding and length.

Section 6.5 discussed possible future implementations of the sub-operations, describing those which are potentially suitable for automatic processing before the extract is presented to the user of a computer-aided summarisation system. Sub-operations whose triggers can be reliably identified and are not problematic in terms of number of occurrences, such as DELETE: REPORT, REPLACE: PRONOMINALISE and REORDER: COHERENCE, could be automatically identified and the units to which they may be applied could be indicated to the user via a dialogue box or highlighting in the extract. However, not all sub-operations are suitable for this approach due to their subjective nature or issues of frequency, for example, DELETE: DETERMINER, INSERT:

PUNCTUATION and MERGE: RESTRUCTURE, and therefore are only suitable for discussion in summary production guidelines. Guidelines should include information about *all* the operations and sub-operations classified on the basis of the corpus analysis, to ensure that the summariser is aware of a wide variety of useful transformations for creating an abstract from an extract. The guidelines developed from the classification were described in Section 6.6.

The classification of human summary production operations discussed in Chapter 5 and Chapter 6 is assessed in the next chapter. The guidelines developed from the classification are used to transform various extracts into abstracts, which are then evaluated using a theory of local discourse coherence and human judgment.

Chapter 7. Evaluation of human summary production operations

7.1 Overview

The previous two chapters identified and classified a number of human summary production operations resulting from a corpus analysis of extracts and abstracts. On the basis of this classification, a set of guidelines was formulated in order to apply the operations to other extracts to produce abstracts. This chapter evaluates the operations applied to a different set of extracts to assess their generality and their suitability for the task and whether they should be employed by users of a computer-aided summarisation system. The evaluation centres on whether the operations result in an abstract which is more coherent than the extract taken as the starting point. To achieve this, an evaluation method which utilises a discourse theory of local coherence and salience is proposed as an alternative to existing methods deemed inappropriate. However, because the method is developed for this particular evaluation, certain parameters first need to be specified. Evaluating the operations applied to a different set of texts to those in the corpus analysis will establish the generality of these operations, and whether they can be applied to other text domains successfully. The results of the evaluation using the discourse theory are complemented by evaluation using a human judge, and the two evaluations are compared.

Section 7.2 provides a brief overview of existing evaluation methods in summarisation and considers why these are not suitable to evaluate the readability and coherence of human summary production operations investigated in this thesis. An alternative evaluation method using Centering Theory (Grosz, Joshi and Weinstein 1995) is proposed in Section 7.3, and the basic notions of the theory, along with reasons for its suitability, are detailed. Section 7.4 specifies appropriate parameters of the theory for the evaluation of summary coherence, as well as an evaluation metric developed for the task, and presents the evaluation results and discusses them. The chapter finishes with conclusions.

7.2 Evaluation in the field of summarisation

Evaluation is a vital issue in the field of automatic summarisation. If summaries produced by automatic systems are not evaluated, there is no way of knowing how well they perform and consequently how useful they are. To date, the vast majority of evaluation in automatic summarisation has focused on the information content of summaries, either in comparison with a human-produced summary of the same text or as a way of completing another task, such as reading comprehension. This reflects the current preference for automatic extraction as opposed to abstraction, with little consideration for issues of coherence and readability (see Section 3.4). Evaluation is not as explicit an issue in the field of human summarisation, possibly because professional summarisers often work for an abstracting organisation and their summaries are subject to editing by others rather than to evaluation as such.

The human summarisation literature does not address evaluation in the same way as the automatic summarisation literature, focusing instead on the checking and editing of abstracts to ensure that they adhere to the conventions and specifications of the organisation or publication for which they are produced. On the basis of the corpus analysis and resulting classification described in Chapter 5 and Chapter 6, the view is taken that it is not desirable to assess abstracts solely in terms of the guidance given to professional summarisers, for example, regarding tense and voice. This is too restrictive and would result in abstracts being negatively evaluated when they are considered acceptable to a human judge and a human summariser working with guidelines. The discussion in those chapters highlighted the fact that such conventions are not always strictly adhered to, because texts have idiosyncracies in the way that they realise information.

The context factors (Section 2.3) which have most influence on how a particular summary is produced affect the type of evaluation that should be used. For example, if a system is developed with the aim of automatically producing summaries which enable the user to decide whether they want to read the full text, then a standard way to measure its performance is to carry out a classification task using human judges who decide if the summary provides enough information to allow them to do this. Therefore, context factors dealing with content can be considered most important in the overall process and this should be reflected in the evaluation methodology employed. However, this is not the case with the summaries evaluated in this thesis, because the part of the summarisation process under consideration is the third stage of summary production. As mentioned in Section 5.3, the context factors affecting the abstracts investigated in this thesis are mainly output factors, and particularly

form. This means that it is only fair to assess the abstracts in terms of these factors. To evaluate them in a different way, for example, in terms of their informativeness (which relates more to purpose factors such as *situation*, *audience*, *use* and *coverage*) would make the evaluation unfair because the summary production operations applied to form abstracts are not concerned with improving the level of information content with reference to either the extract or the source, but with improving the coherence and readability of the final summary.

7.2.1 Evaluation in automatic summarisation

The area of evaluation in automatic summarisation is extensive, and due to space restrictions cannot be addressed in its entirety here. The nature of the investigation in this thesis means that the vast majority of available means of evaluating summaries produced by automatic systems are unsuitable because they focus on informativeness. A brief review of the main aspects of evaluation in automatic summarisation is given in this section, to show why existing evaluation methods are not appropriate for the evaluation carried out here. Some initial observations, which are relevant to NLP in general as well as to this evaluation, concern the extent of human interaction in the evaluation process. *On-line* evaluation occurs when humans are involved in the process, for example, in assessing how well a summary represents its source, and *off-line* evaluation is where evaluation is conducted automatically, without any human involvement (Hirschman and Mani 2003). This type of evaluation has recently become popular in automatic summarisation, particularly in large-scale evaluation conferences such as the Document Understanding Conferences (DUC) (http://duc.nist.gov/). A well-established way of classifying different types of

evaluation is to distinguish between *intrinsic* and *extrinsic* evaluation, where *intrinsic* evaluations are those which assess a system in itself, and *extrinsic* evaluation involves assessing this system in terms of how useful it is in achieving another task (Sparck Jones and Galliers 1996). The intrinsic-extrinsic distinction in terms of automatic summarisation is discussed in more detail below.

Intrinsic evaluation

Intrinsic evaluation assesses a summarisation system by examining the product of the system, i.e., a summary, either alone or in comparison with its source or with another summary, most often a gold standard produced by a human. This type of evaluation can be on-line or off-line, and there are two main aspects of summaries assessed using intrinsic evaluation measures: *quality* and *informativeness* (Hirschman and Mani 2003). As mentioned above, most evaluations focus on informativeness, because this is deemed the most important aspect of a summary. Quality (or readability) is seen as an added bonus because the main function of a summary is to convey relevant information, and it is pointed out that "it is possible to have beautifully-written but incorrect or useless output" (Mani 2001: 227). However, as this thesis focuses on improving the quality of summaries in terms of coherence and readability, this is precisely the aspect of summarisation which is of interest.

Quality evaluates how well a summary reads, by taking into consideration style via phenomena such as dangling anaphors and connectives, discourse ruptures and grammaticality. An example of intrinsic quality evaluation is Minel, Nugier and Piat (1997)'s FAN protocol. As mentioned in Section 3.4.7, Minel, Nugier and Piat

(1997) conducted experiments to assess the quality of automatic summaries independently of their source, using the FAN protocol which comprised the following criteria: number of anaphora deprived of referents, rupture of textual segments organised by linear integration markers, presence of tautological sentences and legibility of the extract. This type of experiment requires human involvement and so exemplifies on-line evaluation. Another example is an experiment run by Saggion and Lapalme (2000), who used Rowley (1988)'s criteria for acceptability, such as good spelling and grammar, impersonal style, clear indication of the topic of the source document, conciseness, etc. as criteria by which human judges graded summaries. The SEE tool (Lin 2001) allows humans to manually assess extracts for a variety of quality and informativeness phenomena, including coverage, completeness and grammatical fluency. Because the corpus used in this thesis contained abstracts as well as extracts, it is assumed that the abstract which was produced by human post-editing of an extract will fulfil these quality criteria due to the effect of human involvement in the summary production stage. Therefore, it is inappropriate to use an evaluation method which will almost certainly always score the abstracts at 100% because the criteria are very similar to the guidelines used to produce the abstracts. This similarity arises because the guidelines are based on an analysis of summary production operations applied by a human summariser.

Standard readability measures such as the Gunning-fog index (Gunning 1952) and the Flesch-Kincaid index (Kincaid et al. 1975), which assess ease of reading based on average word and sentence length, are examples of off-line intrinsic evaluation. However, these have been criticised as extremely coarse methods due to their simplicity (see Mani (2001)): word and sentence length do not determine a 'good'

summary, and do not give many insights into how or why one summary is of a higher quality than another. Indeed, in the corpus analysis and classification described in Chapter 5 and Chapter 6, there were many instances of merging where sentences in an abstract were made longer than their corresponding ones in the extract so that the text did not seem as 'choppy', and to shorten the abstract overall by incorporating one sentence into another by means of a subordinate clause or other modifier. In terms of summarisation, if extracts are evaluated using these measures, the score will reflect the source text more than the extract, because the sentences in extracts are taken from the source without any modification.

Informativeness is the second and most often addressed kind of intrinsic evaluation discussed in the automatic summarisation literature. It assesses the information content of a summary in comparison with a reference text, either the source text or an *ideal* summary, created by a human. In on-line informativeness evaluations, humans are asked to judge how well the units in a summary reflect, or cover, the content of the source or the content of an ideal summary. However this is time-consuming as judges have to read sources and summaries, and it can be a difficult task. In addition, issues such as subjectivity, even when guidelines are issued, background, mood and tiredness can all affect a human judge's performance in an evaluation task.[44] Brandow, Mitze and Rau (1995) and Saggion and Lapalme (2000) used human judges to assess the informativeness of summaries in relation to their sources, issuing the judges with a scale in order to evaluate them.

[44] Although not explicitly, these issues also come into play in certain off-line evaluations, particularly those where corpora annotated by humans or human summaries are used as reference texts.

One of the most common ways of carrying out off-line intrinsic evaluation of informativeness is to automatically compare a system with an annotated corpus containing texts with units marked for importance, used as a gold standard. Evaluating informativeness automatically means that the evaluation can be done on a much larger scale in much less time (see Orasan, Hasler and Mitkov (forthcoming) for an overview of corpora in summarisation, including for evaluation). Alternatively, the summary can be evaluated against its source, as in CAST (see Section 3.5.4). Examples of measures which can be used to automatically compute the overlap of information present in different texts are *precision*, *recall* and *f-measure*, and the *cosine* similarity (Salton and McGill 1983). For a comparison of recall- and content-based evaluation measures, see Donaway, Drummey and Mather (2000). Related to this, although not necessarily making use of annotated corpora, is the evaluation method ROUGE (Lin and Hovy 2003; Lin 2004), which has become widely used in evaluation since 2004 when it was incorporated in DUC. ROUGE counts the number of overlapping units such as n-grams, word sequences and word pairs between an automatic summary and a human-produced 'ideal' summary. Harnly et al. (2005) automated the *pyramid method* (Nenkova and Passonneau 2004), used in DUC 2005, which addresses the fact that there is no single 'best' possible summary for a text and uses instead the information content from a *pool* of human summaries, recognising that, although each summary is different, they can be equally informative. The pyramid method works by semantically matching content units across summaries and assigning weights to them based on their frequency.

Extrinsic Evaluation

Extrinsic evaluation assesses the summaries produced by a system, and therefore the system itself, in terms of how useful they are in achieving another task. Common tasks used to measure informativeness are relevance assessment,[45] where a summary is evaluated in terms of its relevance to a particular topic, and reading comprehension. These both usually require human participants and are therefore examples of on-line evaluation. Evaluation using **relevance assessment** was used extensively in the TIPSTER SUMMAC evaluation conference (Mani et al. 1998), where humans were asked to perform two tasks to investigate whether summarisation saves time in relevance assessment without impairing accuracy. The first task was to determine whether a topic-focused summary was relevant to a particular topic and the second was to use a generic summary to try to categorise a document into one of five categories. **Reading comprehension** involves humans reading a summary (and in some cases the source as well) and then attempting to answer questions (e.g. Morris, Kaspar and Adams (1992), Orasan, Pekar and Hasler (2004)). The questions are based on the source text in order to assess the extent to which the summary contains important information from the source. To evaluate the usefulness of automatically-produced summaries in relation to allowing the user to decide whether or not to read the full text and in helping the user to write a synthesis of a document, Minel, Nugier and Piat (1997) asked human judges to identify the field of the summary, check the presence of essential ideas, identify parasitic ideas and highlight the logical linking of ideas.

---

[45] *Relevance assessment* in evaluation must not be confused with *relevance assessment* identified by Endres-Niggemeyer (1998) as the second stage of the human summarisation process and discussed earlier in this thesis. The term is retained in this chapter on evaluation because it is an established term in the evaluation literature.

Post-edit measures, commonly used in machine translation, are another option for extrinsic evaluation. They are based on the number of corrections necessary to transform the output of a system into an acceptable state. Whilst an analysis of post-edit measures may seem suitable for the evaluation of changes made when transforming an extract into an abstract, it is impractical due to the complexity of the operations applied during the process. It would be a relatively simple task to measure the number of permanent deletions and insertions made from scratch in the abstracts, but dealing with any other operation, particularly *merging*, which was extremely widespread in the corpus (see Chapter 6), would be time-consuming and difficult on a large scale. During the corpus analysis, a program which indicates the edit distance between sentences was experimented with, but because there were so many complex changes between extracts and abstracts, especially due to merging and reordering, this proved unsuitable.

7.3 Centering Theory: an alternative evaluation method

The most relevant type of evaluation discussed above for this thesis is intrinsic evaluation measuring quality. However, current methods are not particularly suitable for assessing the human summary production operations applied to extracts. One reason for this is the coarseness of methods based on word and sentence length, as sentences were often lengthened rather than left the same length or made shorter during abstract production. The second reason is that methods based on aspects of professional summarisation guidelines such as grammatical fluency, impersonal tone, and other standard 'quality' measures such as the presence of dangling anaphors and

discourse ruptures are insufficient because the comparison is between extracts and abstracts created by post-editing them. This means that all abstracts should fulfil the quality criteria because a human summariser was involved in creating them and was concerned with these aspects of extracts because they affect readability and coherence. Therefore an evaluation method whose criteria are very similar to the guidelines used to produce the abstracts, which are based on the observation of human summary operations in the first place, is insufficient. Thirdly, the operations observed during the corpus analysis are too complex to subject to post-edit measures which count the number of alterations an extract needs to make it acceptable.

In light of this, an objective evaluation method is proposed to assess the 'quality' of extracts, and of abstracts produced by applying human summary operations to these extracts. This method is then compared with the intuitive judgment of a human, as human users of computer-aided summarisation systems will ultimately decide how coherent and readable the extract presented to them is. They will also decide to what extent the extract should be post-edited before it is considered acceptable. The human judge used in the evaluation is not the human summariser who annotated source texts to produce extracts and then created abstracts from them. This ensures that their assessment of the summaries is not biased by any knowledge of the operations applied.

The proposed evaluation method utilises aspects of Centering Theory (CT) (Grosz, Joshi and Weinstein 1995) and is a form of on-line intrinsic evaluation, as a human needs to participate in the process by analysing texts using CT. As Centering Theory attempts to explain local coherence and salience within a discourse, it is a prime

candidate with which to evaluate summaries, where both coherence and salience are issues and the text is short and usually only about one (or two) main topic(s). Hasler (2004a) proved its potential suitability for the evaluation of summaries in a small-scale experiment assessing pairs of extracts and comparing the CT analysis with human judgment.

7.3.1 An introduction to Centering Theory

> "Centering is simultaneously a theory of discourse *coherence* and of discourse *salience*. As a theory of coherence, it attempts to characterize ENTITY-COHERENT discourses: discourses that are considered coherent because of the way discourse entities are introduced and discussed. At the same time, Centering is also intended to be a theory of *salience*: i.e., it attempts to predict which entities will be most salient at any given time." (Poesio et al. 2004: 4)

As mentioned above, Centering Theory (Grosz, Joshi and Weinstein 1995) is a parametric theory of discourse structure encompassing local coherence and salience, the development of which started in the 1980s but was not fully published until 1995. Before Grosz, Joshi and Weinstein fully published their work on CT, several other authors extended and challenged the basic notions and parameters, for example, Brennan, Friedman and Pollard (1987), Gordon, Grosz and Gilliom (1993), Suri and McCoy (1994). Since then, there have been numerous other works concerning Centering Theory, some of the most influential being Kameyama (1998), Walker (1998), Kibble (1999), Strube and Hahn (1999) and Poesio et al. (2004). Whilst the most popular application for Centering Theory in the past has been anaphora resolution (for example, Brennan, Friedman and Pollard (1987)), the suitability of

Centering variations for other computational linguistics tasks has been shown in recent years by its application in natural language generation (e.g., Kibble and Power (1999), Karamanis (2003)) and automatic summarisation (Orasan 2006). Hasler (2004a), Barzilay and Lapata (2005) and Lapata and Barzilay (2005) also prove CT's usefulness in evaluation.

Basic concepts and notions

The main concepts and assumptions introduced in the earliest versions of Centering Theory, (Brennan, Friedman and Pollard 1987; Grosz, Joshi and Weinstein 1995), are presented in this section. As the theory is one of *local* coherence and salience, only two consecutive utterances[46] are considered at any one time (U_n and U_{n+1}). Each utterance in a text introduces a number of *forward looking centers* (*Cf*s), which are NPs referring to an entity. These *Cf*s must be realised explicitly in the utterance. In addition, each utterance except the first has precisely one *backward looking center* (the *Cb*), which is the link between one utterance and the previous utterance in the text. The *Cb* of any current utterance (U_{n+1}) is the most highly ranked *Cf* of the previous utterance (U_n) which is realised in the current utterance (U_{n+1}). In later papers, a weaker version of this constraint is posited, asserting that each utterance has at most one *Cb* (e.g. Walker, Joshi and Prince (1998)).

The *Cf*s are ranked, usually according to grammatical function (see below for more details). The more highly ranked a *Cf*, the more likely it is, in a 'coherent' text, to be the *Cb* of the next utterance, U_{n+1}. The most highly ranked *Cf* of an utterance is

[46] What constitutes an utterance is not fully defined in the original formulation of the theory. However, different possibilities for *utterance* are considered below.

known as the *preferred center* (*Cp*), so the theory predicts that the *Cp* of U_n is most likely to be the *Cb* of U_{n+1}. Centering also states that if an entity within an utterance is pronominalised, it is most likely to be the *Cb*. Table 2 summarises the basic notions of Centering Theory as presented by Kibble (1999), in his discussion of rules and constraints:

Constraint 1	Each utterance has precisely 1 *Cb* (Weaker version: each utterance has at most 1 *Cb)*
Constraint 2	Every element of $Cf(U_n)$ must be realised in U_n
Constraint 3	$Cb(U_{n+1})$ is the highest-ranked element of $Cf(U_n)$ which is realised in U_{n+1}
Rule 1	If some element of $Cf(U_n)$ is realised as a pronoun in U_n, then so is $Cb(U_{n+1})$, (Strong version: if $Cb(U_{n+1}) = Cb(U_n)$, a pronoun should be used)
Rule 2	In transitions, CONTINUE is preferred over RETAIN, which is preferred over SMOOTH SHIFT, which is preferred over ROUGH SHIFT

Table 2: Centering Theory rules and constraints

The relationships between *Cf*s and *Cb*s of utterances result in *transitions* between utterances, which have a definite order of preference, meaning that texts demonstrating certain transitions are considered to be more coherent than those demonstrating others. In the original formulation of the theory, three types of transition are described: CONTINUE, RETAIN, SHIFT. However, following Brennan, Friedman and Pollard (1987) and Walker, Joshi and Prince (1998), amongst others, this discussion splits the SHIFT transition into two: SMOOTH SHIFT and ROUGH SHIFT. CONTINUE is preferred over RETAIN, which is preferred over SMOOTH SHIFT, which is in turn preferred over ROUGH SHIFT. Table 3 presents CT's transitions in terms of the relationship between *Cb*s and *Cp*s. The ordering of transitions reflects the idea that it

is preferable for consecutive utterances to have the same *Cb*, i.e., for the same entity to provide the link between two utterances, and also for the most salient entity (the *Cp*) in one utterance to be the *Cb* of the next utterance.

	$Cb(U_{n+1}) = Cb(U_n)$ or $Cb(U_n)$ undefined	$Cb(U_{n+1}) \neq Cb(U_n)$
$Cb(U_{n+1}) = Cp(U_{n+1})$	CONTINUE	SMOOTH SHIFT
$Cb(U_{n+1}) \neq Cp(U_{n+1})$	RETAIN	ROUGH SHIFT

Table 3: Centering Theory transitions

The interaction and positioning of entities (*Cb*s and *Cf*s) in consecutive utterances, which is encoded by the transitions in the table above, helps to create the impression that a text is about the same entity (in terms of summarisation, the 'main topic'). Consider the following examples taken from Grosz, Joshi and Weinstein (1995),[47] who argue that the same information is present in both, but that (1) is more coherent because it suggests that the discourse is about the same thing (John). The changes in the subject (or ranking) of each sentence in (2) make it difficult for the reader to decide whether this particular example is about John or the store. Examples such as these help to validate Grosz, Joshi and Weinstein's argument for the ranking of the *Cf* list and of transitions. The *Cb*s and *Cp*s have been indicated for ease of reference.

(1) a. **John**$_{[Cp]}$ went to his favorite music store to buy a piano.
b. **He**$_{[Cp],\,[Cb]}$ had frequented the store for many years.
c. **He**$_{[Cp],\,[Cb]}$ was excited that he could finally buy a piano.
d. **He**$_{[Cp],\,[Cb]}$ arrived just as the store was closing for the day.

(2) a. **John**$_{[Cp]}$ went to his favorite music store to buy a piano.
b. **It**$_{[Cp]}$ was a store **John**$_{[Cb]}$ had frequented for many years.

[47] This example is also used by Poesio et al. (2004: 5) to highlight the same point.

c. **He**$_{[Cp],\,[Cb]}$ was excited that he could finally buy a piano.

d. **It**$_{[Cp]}$ was closing just as **John**$_{[Cb]}$ arrived.

(Grosz, Joshi and Weinstein 1995: 206)

In terms of a CT analysis, adhering to the weak version of Constraint 1, (1) displays CONTINUE transitions, because the subject of the first utterance is kept as the *Cp* throughout and is also the *Cb* in the subsequent utterances. (2), on the other hand, displays a RETAIN followed by a CONTINUE, followed by a RETAIN, due to the fact that the *Cb* is the same throughout the utterances, but it is not the same entity as the *Cp*.

CT is a notoriously underspecified theory, and this underspecification has prompted a substantial body of research into optimal parameters depending on text type, language and tasks (see Poesio et al. (2004) for a comprehensive overview and a corpus analysis). Due to its parametric nature, there are a wide variety of possible instantiations of Centering Theory, and parameters need to be specified before the theory can be used for any task. In earlier work (Grosz, Joshi and Weinstein 1995), even the most basic notion of *utterance* is not defined, although an utterance is often considered to be a sentence because it is the simplest option in a theory where the choice of specifications for the various parameters can at times be confusing. This view has been criticised by some researchers, and using different types of clauses as utterances has been proposed as an alternative (for example, Kameyama (1998)), although this too has its critics.

Realisation is another parameter which needs to be defined before the theory can be employed in any task. It is possible to have *direct* or *indirect* realisations of *Cf*s, *Cp*s and *Cb*s. Direct realisations must be coreferential, whilst indirect realisations

encompass other relationships between entities, such as part-whole and set-membership. Direct realisation is easier to employ due to the high number of possibilities of indirect realisations of an entity, although van Deemter and Kibble (2000) argue that there is a widespread lack of awareness of the true nature of coreference. Therefore, in CT analyses, direct realisation may not always be as strict as it should be.

Ranking of the *Cf* list is the third parameter to be specified. This is traditionally taken to be a grammatical/linear preference order: subject > object > others, which is sometimes further split to distinguish direct and indirect objects: subject > direct object > indirect object > others. However, it has been argued that this is inappropriate because the 'topic' of an utterance is not always the subject, and Strube and Hahn (1999) propose *functional centering* where information structure is taken into account and ranking is based on hearer-old and hearer-new information.

In recent years (see Kibble and Power (2000)), Centering Theory has been discussed in terms of *principles*: COHESION (which other researchers seem to have renamed COHERENCE), SALIENCE, CHEAPNESS and the need to ensure that two consecutive utterances have at least one entity in common (this has been termed CONTINUITY). CONTINUITY has been the most used in NLP, with applications in text structuring (e.g. Karamanis and Manurung (2002)) and summarisation (Orasan 2006). However, as these principles are not utilised in the present evaluation, they are not discussed any further here.

7.3.2 Centering Theory for summarisation evaluation

A preliminary experiment conducted by Hasler (2004a) proved that even a simple method which is based on the presence of CONTINUE transitions in an extract could be a useful way of distinguishing the better extract out of a pair. A more extensive study of aspects of CT for coherence and evaluation is described in Lapata and Barzilay (2005), who propose two models of coherence for evaluation. Their findings also show that human judgments regarding coherence agree with their entity-based model which takes into account Centering transitions. In a related paper, Barzilay and Lapata (2005) present local coherence as a ranking problem, similar to work on text structuring in NLG, and describe an automatic method of evaluating multi-document summaries. Their program produces different rankings of the same information based on the transitions between sentences and then a ranking model decides which ranking of sentences produces the most coherent summary.

Although Barzilay and Lapata (2005)'s findings are relevant to the evaluation of the summaries, the method is not wholly appropriate because it deals with different sentence rankings to select the most coherent text. Whereas their work uses Centering to assess the most coherent organisation of a set of sentences in order to produce a summary, the evaluation in this chapter uses the theory to assess the end product, i.e., the summary. Lapata and Barzilay (2005)'s development of *entity grids* for texts, which show the presence of an entity across sentences, also uses elements of Centering Theory, but again this is inappropriate for the current evaluation because much more needs to be taken into account. Lapata and Barzilay are only concerned with entities and not their position in a sentence. This evaluation uses the

whole of CT, aspects of coherence *and* aspects of salience as represented by its transitions, to assess the coherence of summaries.

For this reason, it was decided to adapt Hasler (2004a)'s basic idea of considering a text more coherent in terms of Centering Theory if it displays more instances of a 'more coherent' type of transition than the text it is being compared with. However, rather than considering CONTINUE transitions as the most influential transition for summary coherence, all transitions are considered and a distinction is made between those considered positive and those considered negative for coherence in summaries. A metric is developed to assign coherence scores to summaries based on the transitions they display, reflecting the level of 'harmfulness' of each transition. Section 7.4.3 describes the evaluation metric.

As mentioned above, Centering is a theory of local discourse coherence and salience, both of which are important in summarisation, and work together to produce coherent texts (see Section 7.3.1). Because Centering Theory associates more coherent texts with those displaying certain types of transitions, the coherence of texts can be evaluated by examining the transitions they demonstrate. In Section 7.4.4, a comparison of the sets of transitions in extracts and their corresponding abstracts is used to evaluate the summaries in terms of coherence by selecting the more coherent of the pair. As well as appraising the summaries themselves, this evaluation also assesses the guidelines used by a human summariser to transform extracts into abstracts, and by definition, the classification of operations from which the guidelines were derived. If the abstracts created by applying them are more coherent, then the operations are successful and can be applied across different

domains of texts. This would mean that the guidelines used by a human summariser to complete this task are also useful and can be issued to users of computer-aided summarisation systems to facilitate consistent and high quality post-editing.

7.4 Evaluation of human summary production operations

Having introduced Centering Theory and established its appropriacy for this evaluation task, this section develops CT as an evaluation method to assess the coherence of extracts and abstracts. The parameters best suited to the evaluation of summaries are specified, and an evaluation metric which reliably reflects the effects of certain transitions on the coherence of summaries is detailed. The texts used for the evaluation are also described, as they are different from those investigated in the corpus to ensure a fair evaluation and to test the suitability of the operations on different domains. Once these preliminary matters have been addressed, the results of the evaluation itself are presented and discussed. In addition to Centering, human judgment is also used in the evaluation.

7.4.1 Suitable parameters for summarisation

As discussed previously, Centering Theory is a parametric theory which allows different instantiations depending on various factors, and there have been a number of studies addressing this issue and attempting to find the 'best' instantiation (e.g. Poesio et al.(2004), Mitkov and Orasan (2004)). The specification of parameters adhered to in this evaluation is that which best fits with the investigation undertaken in this thesis and with current practices in computer-aided and automatic

summarisation. However, this is not to say that these parameters are the most appropriate for Centering Theory applications in general. Indeed, different instantiations of the theory can be better for processing or analysing different text types and languages.

The major part of this thesis is not concerned with developing Centering Theory as a fully-fledged evaluation method, and the aim of using CT in this chapter is to take a step in the right direction to find a more objective evaluation than those which currently exist suitable for the current investigation. Due to the number of possible instantiations of the theory, it is best to start simply and address issues as and when they arise; there is no point in making the evaluation using CT more complicated than necessary. Other researchers have taken a simpler approach than the one employed here by just using the *continuity principle*. This is the most basic of the four principles identified by Kibble and Power (2000) and only requires that at least one entity be present in consecutive utterances (see Section 7.3.1).

The first parameter to be specified in the current instantiation is *utterance*. An utterance is considered to be a sentence, as that is the unit used to create the extracts in the corpus via human annotation so it would not be practical to equate an utterance with certain types of clause, for example. In addition, this enhances the generality of the evaluation because it can be applied to summaries produced by other methods and systems, which mainly operate at sentence level (see Chapter 3). It is also not feasible at this stage to use full indirect *realisation* between a *Cb* and a *Cf* due to the fact that it is not easy to define what exactly indirect realisation could encompass. In addition, despite the widespread use of replacement sub-operations in the corpus, the

entities subject to them in one text usually form a coreferential chain, and therefore using only direct realisation should not adversely affect the evaluation. The exception to this is possessive pronouns, which occurred quite frequently in the evaluation texts. An entity such as a country, government or other 'organisation' would be introduced in the first sentence, and then throughout the summary possessive pronouns would be used to discuss slightly different aspects of this entity, but the entity itself could still be considered as the main topic, or very closely related to it. For these reasons, direct realisation between entities, plus indirect realisation where possessive pronouns are involved, is used to identify the *Cb* of an utterance.

The grammatical/linear ranking of the *Cf* list for each utterance is employed: the subject of the sentence is most preferable, followed by the direct object, indirect object, and finally any other noun phrases in the sentence. Because sentences can be complex and do not always comprise only one clause, where there is a main and subordinate clause, the grammatical classes in the main clause appear higher in the *Cf* list, with those in the subordinate clause coming later. This reflects the assumption that the most important information in a sentence tends to be presented in its main clause. In a summary, the most important information from the source text is retained, and in the corpus, there are fewer instances of main clause deletion than the deletion of subordinate clauses, suggesting that the most relevant information does indeed appear in main clauses. Other information which is inserted via relative clauses in the abstracts will usually relate to the information in the main clause rather than being in itself the most relevant part of the sentence. Where there is more than one clause of the same type in a sentence, the *Cf* list ranking reflects the linear organisation of the sentence.

The weak version of Constraint 1 (each utterance has at most one *Cb*) is used, which allows two utterances to display a NO TRANSITION between them, in cases where there is no *Cb*. The instances of NO TRANSITION are split into two groups: NO TRANSITION (NO *Cb*), where there is no entity at all in common with a consecutive utterance, and NO TRANSITION (INDIRECT), which are due to the indirect realisation of an entity. Although NO CB (NO TRANSITION (NO *Cb*) in this thesis) is cited as a common transition, representing on average 20% of transitions found in various corpora (Karamanis et al. 2004), in terms of summaries it is the most damaging transition because it means there is no entity in common between two utterances (it is, essentially, a violation of the *continuity principle*). In this type of short text, which should have one or perhaps two main topics over an average of 6 sentences, for even one pair of utterances to not have an entity in common has a negative effect in terms of coherence and readability. This is the reason for separating the NO TRANSITION transitions into two groups. NO TRANSITION (INDIRECT) is relatively harmless because the reader can easily infer the relationship between the two entities, and the only reason it results in a NO TRANSITION is that indirect realisation (with the exception of possessive pronouns) is not considered suitable for the evaluation of summaries.

7.4.2 Texts used for evaluation

The texts used in the evaluation are different to those in the corpus used for the development of the classification and guidelines to assess whether these resources can also be applied to other domains of texts. Twenty two human-produced extracts of news texts were taken from the CAST corpus (Hasler, Orasan and Mitkov 2003)

which was created using the 2003 annotation guidelines presented in Section 4.4.2. Whilst these texts and the texts used in the corpus analysis are classed as news texts, the CAST corpus contained Reuters newswire and *New Scientist* texts, meaning that there is a wider variety of domains because not all the newswire texts are about science. Therefore, if the abstracts created by applying the summary production operations are evaluated positively, the suitability of the operations and the guidelines for other domains is proved. However, finance articles from the CAST corpus were discarded because these contain lots of figures relating to prices and share indexes, which are notoriously difficult to deal with from a coreference point of view (van Deemter and Kibble 2000), thereby making the evaluation more reliable. A further 3 texts from *New Scientist* which had previously been annotated for summarisation in another project were added to make 25 in total. It was not feasible to have more than 25 extracts to start with because of the way the evaluation was conducted. The 25 extracts were used to produce 3 more sets of 25 summaries each (see below), which was necessary to ensure an effective evaluation. A CT analysis of 100 texts is not a trivial task. These extracts were then transformed into abstracts using the summary production guidelines presented in Section 6.6. As in the corpus used to develop the classification, the extracts adhere to a specific compression rate of 30% of the source text, and the abstracts to 20%.

The source texts of the 25 human-produced extracts were fed into the CAST system, which produced 30% automatic extracts of them. *Term weighting* was selected as the automatic summarisation method because experiments showed that a professional summariser using the CAST system selected this method to produce extracts for post-editing (Orasan and Hasler 2006). This means that the automatic extracts used

in the evaluation are more likely to be similar to those which a user of a computer-aided summarisation system would work with. These automatic extracts were then transformed into abstracts by a human summariser using the summary production guidelines. As a result, 100 summaries in total were analysed using Centering Theory: 25 human-produced extracts, 25 abstracts corresponding to these human-produced extracts, 25 automatically produced extracts, and 25 abstracts corresponding to these automatically produced extracts. The evaluation of the operations applied to both automatically and human-produced extracts means that their success can be measured on extracts produced in different ways.

Using the parameters discussed in the previous section, these 100 summaries were analysed using Centering Theory. In order to perform a CT analysis, first of all, a summary was split into its constituent utterances and *Cf*s of the utterances identified. Based on grammatical ranking, the *Cp* of each utterance was marked and the *Cb* for each utterance (where present) was established. Transitions between consecutive utterances were assigned based on the relationship between the *Cb* and *Cp*. The results of the analysis are presented and discussed in Section 7.4.4. To illustrate the task, the following example is analysed using CT. This text is taken from the evaluation set and was chosen as an example because it is a relatively short text which displays different types of transitions. Utterances are marked with {curly brackets}, *Cf*s with (normal brackets), and *Cp*s and *Cb*s are indicated by sub-script text in [square brackets] and are highlighted in **bold**. The transition holding between two utterances is indicated in between each pair.

U1: {**(Everybody)**$_{[Cp]}$ should be ready for ((Monday)'s national championship game), despite (casualties in ((Saturday night)'s NCAA semifinal battles)).}

NO TRANSITION (INDIRECT)

U2: {**(Jason Terry of (Arizona))**$_{[Cp],\ [Cb]}$ was injured.}

RETAIN

U3: {"**(We)** $_{[Cp]}$ were going to put **(him)** $_{[Cb]}$ in late in (the game)," said (Arizona coach (Lute Olson)).}

ROUGH SHIFT

U4: {"**(He)**$_{[Cp]}$ had played a lot before (that), of course, but when (we)'re protecting (a lead), **(we)**$_{[Cb]}$ like getting (four perimeter guys) in there and that gives (us) (another ball handler), gives (us) (another free throw shooter)."}

RETAIN

U5: {**(Kentucky coach (Rick Pitino))**$_{[Cp]}$ predicted that ((Monday)'s championship game) would be also be physical, in view of ((Kentucky)'s all-out pressure defence) and (**(Arizona)**$_{[Cb]}$'s blazing speed)).} (475997_aut_abstract)

7.4.3 The evaluation metric

For the evaluation to accurately reflect the relation between Centering Theory transitions and coherence, and summarisation, a metric needs to be formulated which represents the positive and negative effects of the presence of certain transitions in summaries. Section 7.3.1 describes how the transitions are derived and the order of preference in which they reflect a coherent text: CONTINUE, RETAIN, SMOOTH SHIFT, ROUGH SHIFT. As mentioned above, there is also the option, depending on which version of the theory is used, for utterances to display a NO TRANSITION, or NO CB transition, which is useful if one wishes to take into account new discourse segments

within a text. This idea of formulating a metric for CT is not novel (see Karamanis (2003) for an overview), but the development of a summary-motivated CT metric is.

In order to reward those transitions which add to the coherence of a summary and to penalise those which negatively affect it, weights are assigned to each type of transition, and these weights are taken into account when evaluating abstracts against the extracts from which they were produced. The traditional order of preference for transitions is kept in this evaluation, although NO TRANSITIONS are treated differently. Because only *direct realisation* is used (except for possessive pronouns; see Section 7.4.1), it is possible to distinguish NO TRANSITIONs due to an indirect realisation of a *Cb* or *Cp* and NO TRANSITIONs due to no *Cb* at all. These transitions are labelled NO TRANSITION (INDIRECT) and NO TRANSITION (NO *Cb*), respectively.

Although the presence of a NO CB in other text types can indicate a new discourse segment and therefore not be considered damaging for coherence, summaries are necessarily shorter than full texts, those produced and evaluated in this thesis being 20% or 30% of their source. Due to the nature of a summary, summary sentences should be about the same topic (or possibly two topics). A sentence which does not contain even an indirect mention of an entity which has been repeated in other utterances throughout is detrimental to the flow of the text as the reader has to stop and work out why that particular sentence is included. In such a short text, the presence of even one or two utterances which do not have an entity in common is noticeable and affects coherence negatively. A preliminary investigation of the evaluation texts indicated that parts of extracts which are viewed intuitively as less coherent can be attributed to the presence of a NO TRANSITION (NO *Cb*). Therefore, in

terms of summarisation, NO TRANSITION (NO *Cb*) is the most damaging to coherence and is weighted accordingly. As mentioned previously, NO TRANSITION (INDIRECT) does not damage a summary's coherence because the reader can easily infer that there is some kind of relationship between two entities. To obtain the average transition score per summary, the weights for each transition identified are added, and then divided by the number of transitions, which is the total number of utterances-1. Table 4 shows the scores assigned to each transition, based on their relation to coherence. According to this metric, the abstract used in the example above (Section 7.4.2) would receive an average transition score of 0.8, based on the scores for its transitions (1 NO TRANSITION (INDIRECT), 2 RETAIN, 1 ROUGH SHIFT) and the number of transitions present (4).

Transition	Weight
CONTINUE	+3
RETAIN	+2
NO TRANSITION (INDIRECT)	+1
SMOOTH SHIFT	-1
ROUGH SHIFT	-2
NO TRANSITION (NO *Cb*)	-5

Table 4: Transition weights for summary evaluation

It should be pointed out here that the numbers themselves assigned to the transitions are subjective. The most important part of the metric is the difference between the scores for each transition because that represents how positive or negative the effect of a particular transition is on a summary. It is clear that the traditional preferred ordering of transitions is reflected in the order of transitions in the table, and that whilst there is a difference in the level of coherence each adds to a summary, none display as great a difference in comparison with the others as NO TRANSITION (NO

Cb). NO TRANSITION (NO *Cb*) is easily identified as being the most detrimental transition to summary coherence.

7.4.4 Results and discussion

This section discusses the results obtained by evaluating the texts described in Section 7.4.2 using the parameters of Centering Theory as stated in Section 7.4.1 and the metric detailed above. The interest is primarily in comparing the average transition scores of pairs of extracts and the abstracts produced from them. This allows an assessment of the extent to which the application of summary production operations using the guidelines improves the coherence of an abstract in comparison with its extract. However, to carry out this comparison and to support the decisions regarding the weighting of summaries proposed above, the transitions present in extracts and abstracts must firstly be identified in order to prove that certain transitions have a greater effect on the coherence of extracts and abstracts. Tables 5 and 6 show the transitions found in the extracts and abstracts used for evaluation. The transition types are split into two categories, those considered to indicate a higher level of coherence and those considered to indicate a lower level, to reflect CT's assumptions about them. It is necessary to present both the number of transitions identified and the percentage of total transitions to allow a fair comparison because the abstracts usually contained fewer sentences than the extracts.

Transitions	Extracts		Abstracts	
	Number	%	Number	%
CONTINUE	37	20.2	23	26.1
RETAIN	37	20.2	27	30.7
NO TRANSITION (INDIRECT)	39	21.3	24	27.3
Total	113	61.7	74	84.1
SMOOTH SHIFT	8	4.4	3	3.4
ROUGH SHIFT	21	11.5	5	5.7
NO TRANSITION (NO *Cb*)	41	22.4	6	6.8
Total	70	38.3	14	15.9
Grand total	183	100	88	100

Table 5: Transitions in human extracts and corresponding abstracts

Transitions	Extracts		Abstracts	
	Number	%	Number	%
CONTINUE	31	19.4	12	16
RETAIN	41	25.6	31	41.3
NO TRANSITION	31	19.4	12	16
Total	103	64.4	55	73.3
SMOOTH SHIFT	11	6.9	9	12
ROUGH SHIFT	29	18.1	11	14.7
NO TRANSITION (NO *Cb*)	17	10.6	0	0
Total	57	35.6	20	26.7
Grand total	160	100	75	100

Table 6: Transitions in automatic extracts and corresponding abstracts

The CT analysis of summaries indicates that, generally, 'better' transitions[48] appear more frequently in abstracts than in the extracts from which they were produced. Where this is not the case with individual transition types, the other 'better' transitions occur more frequently so as to balance the distribution of 'better' and 'worse' transitions. For example, there are more NO TRANSITION (INDIRECT) transitions in automatic extracts than in the abstracts produced from them, but the high percentage of the RETAIN transition in the abstracts balances this out. For ease of reference in the remainder of this section, the group of human-produced extracts and

[48] 'Better' transitions are those which are traditionally viewed as indicating a more coherent discourse. This is in contrast to 'worse' transitions, which are traditionally viewed as indicating a less coherent discourse.

corresponding abstracts is referred to as *set 1*, and the group of automatically produced extracts and corresponding abstracts as *set 2*.

The figures show that, particularly for set 1, there are fewer occurrences of transitions which are considered by CT to indicate less coherent texts in abstracts than in extracts. The most striking differences concern NO TRANSITION (NO *Cb*), which means that there was no entity in common in consecutive utterances. The fact that there were far more of these transitions present in extracts than in abstracts for both sets of summaries indicates that these are indeed the most damaging type of transition to appear in a summary, and justifies the -5 weight assigned to instances of that transition in the summaries. Whilst the percentage difference is not as large for set 2, the fact that the set 2 abstracts displayed no NO TRANSITION (NO *Cb*) transitions at all strengthens the argument that they are not desirable in coherent summaries. Although there is a substantial difference in instances of RETAIN, especially in set 2, because this is considered as only the second best coherence indicator by CT, its weight remains slightly lower than that of CONTINUE.

Table 7 presents the normalised transition scores per summary for all summaries evaluated. Whilst it is unfair to compare the results for sets 1 and 2 due to the difference in information content between them (see below for a fuller discussion of this), a comparison between an automatically or human-produced extract and its corresponding abstract usually shows an improvement in coherence. This means that the summary production operations based on the classification in Chapter 5 and Chapter 6 applied to extracts using the summary production guidelines (Section 6.6) improves the coherence of the resulting abstract in most cases.

Text	Human-based		Automatically-based	
	Extract	Abstract	Extract	Abstract
475968	2.4	2.5	0.6	1
475997	0.1	-2	-0.1	0.8
476016	-0.2	0.8	0.1	1
476032	-1	1.7	0.6	1
476038	-0.3	1.7	1	2
476040	-0.9	2.3	1	2.3
476052	-1.5	-2	0.7	2
476056	-0.7	2	0	0.3
476057	-0.6	1.7	0.8	1.3
476058	0.3	1	0.2	1.3
476059	0	0.8	-1	1.3
476062	0.2	1.5	1.3	1
476074	1.3	2	0	2
476086	0.9	2.3	0.3	0.6
476093	-0.8	0.2	0.2	0
476097	-1.4	2.7	3	3
476143	2.8	2.5	1.8	0
476183	2.2	2	0.7	1.5
476316	-0.7	0.3	1	0.7
476501	-1.1	0.8	0.2	1
476208	-0.3	2.8	0.5	0.3
476520	-1.7	-0.3	-0.3	1.5
sci01	-0.1	0	-0.7	-0.3
sci03	-0.2	3	-0.2	0
sci37	0.6	1.7	1.5	1.3
Total score	**0**	**1.3**	**0.5**	**1.1**

Table 7: Transition scores for all summaries

In total, CT evaluated 78% of abstracts as more coherent than the extracts from which they were produced. 2% (1 instance) of pairs were considered to be of equal coherence, leaving 20% of extracts evaluated as more coherent than the human post-edited abstracts. If only set 1 is considered, the abstracts are considered better in 84% of cases, and the extracts in 16%. For set 2, extracts are considered more coherent than their corresponding abstracts 24% of the time, and 4% are evaluated as demonstrating equal coherence, leaving 72% of abstracts classed as more coherent than their extracts.

There are several reasons for the unexpected results regarding the evaluation of the set 2 abstracts in comparison with set 1. First of all, CT considers representations of the same information (an entity in consecutive utterances). The sentences extracted from source texts automatically are more likely to be repetitive, i.e., to focus on exactly the same aspects of the same topic, due to the way that term-based summarisation works. Such automatic methods are not concerned with the way sentences fit together, or how they repeat information. If repetitive sentences are presented to the human summariser, they are very likely to change these sentences to make them more readable, which will involve some modification of the information or sentences to try to avoid repetition, as advised in the summary production guidelines. However, the resulting abstract is considered less coherent by CT because changes have been made which affect the ranking of elements within the sentences and delete some of these altogether. If the same sentence is taken as the starting point in the extract and in the abstract, it will mean that the abstract is less coherent because the links in subsequent sentences to the initial ordering are disrupted. This illustrates the fact that information content cannot be completely divorced from issues of readability and coherence, which is emphasised by discussions with a human judge who evaluated the coherence and readability of the summaries using intuitive judgment.

Related to these repetitive sentences is the presence of a 'headline' in automatically produced extracts. This headline can be repeated, once with a location and once without, for example:

RUSSIA: Threats get Russians to pay tax – but not much.
Threats get Russians to pay tax – but not much. (476520_aut)

This would be analysed by CT as displaying the transition CONTINUE, although it is obviously not desirable to have such repetition in any kind of summary. This increased the number of CONTINUE and RETAIN transitions in automatic extracts (allowed because of the adoption of the weaker version of Constraint 1, see Section 7.4.1) because if the same sentence is repeated immediately after its first instance, the utterances will be deemed optimally coherent due to the fact that the same entity is mentioned in exactly the same position. Because the transition weights in the evaluation metric reward the more coherent transitions, the average transition score is higher than it might have been had one of these headlines not been included. This should be taken into account in future uses of CT in summarisation evaluation.

The third reason for these results is that the transformation of automatic extracts into abstracts was not as simple as that of human-produced extracts. Where human annotators have the option to include *referred* sentences, which contain information vital to the full understanding of an important sentence such as the antecedent of a pronoun, this is not the case with extracts produced automatically. During the transformations of automatic extracts, the human annotator had to access the source on a number of occasions in order to resolve a pronoun and replace it with the full NP in the abstract so that the text was understandable. This did not happen with human-produced extracts because the information was already contained in the extract itself via a referred sentence. However, improvements such as these are not reflected in the CT evaluation, because it does not consider that such operations may have to be applied, being developed for whole discourses rather than ones which are

produced from parts of a whole. It assumes that the text under analysis will be understandable in terms of this type of information. Issues such as whether a pronoun has an antecedent or not, and whether this has to be remedied to make the text coherent from the point of view of a human reader, are simply not addressed by the theory because it was not designed to deal with such issues. In these cases, the same transition will hold between utterances regardless of whether the subject is an unresolved pronoun or a full NP mention. The limitations of the evaluation using Centering demonstrated by this discussion need to be addressed, and therefore intuitive human judgment was obtained for the coherence and readability of the summaries.

Human judgment

Because the evaluation using CT is a new method, not having been used in this form for this type of task before, human judgments about the summaries' coherence and readability were also obtained in order to validate the evaluation method. Human readers are the ultimate users of computer-aided summaries which have been created by post-editing automatically produced extracts and so their judgment matters. In addition, being based on entity repetition, CT does not take into account aspects of texts such as the use of connectives to signal relations between units or the restructuring of noun phrases, which were identified in the corpus during the classification of human summary production operations.[49] The human judge[50] was

[49] However, it does take into account many of the other sub-operations which concern different realisations of NPs, and the restructuring and reordering of information, and is therefore deemed a suitable theory with which to evaluate the coherence of texts produced using these kinds of operations.

asked to select the more readable and coherent summary out of an (extract, abstract) pair, but was not told which text was which. Table 8 presents the human judgment of the extracts and abstracts, and whether this judgment agrees with the Centering evaluation discussed above. Cases of uncertainty are indicated by a question mark (?), and other relevant notes also appear in the table.

Text	Human		Automatic	
	Extract/Abstract	CT agrees?	Extract/Abstract	CT agrees?
475968	A	Y	E	N
475997	A	N	A	Y
476016	A	Y	A	Y
476032	A	Y	E	N
476038	A	Y	A	Y
476040	E	N	A	Y
476052	A	N	A	Y
476056	A	Y	A	Y
476057	A	Y	A	Y
476058	A	Y	A	Y
476059	A	Y	A	Y
476062	A	Y	E	Y
476074	A	Y	A	Y
476086	A	Y	A	Y
476093	A	Y	A	N
476097	A	Y	E (v. similar)	Same
476143	A	N	E	Y
476183	A	N	E?	N
476316	E	N	A	N
476501	A	Y	A	Y
476208	A	Y	A?	N
476520	A	Y	A	Y
sci01	A	Y	A	Y
sci03	A	Y	?	Abstract
sci37	A	Y	A	N

Table 8: Human judgments on readability of extracts and abstracts

[50] The human judge was a lecturer in the departments of English, Media and Sociology at the University of Wolverhampton. This judge is not the human summariser who produced the corpus for investigation and the abstracts for evaluation.

In total, 82% of abstracts were judged to be more readable/coherent than the extracts from which they were produced. The human judge expressed complete uncertainty about one pair (2%), leaving 16% of extracts evaluated as more coherent than their corresponding abstracts. Similar to the CT evaluation of coherence, examining the two sets separately demonstrates a better evaluation of set 1 than set 2. As mentioned above, automatic extracts do not always focus on the main topic or the same aspects of the main topic, and in certain cases the human judge preferred summaries which gave information about various topics or aspects rather than focusing on the main topic in more detail. Where operations had been applied to delete units about different topics or different aspects of the same topic in the abstract, the extract contained different information. The judge commented that it was impossible to assess the summaries on readability and coherence alone, and that the information present in the summary was an important element in the evaluation. This related to the discussion of the CT evaluation where it was pointed out that information content and coherence/ readability cannot be completely separated. Set 1 abstracts are judged as better in more cases: 92% of the time. Only 2 human-produced extracts (8%) were judged as more readable/coherent than the abstracts created by post-editing them. Automatic extracts are evaluated as better than their corresponding abstracts in 24% of cases, and there is one case of complete uncertainty (4%), meaning that 72% of the set 2 abstracts are considered to be more readable/coherent.

In terms of agreement between the CT evaluation and the evaluation by the human judge, their total agreement was 70%, total disagreement 26%. Two cases (4%) were unable to be compared because in one case the human judge could not select the better text, and in the other case, CT evaluated the abstract and the extract as exactly

the same in terms of coherence. As with the other results, the evaluation of set 1 was better on the whole than that of set 2. The human judge and CT agreed in 76% of cases in set 1, and disagreed in 24%. The summaries which could not be compared in terms of judgment belong to the automatic group, constituting 8% of cases. However, in the instance of the CT evaluation of equal coherence, the human judge commented that it was extremely difficult to decide on a better summary because they were so similar, which is in keeping with the CT evaluation. For set 2, the human judge and CT agreed in 64% of cases and disagreed 28% of the time.

To establish whether the disagreement between the human judge and Centering Theory indicates a problem with the reliability of CT, chi-square was calculated. Chi-square is normally calculated between a set of expected results and a set of observed results, to see whether there is a statistically significant difference. In the case of this particular evaluation, the *expected result* was that the human would agree with the CT evaluation on all pairs of abstracts and extracts. However, Table 8 shows that this was not always the case. The number of agreements and disagreements on pairs between CT and the human judge is the *observed result*. For the purposes of this calculation, the result for text sci03 could not be taken into account because there was no human judgment, but the result for 476097 was interpreted as one instance of agreement and one instance of disagreement because CT evaluated the coherence of the extract and abstract as equal. This gave an observed result of 33 agreements and 8 disagreements on abstracts (out of 41 expected), and 3 agreements and 6 disagreements on extracts (out of 9 expected). Chi-square revealed that there is no statistically significant difference between the CT and human evaluations of pairs, with a confidence level of $p \leq 0.001$. This means that CT is a reliable way of

evaluating coherence because there is no statistical difference between its evaluation of the better summary out of a pair and human judgment on the same pair.

Despite this, the individual cases of disagreement between the human judge and CT need to be addressed. The disagreement illustrates the fact that what a human considers to be a 'readable' text is not necessarily the same as something which is judged more coherent in terms of an 'objective' theory based on entity repetition. As mentioned above, the abstracts observed in the corpus and the abstracts produced for evaluation were subject to operations which did not only concern mentions of entities. However, the comparison of the two aspects of the evaluation correlates with the findings of other researchers (Kibble 2001; Poesio et al. 2004), who claim that CT alone is not always enough to account for the coherence of a text. Indeed, a reader does not assess a text solely on whether an entity is mentioned in consecutive utterances, although this is an important part of their assessment. They also look at aspects such as rephrasing, reordering and conciseness, which were identified in the corpus analysis and classification in Chapter 5 and Chapter 6 and come under the headings *replacement*, *reordering* and *merging*. It also suggests that coherence as measured by CT and coherence/readability as judged by a human are different things, and that perhaps this distinction is something to consider in future. However, CT is still considered a useful tool in evaluating the local coherence of summaries (as chi-square proved), although at this stage in its development as an evaluation method it might be wise to supplement it with other methods to take into account the wide variety of operations a human summariser applies to extracts. Indeed, for a theory which takes into account only part of what humans consider when evaluating texts, CT's evaluation of summary pairs is good.

Despite the limitations of Centering Theory for the evaluation of summaries discussed, its appropriateness for the task was also supported by discussions with the human judge. In most cases the judge found it very difficult to select the best summary out of a pair in terms of readability and coherence *alone*: the information contained in the summary nearly always affected their judgment. Therefore any measure of coherence which is more objective than intuitive human judgment is useful for evaluation.

The generality of the summary production operations

As mentioned in Section 7.4.3, the texts for evaluation are news texts, as in the corpus developed for the investigation, but they are from different domains, including sport, business, politics and science. The fact that the summary production operations improved the coherence and readability of the abstracts produced from extracts of these texts shows that they can be applied to news texts in general, and not just those from the domain of science, which were used in the analysis. The same is also true of the guidelines developed to help a human summariser, or user of a computer-aided summarisation system, consistently apply these operations.

Because the evaluation was conducted on both automatically and human-produced extracts and their corresponding abstracts, it also proved that whatever the starting point, applying human summary production operations to extracts generally results in an improvement in terms of coherence. An interesting observation concerned the type of operations which had to be applied to reach a coherent and readable abstract.

Earlier in this section, it was pointed out that automatically produced extracts tended to contain repetitive sentences. Related to this is the need to apply more DELETE: SENTENCE operations, due to the fact that more whole sentences in the extract contain similar information. Another way of dealing with this repetition was to merge information from several sentences, rather than just two, using MERGE: RESTRUCTURE into one longer abstract sentence. This also involves deletion of redundant clauses and other parts of the merged sentences. DELETE: SENTENCE was also used more frequently because of the presence of sentences which seemed to be unrelated to the main topic, or at least less related than the rest of the sentences, in the automatically produced extracts.

However, regardless of the differences in automatic and human-produced extracts, the same summary production operations were applied to the texts. The only change is that, depending on the way the extract is produced, there is a different level of usage of certain operations, such as DELETE: SENTENCE. This again proves that the operations classified and the guidelines formulated from this classification are useful because they successfully improve coherence and readability regardless of the extract that is transformed, or the way in which it is produced.

7.5 Conclusions

The purpose of this chapter was two-fold. Its main purpose was to evaluate the extent to which applying the human summary production operations identified in Chapter 5 and Chapter 6 improves the coherence and readability of abstracts created by post-editing extracts (Section 7.4). However, to do this, it was necessary, due to the

inappropriacy of existing evaluation methods in the field of automatic summarisation for this particular task, to propose a new evaluation method. Section 7.2 demonstrated inadequacies of existing methods, and therefore justified the development of a suitable alternative, Centering Theory (CT) (Grosz, Joshi and Weinstein 1995), an overview of which was given to indicate its relevance to the task in Section 7.3.

Although the focus of this thesis is not on developing new evaluation methods for the 'quality' of summaries, this was a necessary task to enable a worthwhile evaluation of the investigation conducted in earlier chapters. Therefore, the proposed method is not, perhaps, as accurate as it could be, especially due to the underspecified nature of CT (Section 7.3.1). In order to start developing CT for the evaluation of extracts and abstracts, the simplest specifications of the parameters were used initially, and changed according to any serious problems arising during the analysis. A thorough investigation of all possible instantiations and their effect on the evaluation of summaries is beyond the scope of this thesis, and as the overall results show, the specification used in Section 7.4.1 worked relatively well for an initial attempt. Future work should explore different instantiations of the theory by experimenting with different parameter specifications in order to improve CT for the evaluation of summaries. To ensure a fair evaluation and to prove that the operations can be applied to extracts in other domains, a different set of texts was used for the evaluation (Section 7.4.2). An evaluation metric reflecting the traditional preference order for transitions in CT and their effect on the coherence of summaries was presented in Section 7.4.3.

Two sets of 25 (extract, abstract) pairs, comprising human-produced extracts and abstracts created from them, and automatically produced extracts and their corresponding abstracts, were analysed using Centering Theory. The results showed that most of the time (78%), the application of human summary production operations improves the coherence of the abstract when compared to the initial extract (Section 7.4.4). When the operations did not improve the coherence, this was due to the way that information was presented in the extracts, particularly due to the nature of term-based automatic summarisation methods, and indicated that it is not possible to completely divorce the evaluation of coherence from the information contained in a text. Human judgment was also obtained for the (extract, abstract) pairs evaluated by CT, which again proved that most of the time (82%) summary production operations achieve their purpose: they improve the readability and coherence of extracts whilst transforming them into abstracts. This means that the operations identified during the corpus analysis and the guidelines used to facilitate the summary production process are also appropriate to be employed by users of a computer-aided summarisation system. It also means that, if the necessary automatic methods perform adequately, the sub-operations suitable for (semi-)automation could be implemented in a computer-aided summarisation system in the future.

The agreement between the human judge and the CT evaluation was 70%. The reason for disagreement was that humans take into account much more than mentions of entities in consecutive utterances when assessing readability and coherence. This finding correlates with other research into CT, which states that additional aspects need to be taken into consideration for a full account of coherence (Kibble 2001; Poesio et al. 2004). However, the difference between the CT

evaluation and the human judgment of summary pairs was not statistically significant, meaning that CT is a reliable method by which to evaluate coherence.

Chapter 8. Concluding remarks

8.1 Overview

The main aim of thesis was to identify ways of improving the quality, in terms of coherence and readability, of extracts, which can then be applied in the field of computer-aided summarisation. This chapter considers the extent to which this has been achieved, and how. The aims and contributions of the thesis are revisited in Section 8.2. Section 8.3 reviews the content of the thesis chapter by chapter, and directions for future work are addressed in Section 8.4.

As mentioned in Chapter 1, this thesis provides original contributions to research in the fields of human summarisation and computer-aided summarisation, and within computer-aided summarisation in the area of evaluation. The contributions to computer-aided summarisation are also, by virtue of association, contributions to the field of automatic summarisation. The thesis focused on improving the readability and coherence of extracts by transforming them into abstracts, thereby addressing a major shortcoming of existing automatic summarisation methods and systems. By grounding the research within the framework of Endres-Niggemeyer (1998)'s three stages of human summarisation (*document exploration*, *relevance assessment* and *summary production*), the main aim of the thesis was justified.

8.2 Aims and contributions revisited

This section reiterates the main aim of the thesis and how it was achieved by completing a number of smaller, interlinked goals. The **main aim** of the thesis: identifying ways of improving the quality, in terms of coherence and readability, of extracts, which can then be applied in the field of computer-aided summarisation, has been met. A **classification of human summary production operations** has been formulated and evaluated as useful for improving the coherence and readability of extracts.

These operations, five general classes split into *atomic* and *complex* operations, each comprised a number of sub-operations. Additionally, the atomic operations work together to produce the complex ones. Two atomic operations, *deletion* and *insertion* were identified. *Deletion* involves removing a unit from an abstract either permanently, or non-permanently as part of a complex operation. *Insertion* is where a unit is placed in an abstract sentence, either from scratch, or from somewhere else in the extract as part of a complex operation. These two operations work together in the complex operations *replacement*, *reordering* and *merging*. *Replacement* involves the deletion of one unit and the insertion of another unit in its place. *Reordering* occurs when a unit is deleted from one place and inserted in another place. *Merging* happens when information from more than one unit is presented together as one unit, and often involves elements of all the other general classes of operation.

The operations can be recognised by certain *triggers*, or surface forms. Although the function of the units to which the operations are applied is also vital, such recognisable

forms are necessary because this research is concerned with computer-aided summarisation. In addition, not all users of a computer-aided summarisation system such as CAST will be experts in linguistics or summarisation, therefore a classification which is relatively simple in terms of grammatical classes and other aspects of the English language is necessary. The complex operations become increasingly difficult to further classify into sub-operations, *merging* being the most difficult. However, *merging* was found to be the operation which best captures the essence of an *abstract* in comparison with an *extract*. During *merging*, information is taken from different places in an extract and presented together, often using a different realisation, in the abstract. This operation does not just involve the merging of units realised in the extract, but also uses world-knowledge to infer information related to units in the extract.

Certain sub-operations were identified as potentially suitable for future implementation in a computer-aided summarisation system, and others were identified as suitable for inclusion only in a set of guidelines issued to a human summariser post-editing the output of such a system, mainly due to the frequency of occurrence of their triggers or of the sub-operation itself. The classification is represented in the following table, which gives the type and general class of operation, the sub-operations which are considered as instances of that class, and the units within the summaries or the types of change which are affected by them:

Atomic summary production operations		
Deletion	DELETE: SENTENCE	Complete sentences
	DELETE: SUB_CLAUSE	Subordinate clauses
	DELETE: PREP_PHRASE	Prepositional phrases
	DELETE: ADVERB	Adverb phrases
	DELETE: REPORT	Reporting clauses and speech
	DELETE: NP	Noun phrases
	DELETE: BE	The verb *be*
	DELETE: DETERMINER	Determiners
	DELETE: FORMAT	Specially formatted text
	DELETE: PUNCTUATION	Punctuation
Insertion	INSERT: CONNECTIVE	Connectives
	INSERT: FORMULAIC	Formulaic units
	INSERT: MODIFIER	Modifiers
	INSERT: PUNCTUATION	Punctuation
Complex summary production operations		
Replacement	REPLACE: PRONOMINALISE	Pronominalisation
	REPLACE: LEXEME	Lexical substitution
	REPLACE: RESTRUCTURE_NP	Restructuring of noun phrases
	REPLACE: NOMINALISE	Nominalisation
	REPLACE: REFERRED	Referred sentences
	REPLACE: VP	Verb phrases
	REPLACE: PASSIVISE	Passivisation
	REPLACE: ABBREVIATE	Abbreviations
Reordering	REORDER: EMPHASISE	Emphasising information
	REORDER: COHERENCE	Improving coherence/readability
Merging	MERGE: RESTRUCTURE	Restructuring clauses/sentences
	MERGE: PUNCTUATION-CONNECTIVE	Punctuation and connectives

Table 9: Classification of human summary production operations

In order to arrive at this classification, a number of smaller goals, also original contributions, had to be achieved. A corpus analysis was conducted on **a corpus developed specifically for this task**. The corpus contained 43 (extract, abstract) pairs which allowed the investigation of exactly how a human summariser transforms an extract into an abstract. The extracts for this corpus were produced by a human annotator using a **set of annotation guidelines for extraction** formulated to facilitate quality and consistency in the task. It was important to use a human annotator rather than automatic methods to produce the extracts because it means they are of a higher quality in terms of informativeness, allowing the focus of the investigation to be wholly on the *summary production* stage of summarisation. These guidelines were based on those used in a similar annotation task in 2003, but with a number of amendments to make them suitable for this particular annotation.

Because the summary production operations observed in the corpus were identified as a means of improving the coherence and readability of extracts, it was essential to apply them to different texts to assess whether this was really the case. To make this possible, the classification had to be translated into **a set of summary production guidelines**. These guidelines are novel, as there are no other existing guidelines regarding improving the quality of news extracts by human post-editing. As well as being necessary for the testing of the operations in this thesis, they can be issued to users of a computer-aided summarisation system to help them consistently improve the extracts produced by applying operations which are known to be used by a human summariser.

An evaluation of a different set of (extract, abstract) pairs to which the operations identified in the corpus were applied using the summary production guidelines proved that these operations do improve the coherence and readability of extracts when they are post-edited and transformed into abstracts. An examination of current evaluation methods in automatic summarisation showed that no completely suitable methods were available for this particular evaluation, and so a new evaluation method was necessary. **Centering Theory (Grosz, Joshi and Weinstein 1995) was developed as an evaluation method** to measure the coherence of extracts and abstracts in order to ascertain the 'better' summary. Suitable specifications for the necessary parameters were determined and a metric was formulated which reflects both the traditional preference order of transitions relating to how they affect coherence, and the most damaging transitions for coherence in summaries. It should be noted that as the main purpose of this thesis was not to investigate evaluation and develop fully-fledged evaluation methods, there is still room for improvement in the CT evaluation method.

The evaluation proved that, in most cases (78%), the operations applied to extracts using the summary production guidelines improved the coherence of the resulting abstract in comparison with its extract. To further validate this, a human judge intuitively evaluated the same summaries for coherence and readability and, similar to the CT evaluation, preferred the abstracts 82% of the time. This means that the classification of summary production operations identified in the corpus is useful, and can be successfully applied to extracts produced by a computer-aided summarisation

system during the post-editing stage, providing that the user is also issued with the guidelines formulated to facilitate the task.

This thesis has therefore achieved its main aim: reliable ways of improving the coherence and readability of extracts which can be employed in computer-aided summarisation have been established. To summarise, the **original contributions** of this thesis, presented in the order in which they were achieved, are:

1. A set of guidelines to annotate source texts for important information which results in extracts for a corpus of (extract, abstract) pairs.
2. A corpus of (extract, abstract) pairs to enable the analysis of human summary production operations.
3. A corpus-based classification of human summary production operations which can be successfully applied to extracts to transform them into abstracts by improving their coherence and readability. These can also be successfully applied in computer-aided summarisation during post-editing.
4. A set of summary production guidelines derived from the classification which can be issued to users of a computer-aided summarisation system.
5. The development of Centering Theory (Grosz, Joshi and Weinstein 1995) as a new evaluation method to assess the operations due to the unsuitability of existing evaluation methods for the task.
6. An evaluation of the coherence and readability of abstracts produced using the summary production operations and, by default, the guidelines issued to a human summariser and the operations themselves.

8.3 Review of the thesis

This section gives a brief review of the work completed in each chapter in order to achieve the main aim of the thesis.

Chapter 1 introduced the topic of research for the thesis. A brief background to summarisation was given and the need to address issues of coherence and readability in the fields of automatic and computer-aided summarisation was stated. The chapter also identified the main aim of the thesis and specified how this could be achieved.

Chapter 2 presented an introduction to summaries and summarisation via a discussion of *context factors* (Sparck Jones 1999; Tucker 1999), which were used to describe issues that need to be taken into account when creating any kind of summary. The chapter also addressed the field of *human summarisation*, reviewing major work. The three stages of the summarisation process: *document exploration*, *relevance assessment* and *summary production* (Endres-Niggemeyer 1998) were presented as a general model for automatic and computer-aided summarisation.

Chapter 3 addressed automatic summarisation, providing an overview of major work. Automatic abstracting and automatic extracting were distinguished, and the positive and negative aspects of the summaries produced by each were detailed. The recently developed concept of *computer-aided summarisation* (Orasan, Mitkov and Hasler 2003), was presented as a feasible alternative to fully automatic methods. In combination with

Chapter 2, this chapter provided the necessary background to and justification for the remainder of the thesis.

Chapter 4 presented the annotated corpus of news texts exploited in Chapters 5 and 6 and a set of guidelines used for the annotation. A review of available guidelines showed that none were suitable for this particular task and so the guidelines had to be formulated based on those used in a similar annotation of the CAST corpus (Hasler, Orasan and Mitkov 2003). It was argued that using human annotations rather than automatic methods achieved a higher quality result of the first two stages of summarisation, allowing the focus in the corpus analysis to remain on *summary production*. This chapter also helped to justify the investigation of the summary production issues of coherence and readability.

Chapter 5 introduced the classification of human summary production operations resulting from the analysis of the corpus developed in Chapter 4. Two general types of operation were identified: *atomic* and *complex*, where complex operations are made up of the atomic ones. The chapter focused on *atomic* operations, and two classes were distinguished: *deletion* and *insertion*, each comprising a number of sub-operations which can be identified by certain *triggers*. To contextualise this investigation, a brief review of existing work related to the operations humans use to create summaries was offered, and differences between them and the work in this thesis were established.

Chapter 6 examined cases of the *complex* summary production operations from the corpus: *replacement*, *reordering* and *merging*, and pointed out the increasing difficulty of identifying sub-operations within these. As well as being the most difficult to classify in terms of sub-operations and triggers, *merging* was identified as the operation which best captures the essence of an *abstract* as opposed to an *extract*. A discussion of possibilities for future automation of the sub-operations was also offered, along with an example to highlight the complexities of summary production. In addition, a novel set of summary production guidelines formulated from the classification was presented.

Chapter 7 evaluated the extent to which the classification of summary production operations and the guidelines formulated from it are useful to a human summariser post-editing extracts. The guidelines were used to produce abstracts from extracts for a different set of texts to those in the analysed corpus to assess the suitability of the application of the operations to other domains. These extracts were produced automatically and by a human annotator to show that the operations work on both types of extract. Centering Theory (CT) (Grosz, Joshi and Weinstein 1995) was developed as an evaluation method based on a review of existing evaluation in automatic summarisation. Human judgment was also obtained, which further strengthened the CT evaluation. The evaluations proved that the summary production operations can be applied successfully to extracts to improve their coherence and readability. They can therefore be employed by users of a computer-aided summarisation system during post-editing, providing the guidelines formulated for the task are also issued.

8.4 Possible directions for future work

During the process of investigation in this thesis, a number of possible directions for future work became apparent. These are addressed in the order in which they appear in the chapters of the thesis.

Chapter 6 discussed the operations classified during the corpus analysis in terms of potential implementations within a computer-aided summarisation system. Certain operations were identified as being suitable, providing that the automatic methods involved perform well enough and do not negatively affect the summarisation process. The most obvious next step is to attempt some automation of the appropriate summary production operations. This would involve an investigation into the performance of automatic methods such as coreference resolution, which could help in processing NPs in the extracts in order to determine which can be safely transformed, to establish whether their incorporation would be detrimental or not.

In addition, and very importantly for computer-aided summarisation, if such automation is feasible, the effect of presenting such a potentially large amount of information on the screen to the user would need to be evaluated. Previous experiments with the CAST system have shown that sometimes users find information intended to help them very confusing if too much is presented at once (Orasan and Hasler 2006). As one of the main aspects of the computer-aided summarisation system is to enable users to produce summaries in a user-friendly environment, the user is very important indeed. Their

opinions should be sought as to how much additional information presented to them really helps them to post-edit extracts effectively.

A second idea for future research developed from the adaptation of Centering Theory for the evaluation of summaries (Chapter 7). It has already been mentioned that there is room for improvement in the method because its development was not the main focus of this thesis. Research projects resulting in long technical reports and PhD theses have been published which attempt to identify the best instantiation of CT either in general (Poesio et al. 2004) or for a particular task (Karamanis 2003). It would be interesting to investigate different specifications for the parameters of the theory to see whether this affects its performance as an evaluation method for summaries. Other researchers have found that different instantiations result in different frequencies of occurrence of NO CB transitions, which were found to be the most damaging for summary coherence in this thesis, making this a useful future step.

A more general path for further research is to consider other types of texts in terms of post-editing. Although the classification of human summary production operations was derived from a corpus of news texts, it would be interesting to explore other types of texts for which they could be used. Scientific articles are another popular text type used in automatic summarisation, therefore this would be a sensible option. This would also allow a comparison between summary production (or editing) operations utilised by professional abstractors and users of a computer-aided summarisation system, because

the majority of existing research into professional summarisation considers scientific articles/research papers as source texts.

Appendix I: Annotation Guidelines - marking important units of text for summarisation

(Author: Laura Hasler; based on 2003 guidelines, last updated March 2007)
Classification: essential, important, referred, removed.
Unit: sentence (essential, important, referred); clause/phrase (removed).
Length restriction: 30%. Mark 15% of the source sentences as *essential*. Mark a further 15% as *important* sentences.

General advice

1. Prior to annotation, read the whole text to familiarise yourself with it and get a feel for what the text is "about" (one or maybe two main topics).

2. Ensure that the annotation is done in one intensive period (relatively easy in this case, as the domain is news texts and the documents are not too long), as sporadically annotating a file can lead to the annotator having to re-read the document for familiarisation several times.

3. Comment on troublesome cases of annotation and then discuss them with other annotators to decide upon the best solution to tackle them in future.

Marking essential information

Identify the main topic of the text. A good indicator of this is usually the 'headline' and first sentence.

Mark sentences which refer to this topic and give important information about it – mark those you see as **essential**, keeping as near to the 15% length restriction as possible.

Do not include sentences which contain the same information as others you have marked. If there are several sentences containing similar information, pick the most appropriate one. This is not necessarily the most descriptive or the longest sentence, but that which most succinctly expresses the essential information.

Example: prefer i) to ii)
i) *The diversion was part of a plan to make the project feasible without the technical help of Hungary.*

ii) *While talks between the Hungarian and the then Czechoslovak governments dragged on, Binder's company was preparing a huge 25 km-long, 300-metre (900 foot) wide channel to feed the Danube waters into the massive Gabcikovo turbines.*

Indicate sentences which should be included if other sentences are marked. Do not just mark them as very important, but also indicate that there is a relationship between them. For example, if you mark a sentence containing an anaphoric reference, you should also mark the sentence containing its antecedent.

Example: If ii), containing *it*, is marked as important, then i), with the noun phrase *the Festival*, also has to be marked.
i) *For film lovers the Festival's the place to be in September.*
ii) *It grows from strength to strength each year…*

Do not mark sentences concerning sub-topics unless they directly influence the main topic (or present new, essential information on it) and do not repeat information – the main topic is essentially what the text is about.

Do not mark sub-headings, as they tend to be subjective, are not generally relevant to the main topic and cannot be taken out of context of the full text.

Do not include examples, including constructions starting with *e.g.*, *for example*, *such as*, *like*, *for instance*, etc.

Do not include tables and figures/illustrations such as graphs, diagrams etc.

Direct speech presents an interesting case. It sometimes contains relevant information but not always. These units should be treated with care and neither marked nor omitted purely because they are speech. The annotator should decide in individual cases. It is important to distinguish between speech and other text in quotation marks, which may be important.

Example:
But its "killer app" is reinventing how the software industry works.

Marking segments for removal

Once the very important/essential sentences have been marked, indicate those segments of these sentences which are not vital and can be **removed**. The examples of units given below should not be indiscriminately marked for removal. Consider each case carefully in its given context. So, **within the sentences already marked as essential**:

Mark irrelevant subordinate clauses as removed.
Example: the following "which" clause could be viewed as unimportant.
The US government's Nuclear Regulatory Commission and the private Electric Power Research Institute will also pay into the pot~~, which totals $93.5million~~

Mark text in brackets and text occurring between dashes as removed, unless vital to the main topic. This kind of formatted text should not be marked as removed solely on the basis of its formatting – the function of the text is much more important.
Example: the following bracketed text would be considered important because it is an abbreviation which will replace a noun phrase later in the text:
A Poverty Reduction and Growth Facility (PRGF)…It is intended that PRGF-supported programmes…

Mark examples (see above) as removed.

Mark constructions such as *a spokesman said*, *it was claimed*, *she told The Guardian*, *he explains* as removed, unless the point of the text is to offer differing opinions/reports/findings etc.

Mark phrases which elaborate on information, such as *in addition to…*, *due to…*, *compared to…* as removed unless vital to the main topic.

Adjuncts containing dates, times, places, etc. can be important or irrelevant depending on the text. They should not always be marked for removal, but should be considered as individual cases. The decision to remove them lies with the annotator. If they are not vital to the main topic, mark these adjuncts (not just single words within sentences) as unimportant as removed.

Example:
Units of the Fleet, sailing south from Gibraltar ~~on Monday, 29 March, 1982,~~ were carrying nuclear weapons.

Other instructions

Once you have completed this, if the final amount of text marked is substantially below 15% of the full text, try to add more units which you consider essential to bring the percentage up.

Then go through the text again and use the same instructions to mark sentences which you would classify as **important**, again keeping as close to 15% as possible.

Comment on the annotation – any problems, indecisions, observations etc.

Appendix II: A selection of (extract, abstract) pairs from the corpus

e02-ljh

Extract

A devastating tropical cyclone has ripped through north-eastern Australia, injuring people and destroying homes with gusts of up to 290 kilometres per hour. The first cyclone, named Larry, reached maximum (Category 5) intensity at about the time of landfall. It hit the coast at the town of Innisfail at 0700 Eastern Standard Time (AEST) on Monday. By 2000 AEST it was about 400 km inland and had downgraded to a Category 2 storm. no serious injuries have been reported. About half the houses in Innisfail have been damaged. Millions of dollars worth of sugar cane and banana crops have been destroyed. A tent city will now be erected to house the homeless. it could turn out to be the most destructive cyclone ever to hit Queensland.

Abstract

A devastating tropical cyclone ripped through north-eastern Australia injuring people and destroying homes with gusts of up to 290kmph. Possibly the most destructive cyclone to ever hit Queensland, Larry reached maximum (Category 5) intensity when it hit Innisfail at 0700 Monday. By 2000 it was 400km inland and downgraded to Category 2. No serious injuries were reported, although about half of the houses in Innisfail were damaged and millions of dollars worth of sugar cane and banana crops destroyed.

h01-ljh

Extract

World's second face transplant performed in China. Experts predict the number of these operations will rise rapidly as centres around the world gear up to perform the procedure. Thirty-year-old Li Guoxing received a new upper lip, cheek and nose from a brain-dead donor to repair injuries sustained after an attack by a black bear. He was reported to be in a stable condition and taking liquid food following the 13-hour surgery on Thursday at Xijing hospital in Xian. The surgery scar will not be obvious but there is a difference in the donor's and recipient's skin colour, so that will be noticeable. Guoxing is reportedly happy with his new face, which will be improved by further treatment over time. it will be two months until they are sure that Guoxing has not rejected the new tissue. Rejection of the transplanted facial tissue could have life-threatening consequences, and the immunosuppressant drugs used to keep this from happening can make a person more prone to certain cancers.

Abstract

In the world's second face transplant, Li Guoxing, 30, has received a new upper lip, cheek and nose from a brain-dead donor following a bear attack. He is in a stable

condition after the 13 hour surgery at Xijing hospital, Xian, China, and is happy with his new face, which will be improved by further treatment. Doctors need 2 months to be sure that the new tissue is not rejected. Rejection can be life-threatening and the immunosuppressants used to stop it can increase the risk of cancer. The scar will not be obvious, but the donor's and recipient's skin colour is different. The number of face transplants is predicted to rise rapidly, as centres world-wide prepare for the procedure.

h02-ljh

Extract

UK's bird tests may be missing flu virus. Last week, scientists found H5N1 bird flu for the first time in the UK. DEFRA stated that all wild birds tested so far were negative for flu, so it was unlikely to be widespread. Suspicions have been raised because DEFRA's tests revealed none of the ordinary flu that ducks and geese normally carry. The problem may have been DEFRA's method of collecting samples. DEFRA told WWT samplers to moisten a sterile swab on a stick with saline, take a faecal sample from the bird, then put the swab back in its dry plastic tube. The tubes were kept at refrigerator temperature and taken to the testing laboratories the next day. swabs must be immediately immersed in a saline or preservative solution, and frozen quickly. DEFRA has not done large-scale flu surveys before. DEFRA declined to comment on whether its sampling method would deliver intact virus to the testing labs. H5N1 was most likely carried to the UK by migratory ducks, which could have spread the virus to wintering grounds all over the country.

Abstract

Last week, H5N1 bird flu was found for the first time in the UK, but DEFRA stated that all wild birds tested so far were negative, so it was unlikely to be widespread. Suspicions have been raised because their tests revealed no ordinary flu either - so by now the virus could have spread. The problem may have been the sampling method. Samplers moistened a sterile swab with saline, took a faecal sample, then put the swab back in its dry plastic tube, kept it at refrigerator temperature and took it to laboratories the next day. However, swabs should be immediately immersed in a saline/preservative solution and frozen quickly. DEFRA has not done large-scale flu surveys before and declined to comment.

h03-ljh

Extract

Chernobyl reactor number 4 was ripped apart by an explosion on 26 April 1986. Last September, the IAEA and the WHO released a report. Its headline conclusion that radiation from the accident would kill a total of 4000 people was widely reported. In a report this week, Ian Fairlie and David Sumner, two independent radiation scientists from the UK, say that the death toll will in fact lie somewhere between 30,000 and 60,000. They accuse the IAEA/WHO report of ignoring its own prediction of an extra 5000 cancer deaths in the less contaminated parts of Ukraine, Belarus and Russia, and of failing to take account of many thousands more deaths in other countries. Zhanat Carr, a radiation scientist with the WHO in Geneva, says the 5000 deaths were omitted because the report was a "political communication tool". She also accepts that the WHO estimates did not include predicted cancers outside Ukraine, Belarus and Russia. Fairlie and Sumner's accusations are backed by other experts.

Abstract

Last September, the IAEA/WHO released a report on the explosion of Chernobyl reactor number 4 on 26 April 1986, concluding that radiation from the accident would kill a total of 4000 people. However, this week two independent radiation scientists say that the death toll will in fact be between 30,000 and 60,000. They accuse the IAEA/WHO report of ignoring its own prediction of an extra 5000 deaths in less contaminated parts of Ukraine, Belarus and Russia, and excluding deaths in other countries. Their accusations are backed by other experts. The WHO admits the deaths were omitted because the report was a "political communication tool", and that estimates excluded areas outside Ukraine, Belarus and Russia.

new-sci-B7K-25

Extract

The making of an insect's nervous system

Some recent experiments suggest that in insects a small group of pioneer cells may provide a labelled "streetmap" along which the developing neurons can navigate by using a simple set of directions to locate their specific targets. In 1976 Michael Bate now working at Cambridge University drew attention to the embryonic development of nerve pathways linking these ganglia with the budding limbs. Since Bate's discovery, many other examples of pioneer cells have been found which established a network of criss-crossing links between major parts of the nervous system and the developing limbs. One of the most interesting recent developments had been the discovery that the very first muscle cells to differentiate also erect a kind of scaffold of pathways along which the later "motor" nerves migrate in order to locate their muscle targets. as the nervous system develops and the distances that must be crossed become greater, the tracks traced by individual nerve cell processes become increasingly elaborate, and cease to follow single pioneer pathways. pioneer cells can be stained selectively with certain monoclonal antibodies indicating that their surface membrane contains a unique type of "marker" molecule which could provide one of the cues traced by later developing axons. It is not yet clear whether this "choose-and-follow" mechanism occurs in other species. However, the bundling, or "fasciculation" of axons travelling to similar destinations is a widely observed phenomenon.

Abstract

The making of an insect's nervous system.

Recent experiments suggest that in insects a small group of pioneer cells may provide a labelled "streetmap" along which the developing neurons can navigate by using a simple set of directions to locate their specific targets. As the nervous system develops and the distances that must be crossed become greater, the tracks traced by individual nerve cell processes become increasingly elaborate, and cease to follow single pioneer pathways. Pioneer cells can be stained selectively with certain monoclonal antibodies indicating that their surface membrane contains a unique type of "marker" molecule which could provide one of the cues traced by later developing axons. It is not yet clear whether this "choose-and-follow" mechanism occurs in other species. However, the bundling, or "fasciculation" of axons travelling to similar destinations is a widely observed phenomenon.

new-sci-B7K- 44

Extract

the southern hemisphere has experienced extreme climatic events in the past year. Over the Pacific Ocean, the most significant change was the sudden onset of El Nino in May last year. it arrived five months early and was one of the strongest events of this kind observed in the present century. Widespread flooding, loss of life, property damage and economic disruption has occurred across one-third of Ecuador. Since October last year, South Africa has been in the grip of serious drought, with agricultural production expected to decline by at least 700 million Rand. The pattern of the connected climatic events in the southern hemisphere is nothing unusual, but its timing is and so is the strength of the phenomenon. The temptation to link the change with the disruption to equatorial atmospheric conditions produced by the El Chichon eruption in April is strong A related possibility is that the El Nino of 1982 was not the usual primary pulse of that event, hut an unusually strong secondary pulse. Such secondary pulses have been observed at about this time of year after very strong El Ninos in 1957, 1956 and 1972 - but this 1982 "secondary" would have been stronger than the "primary" it belongs to. the return of the overall pattern in the early 1980s comes as little surprise to some researchers, who have been warning for several years that a new drought cycle was due.

Abstract

The southern hemisphere has experienced extreme climatic events in the past year, from widespread flooding to serious drought. Over the Pacific Ocean, the most significant change was the sudden onset of El Nino in May last year. It arrived five months early and was one of the strongest events of this kind observed in the present century. The pattern of the connected climatic events is nothing unusual, but its timing and strength is. The temptation to link the change with the disruption to equatorial atmospheric conditions produced by the El Chichon eruption in April is strong. Another related possibility is that the El Nino of 1982 was not the usual primary pulse of that event, but an unusually strong secondary pulse, similar to those in 1956, 1957 and 1972. The return of this overall pattern in the early 1980s comes as little surprise to some researchers, who have been warning for several years that a new drought cycle was due.

new-sci-B7L-72-ljh

Extract

Roy Herbert recalls some early scientific influences.

THE VICTORIAN red-brick building which housed the grammar school I went to still stands. It had laboratories on the ground floor, with bronzed taps and bunsen burners. In these labs I made my first acquaintance with physics. I never got an answer to my question, a reasonable enough one, what was magnetism? The question placed me in danger of a box over the ear from the broken-nosed physics master He was the junior physics master. The senior one was a fearsome, mustachioed amateur cello player who addressed every class in a terrifying bawl. He would silence noise by yelling threats of appalling punishment. The physics master, long after John Cockcroft and Ernest Walton had successfully split the atom, was still dictating to his pupils, "Matter can neither be created nor destroyed". The only solace for all this was the weekly reading of Modern Boy Modern Boy included real science. It illustrated up-to-date transport. I still remember the Krooms. Modern Boy also printed fiction.

They were scaly creatures, walking on two feet and they were out to conquer the Earth and every living creature on it. the most desperate measures were necessary to provide even the most slender chance of defeating the malevolent hordes. Fortunately there was a scientist about who had invented a time machine. Fortunately he had at least one dauntless and resourceful boy to assist. The best idea they had was to travel back to the Wars of the Roses and bring two armoured knights into the age of the Krooms. we owe a great debt to these doughty fighters of the past in banishing the menace of the Krooms forever. It seemed to me at the time that the teachers of science at school were, if not actually off their trolleys, a trifle on the demented side and undoubtedly strangers to coolness. The wounds they inflicted left me with a permanent scepticism about science and the progress of research. That, I think, is a healthy enough attitude.

Abstract
Roy Herbert recalls some early scientific influences.
THE VICTORIAN red-brick building which housed the grammar school I went to still stands. It had laboratories on the ground floor, with bronzed taps and bunsen burners. In these labs I made my first acquaintance with physics. I never got an answer to my first physics question, a reasonable enough one, what was magnetism? The question placed me in danger of a box over the ear from the broken-nosed junior physics master. The senior one was a fearsome, mustachioed amateur cello player who addressed every class in a terrifying bawl and silenced noise by yelling threats of appalling punishment. Long after John Cockcroft and Ernest Walton had successfully split the atom, he was still dictating to his pupils, "Matter can neither be created nor destroyed". It seemed to me at the time that the teachers of science at school were, if not actually off their trolleys, a trifle on the demented side and undoubtedly strangers to coolness. The only solace for all this was the weekly reading of Modern Boy. Modern Boy included real science, illustrated up-to-date transport, and also printed fiction - I still remember the Krooms. But the wounds my physics masters inflicted left me with a permanent scepticism about science and the progress of research. That, I think, is a healthy enough attitude.

sci09done-ljh

Extract
ME sufferers are sick of doctors not taking them seriously and have made a cinema ad to win support for their case. WE ALL like to be taken seriously, especially when we are feeling ill. the erosion of the traditional divisions between sickness and health, which have led to, among other things , the medicalisation of "anxiety". Until the launch of the so-called "minor tranquillisers" in the early Sixties, anxiety was seen as an inevitable everyday experience. Doctors are now increasingly reluctant to prescribe tranquillisers for fear of being sued by patients who become addicted to them. What this means is that good old fashioned self-reliance could become a new health vogue of the Nineties. According to one theory, held by some psychiatrists, patients may find doctors increasingly unsympathetic to their plight. They may be neurotics or malingerers or both. Patients claiming to have Myalgic Encephalomyelitis (ME) may come into the latter category. Symptoms include fatigue, night sweats, weight loss and exhaustion. While some doctors recognise the condition, others question its existence, attributing problems to causes such as depression. The ME Action Campaign is waging a major campaign to validate ME as a disease. Featuring a teenage girl whose life is totally disrupted by ME, it is

designed to win public support for ME sufferers. The Action Campaign is also seeking support from medical journalists.

Abstract
ME sufferers are sick of doctors not taking them seriously and have made a cinema ad to win support for their case.
Good old-fashioned self-reliance could become a new health vogue of the 90s, as doctors become increasingly reluctant to prescribe tranquillisers. According to one theory held by some psychiatrists, patients claiming to have ME may find doctors increasingly unsympathetic to their plight. ME symptoms include fatigue, night sweats, weight loss and exhaustion. While some doctors recognise the condition, others question its existence, attributing problems to causes such as depression. The ME Action Campaign is waging a major campaign, featuring a teenage girl whose life is totally disrupted by ME, to validate ME as a disease and to win public support for sufferers. The Action Campaign is also seeking support from medical journalists.

sci15done-ljh
Extract
The camel is without doubt one of the natural world's most remarkable forms of transport. Underpinning all these accomplishments is the camel's efficient use of metabolic fuel. three zoologists led by M. K. Yousef of the University of Nevada gathered recently in Australia to investigate the camel's fuel economy. The researchers' aim was simply to measure the amount of oxygen consumed by the camels under various conditions. As each beast ambled at its own pace along a farm track, the team collected its breath with the help of a gas-tight mask and a weather balloon. Over the same distance, people use two and a half times more energy (per kilogram of their bulk) than the average camel. It uses less than three-quarters of the energy expected of a mammal of its size. Most animals increase their fuel consumption in proportion to the extra baggage they are carrying, but not the camel. Zoologists have yet to establish the precise mechanisms behind the camel's fuel economy. One candidate is the leg musculature. Another candidate is the unique stepping gait.

Abstract
The camel is without doubt one of the natural world's most remarkable forms of transport. Underpinning all these accomplishments is its efficient use of metabolic fuel. Three zoologists led by M. K. Yousef of the University of Nevada recently investigated the camel's fuel economy. Their aim was simply to measure the amount of oxygen consumed under various conditions. Over the same distance, people use 2 1/2 times more energy. Camels use less than three-quarters of the energy expected of a mammal of their size and do not increase their fuel consumption in proportion to the extra baggage. Zoologists have yet to establish the precise mechanisms behind the camel's fuel economy, but candidates are the leg musculature and the unique stepping gait.

sci13done-ljh
Extract
Rapidly growing environmental fears have made climate research more important than ever. Britain is playing a prominent role in a complex satellite which could unveil the mysteries of the Earth's resources. climate research has gained new

prominence, chiming with the rise in environmental concerns the world over. For once, the UK is in the good books of the European Space Agency (ESA) for our support of the European Remote Sensing (ERS) programme. ERS-1 is currently under test in Toulouse, scheduled for launch next October. ERS-1 has five instruments which will scan the Earth at microwave wavelengths to monitor ocean, ice and land resources. So for oceanographers, used to data returned from ships in commercial shipping lanes, ERS-1 will give a totally global picture of the seas. The Active Microwave Instrument radar system, built by Marconi Space Systems, is significant because it can be used over land or ocean: it will return vital information about how the wind interacts with waves. How the wind affects the ocean and modifies the heat exchanged with the atmosphere is not well understood, but is of paramount importance to weather forecasters who can not wait for ERS-1 to begin operations. The Rutherford Appleton Laboratory in Oxfordshire has provided the Along-Track Scanning Radiometer, an infrared instrument which will measure sea surface temperatures down to 0.3C. This greater accuracy will allow the heat transfer between the oceans and the atmosphere to be better computed. Commercially, the ATSR data will allow fishing vessels to "home in" on the edges of sea currents where fish congregate to feed on nutrients brought up from the ocean floor. The way in which the atmosphere interacts with the oceans remains a last great unknown in Earth science. After refusing to sanction the project, the UK is now providing funding. At present, only France has not given its go-ahead for ERS-2 funding.

Abstract
Rapidly growing environmental fears have made climate research more important then ever, and the UK is in the good books of the European Space Agency for its support of the European Remote Sensing (ERS) programme. ERS-1 is scheduled for launch next October and has five instruments which will scan the Earth at microwave wavelengths to monitor ocean, ice and land resources. The Active Microwave Instrument radar system, built by Marconi Space Systems, is significant because it can be used over land or ocean: it will return vital information about how the wind interacts with waves. How the wind affects the ocean and modifies the heat exchanged with the atmosphere is not well understood, but is of paramount importance to weather forecasters, who can not wait for ERS-1 to begin operations. The Rutherford Appleton Laboratory in Oxfordshire has provided the Along-Track Scanning Radiometer, an infrared instrument which will measure sea surface temperatures down to 0.3C. This greater accuracy will allow the heat transfer between the oceans and the atmosphere to be better computed. The research should give insights into the way in which the atmosphere interacts with the oceans, which remains a last great unknown in Earth science.

new-sci-B7L-70-ljh

Extract
YOU COULD be forgiven for not having heard of the Treaty of Tlatelolco. The non-aligned nations have just dragged it on the stage by backing an Argentine call for the withdrawal of all nuclear weapons from the Falklands. The Treaty of Tlatelolco was signed on 20 December, 1967, and ratified on behalf of Britain by the Wilson government on 11 December, 1969. the signatories agree to respect fully the statute of denuclearisation of Latin America. The express aim of the treaty is to keep the whole region, including the Falkland Islands and the adjoining seas "forever free from nuclear weapons". the concern over Britain's importing nuclear weapons into an

area hitherto free of them was real and not fictitious. Britain certainly has various nuclear weapons there. Can we be sure they will never be used? Units of the Fleet, sailing south from Gibraltar on Monday, 29 March, 1982, were carrying nuclear weapons. Some of the ships which left Portsmouth on Monday, 5 April, 1982, were carrying nuclear weapons. some of the nuclear weapons were lifted back by helicopter and other boats. Stenor Inspector and Stenor Sea-Search were despatched to retrieve nuclear devices from HMS Sheffield and HMS Coventry. Even more elusive have been facts about retrieving nuclear depth charges from lost helicopters. All carried nuclear depth charges. Heaven knows what pollution of the ocean is occurring in the form of emission of radionuclides, and building up in the various food chains in which plankton play a part. If negotiation about sovereignty does not take place, there will inevitably be a "Falklands re-play".

Abstract
The non-aligned nations have just dragged The Treaty of Tlatelolco onto the stage by backing an Argentine call for the withdrawal of all nuclear weapons from the Falklands. The Treaty was signed on 20 December, 1967, and ratified on behalf of Britain by the Wilson government on 11 December, 1969. Its express aim is to keep the whole of Latin America, including the Falkland Islands and the adjoining seas, "forever free from nuclear weapons". Britain certainly has various nuclear weapons in the Falklands. Units of the Fleet, sailing south from Gibraltar on Monday, 29 March, 1982, were carrying nuclear weapons, as were some of the ships which left Portsmouth on Monday, 5 April, 1982. Some of these were lifted back by helicopter and other boats. Stenor Inspector and Stenor Sea-Search were also despatched to retrieve nuclear devices from HMS Sheffield and HMS Coventry. Facts about retrieving nuclear depth charges from lost helicopters, which can cause pollution of the ocean due to emission of radionuclides that build up in food chains containing plankton, are more elusive.

new-sci-B7L-64-ljh

Extract
Cave life by David Culver, Harvard UP.
THE animals that live in caves, David Culver believes, should be of particular interest to biologists. "Because of their simplicity, cave communities in many cases are close to the assumptions of various ecological and evolutionary models. models can be tested more completely," Has the study of cave life settled any important questions in evolution or ecology? The answer is always the same. It is always no. So we might ask whether the study of cave life has told us anything at all about the great questions of evolution and ecology. Again the answer is no. The most we learn, from caves, is that the questions are difficult. All this is hardly Culver's fault. The same difficulties are to be found in all other parts of evolutionary ecology. But he did advertise his book saying that caves were an exception to the general difficulties, so he can hardly object if that is how his book is judged. It is an authoritative review of the research that has been done on a fascinating fauna.

Abstract
Cave life by David Culver, Harvard UP.
THE animals that live in caves, Culver believes, should be of particular interest to biologists because their simplicity allows various ecological and evolutionary models to be tested more completely. However, the study of cave life does not settle any

important questions in evolution or ecology. Nor does it tell us anything at all about the great questions of evolution and ecology. These general difficulties of evolutionary ecology are hardly Culver's fault. But he did advertise his book saying that caves were an exception to the general difficulties, so he can hardly object if that is how it is judged. However, it is an authoritative review of the research that has been done on a fascinating fauna.

new-sci-B7L-74-ljh

Extract

the annual consumption of polythene is between 5 and 10lb per head. On 24 March, 1933. at ICI's laboratories at Winnington, Cheshire, Reginald Gibson and Eric Fawcett began a series of high pressure experiments with gas-liquid mixtures. These were difficult to undertake, because of the problems of pressurising gases with the primitive equipment available and maintaining the pressure. In the first experiment Gibson and Fawcett used ethylene and benzaldehyde. part of the reaction tube, through which ethylene had been admitted to the vessel, was coated with a white waxy solid. the two chemists obtained sufficient of the material to indicate that it was a polymer of ethylene. Polythene was virtually forgotten for the next two and a half years. It was a polymer so unlike the polymers known at the time that no one could envisage a use for it. And we couldn't make it consistently. The very idea of a polymer of ethylene went against conventional wisdom. Perrin subsequently took charge of the high-pressure research. In December 1935, Perrin decided to look at ethylene. On 19 December, 1935, Perrin produced several grams of polythene. During 1936, he and colleagues worked out the details of how to make the reaction occur consistently and explored the commercial possibilities. a chemist who had recently worked for a company involved in the manufacture of submarine telephone cables recognised its similarity to the material used to sheathe such cables. polythene was superior to the natural material. The world's first polythene plant came on stream the day that Hitler invaded Poland. polythene played an important part in radar during the Second World War. when ethylene became available on a large scale from oil refining processes the price of the polymer dropped.

Abstract

On 24 March 1933, at ICI's laboratories at Winnington, Cheshire, Reginald Gibson and Eric Fawcett began a series of high pressure experiments with gas-liquid mixtures. After the first experiment, in which Gibson and Fawcett used ethylene and benzaldehyde, part of the reaction tube through which ethylene had been admitted to the vessel was coated with a white waxy solid. They obtained sufficient of the material to indicate that it was a polymer of ethylene: polythene. Polythene was then virtually forgotten for the next two and a half years, mainly because it was so unlike existing polymers that no-one could envisage a use for it and it couldn't be made consistently. In 1935, Perrin took charge of the high-pressure research and on 19 December 1935, he produced several grams of polythene. During 1936, he and colleagues worked out the details of how to make the reaction occur consistently, and explored the commercial possibilities. A chemist recognised its similarity to the material used to sheathe submarine telephone cables - tests showed that polythene was superior, making large-scale production worthwhile. The world's first polythene plant came on stream the day that Hitler invaded Poland.

new-sci-B7L-54-ljh

Extract

It took four years to develop a CAD system for shoe making. The software, which was written by CAD Centre in Cambridge, gives pictures and patterns of finished shoes on a high-quality monitor screen. It uses the Centre's Polysurf program which can define free-form surfaces. The program also has an art "module" which allows a designer to draw lines and paint up to 16 million colours on the screen. Pattern flattening is done by additional mathematics on a specially written part of the Clarks program. Shoemaking with CAD starts with a last marked with a grid and clamped in a digitizer. A pointer is touched against each point of intersection on the grid and the digitizer records the position. the computer displays a wire-frame model - hundreds of rectangular planes joined together in a last-like shape. The planes are smoothed out, the surface is coloured and hidden areas are removed from view. The last displayed can be rotated, enlarged and panned. The designer draws on the last by using a digitizing pen on a tablet. When the basic style is agreed, the computer draws a two-dimensional pattern of it complete with stress points marked. No one thought the patterns would work. they were stitched into shoes that required 3 per cent less leather. Problems with a pattern can be sorted out before detailed styling begins. Clarks' system will draw details, change colours, experiment with textures and vary the shading at will. It is simple to drop unsuccessful ideas because the computer retains the last's shape and specific designs. Potential and actual savings which have been identified include: shorter lead times, presently as long as six months; producing better, more up-to date styles; and shoes that fit better. CAD had, and still has, its doubters at Clarks.

Abstract

A CAD system for shoemaking at Clarks, developed by CAD Centre in Cambridge, gives pictures and patterns of finished shoes on a high-quality monitor screen using the Centre's Polysurf program which can define free-form surfaces. It also has an art "module", allowing a designer to draw lines and paint up to 16 million colours on the screen. Pattern flattening is done by additional mathematics on a specially written part of the program. The process starts with a last marked with a grid and clamped in a digitizer. A pointer is touched against each point of intersection on the grid, the digitizer records the position and the computer displays a wire-frame model. The planes are smoothed out, the surface coloured and hidden areas removed from view. The last can be rotated, enlarged and panned. The designer draws on the last using a digitizing pen on a tablet. When the basic style is agreed, the computer draws the 2-D pattern with stress points marked. No one thought the patterns would work - they were actually stitched into shoes that required 3% less leather. Potential and actual savings include shorter lead times, better, more up-to date styles and better-fitting shoes. However, CAD had, and still has, its doubters at Clarks.

Appendix III: Summary production guidelines for the computer-aided summarisation of news texts

(Author: Laura Hasler; last updated: March 2007)
Field: computer-aided summarisation
Domain: news texts
Purpose: to create an abstract from an extract via post-editing (summary production) operations
Users: human post-editors of computer-aided summarisation systems

Description
These guidelines are designed to facilitate consistency and quality when post-editing computer-aided extracts of news texts to transform them into an abstract. They are concerned with style, conciseness, coherence and readability. They are also concerned with further text reduction, as an abstract should be the best possible representation of relevant information in the smallest possible space.

Length restriction: 20% of the source text for the 2006-7 task

1. General strategy and information
• Prior to post-editing, read the extract to familiarise yourself with it and make sure you understand what it is about. Identify the main topic of the extract – the 'headline' and first sentence are often good indicators.

• If there are any doubts about meaning, or any other uncertainties, refer to the source text for clarification - remember that there is *always* the opportunity to access the source text if you feel that it may help.

• Read the extract through a second time, and try to identify units which can be altered to produce a concise abstract which is coherent and easy to read without deleting relevant information.

• Ensure that the abstract is created in one intensive period, as sporadically editing a text can lead to a lack of accuracy.

• Having completed the annotation, check through it to see if there are any mistakes or
additional changes you wish to make. Check that the abstract reads well.

• The operations described in the guidelines should only be applied if you are sure that they do not change the meaning of the summary in any way, and if they are appropriate in the particular transformation of the extract under consideration at any given time.

• There are *always* exceptions to the guidelines, and cases where they do not need to be applied: you must use your expertise and the context as well as the guidelines in order to produce the best abstract.

• The guidelines should be taken as *guidance* rather than strict instructions, and indicate rather than exhaustively describe operations which *can* be applied to produce a good abstract. The basic idea is that if you need to edit the computer-aided extract, do so using the operations listed here.

2. Specific summary production operations

These operations are based on a classification of human summary production operations used to transform human produced extracts into abstracts. They cover most cases of operations in a corpus of news texts developed specifically to analyse these types of operation.

It is difficult to present these guidelines in the order in which the operations might be applied to a text, because during summary production the extracts are not necessarily dealt with sentence by sentence. Therefore the guidelines are divided into headings which cover broad classes of operations: *deletion*, *insertion*, *replacement*, *reordering* and *merging*, starting with the 'simplest'. The sub-operations of these broad classes can be applied to the same unit in a text at virtually the same time.

The first two classes of operations discussed are labelled *atomic* operations, because the last three classes, labelled *complex*, comprise them. *Atomic* operations can be applied either alone or as part of the *complex* operations.

Atomic operations

Deletion

Complete sentences

If a distinction is made between *essential* and *important* sentences in the extract (i.e., if the 'more important' sentences are indicated in some way), delete irrelevant *important* sentences rather than *essential* ones.

Delete complete sentences which are not directly related to the main topic of the extract.

Delete complete sentences which introduce examples, additional detail or explanations, or that repeat information which is better presented elsewhere.

Delete complete sentences offering speculation or opinion, unless one of the main points of the extract is to present speculation or opinion.

It may also be necessary to delete sentences which have been wrongly selected by automatic methods for inclusion in the extract because they do not contain information about the main topic of the source text.

Example:
According to one theory, held by some psychiatrists, patients may find doctors increasingly unsympathetic to their plight. ~~They may be neurotics or malingerers or both.~~

Subordinate clauses
Delete subordinate clauses which contain non-essential information such as examples, explanations, temporal and spatial information, and information which can be inferred from elsewhere in the text.

Subordinate clauses can be recognised by their subordinators and *wh*-elements, and their lack of a finite verb.

Do not delete subordinate clauses which are necessary to avoid ambiguity.

Example:
A chemist ~~who had recently worked for a company involved in the manufacture of submarine telephone cables~~ recognised its similarity to the material used to sheathe such cables.

Prepositional phrases
Delete prepositional phrases which add too much detail and introduce redundant information by referring to something previously mentioned in the extract.

These PPs include temporal and spatial adjuncts and postmodifiers of NPs, and can be recognised by the presence of a preposition immediately after a noun.

Do not delete PPs which are necessary to avoid ambiguity.
Example:
Last week nuclear scientists ~~from Britain, Japan, and several West European nations~~ met in the US to haggle over a new joint programme of research…

Adverb phrases
Delete adverb phrases (a phrase introduced by an adverb or with an adverb as the head) and adverbs which add unnecessary information, such as additional detail, to the extract. This is similar to the deletion of PPs mentioned above, but an adverb is the trigger by which to recognise the adverb phrase.

Do not delete adverb phrases which are necessary to avoid ambiguity.

Example:
Potential and actual savings which have been identified include: shorter lead times, [***presently*** *as long as six months*]*…*

Reporting clauses
In indirect speech and reporting, delete the reporting clause plus *that* (or other word introducing the reported clause) when it is not necessary to identify the person associated with the speech or claim being made, or when this person or organisation has already been mentioned in the text. This is not restricted just to speech or

thought, but to any kind of report such as findings, claims, etc. which follow the same syntactic pattern as reporting indirect speech. In some cases, the retained text will need to be reformulated to ensure grammaticality.

Do not delete reporting clauses when it is important that the speaker is known, for example in extracts about conflicting opinions.

Example:
~~*We know with some confidence that*~~ *if greenhouse gasses continue to be emitted in their present quantities, we will experience unprecedented rates of sea-level rise.*

Noun phrases
Delete modifiers of nouns, such as adjectives, and other parts of noun phrases (including heads where appropriate) which are not necessary to avoid ambiguity or are repeated elsewhere in the extract in order to shorten the text. Whole NPs can be deleted where these serve to further explain or modify a unit in the text.

Do not delete parts of NPs when they are necessary to avoid ambiguity.

Example:
It hit ~~*the coast at the town of*~~ *Innisfail at 0700* ~~*Eastern Standard Time (AEST) on*~~ *Monday.*

Determiners
If you are struggling to reach a predetermined compression rate, it is possible to delete determiners from different places in the extract in order to reach the desired compression. Definite determiners are most suitable for deletion without causing incoherence. This is particularly useful in the first sentence of news texts as it also transforms the sentence into a 'headline' (as in the example below).

Example:
Britain is among ~~*the*~~ *frontrunners as tomorrow's supercomputers take shape.*

The verb *be*
Similar to determiners, *be* can be deleted in the last stages of trying to reach a specified compression rate and also functions to transform text into a 'headline' (as in the example below).

Example:
Britain ~~*is*~~ *among the frontrunners as tomorrow's supercomputers take shape.*

Specially formatted text
Text in brackets, or following punctuation such as a colon, semi-colon or dash can *sometimes* be suitable for deletion, but not always. Care should be taken with these units as they are often considered unsuitable for inclusion in an abstract, but the corpus analysis on which these guidelines are based proved different. If the text in these formats provides additional detail or redundant material which can be inferred from elsewhere, they can be deleted, but they should not be deleted as a matter of course just because of their formatting as vital information can be lost.

Do not delete bracketed text which introduces an abbreviation (an acronym or initialism) repeated later in the text.

Example:
Over the same distance, people use two and a half times more energy ~~(per kilogram of their bulk)~~ than the average camel.

Punctuation
Extracts can contain misplaced punctuation such as commas which can be safely deleted. There may also be cases where you prefer not to have certain punctuation in a certain place in a unit due to style. The most common deletion should be of commas.

Example:
A devastating tropical cyclone has ripped through north-eastern Australia~~,~~ injuring people and destroying homes with gusts of up to 290 kilometres per hour.

Insertion
Connectives
Insert connectives to make the text flow better by explicitly signalling coherence relations. Connectives include coordinators, subordinators and adverbs functioning as conjuncts and are more commonly inserted in sentence-initial position, but can also be inserted mid-sentence.

Example:
The tubes were kept at refrigerator temperature and taken to the testing laboratories the next day. [***However,***] *swabs must be immediately immersed in a saline or preservative solution, and frozen quickly.*

Formulaic units
Insert standard patterns or units typical of the style of news texts in order to make relationships between units more explicit. This also usually involves some reformulation of the following text to ensure grammaticality.

Example:
[***And we should bear in mind that***] *however we may feel about overseas aid, the requirement for agricultural research is a matter of tens of millions of dollars, not billions.*

Modifiers
Insert modifiers in order to avoid ambiguity between noun phases or to clarify meaning, and to save space, by introducing a concept mentioned later in a concise form earlier in the abstract.

Example:
The four volumes of this work [***published so far***] *are to be accompanied soon by a further five…*

Punctuation
You may consider it desirable from a style point of view to insert punctuation to make the text easier to read. Commas are the most suitable type of punctuation to insert for this purpose.

Example:
Faced with these problems, computer scientists began[**,**] *in the mid 1970s*[**,**] *to experiment with using many processors in a single machine to work in parallel on a single problem.*

Complex operations
Replacement
Pronominalisation
Replace noun phrases with pronouns after the first mention to avoid repetition and to shorten the abstract.

Replace the object of a sentence with the relative pronoun *which* when it is merged to avoid repetition.

Example:
It uses the Centre's Polysurf program which can define free-form surfaces. ~~*The program*~~ [***It***] *also has an art "module"…*

Lexical substitution
Replace lexical items with alternative lexical realisations to avoid repetition and to shorten the abstract. Alternative lexemes can be related to the original by relations such as synonymy, hypernymy, hyponymy and metonymy, and they can be more or less specific than the original as required. This can include the replacement of person names with names of organisations they represent.

Example:
~~*Shoemaking with CAD*~~ [***The process***] *starts with a last marked with a grid and clamped in a digitizer*.

Restructuring of noun phrases
Postmodifiers of NPs should be restructured as premodifiers by deleting their preposition or relative pronoun and relocating what is left of them immediately before the head. This functions to shorten the text and make it more concise.

Example:
Where the two configurations meet, a disturbance occurs in the ~~*orderly structure of the polymer*~~ [***polymer structure***]*…*

Nominalisation
Nominalise sentences or clauses which it is possible to nominalise without disrupting the overall flow of the abstract. This is useful for shortening the text, but must be done in conjunction with merging information from different units in order to result in a grammatical unit/sentence.

Example:
Fairbanks found that [*when sea level first began to rise as the ice sheets melted, 17,000 years ago, it did so at a rate of about 4mm per year*]. --- [***The initial sea-level rise at this point***] ***was about 4mm per year.***…

Pronoun expansion
Expand pronouns on their first mention if they have not previously been realised as a noun phrase. This operation usually involves *referred* sentences, which are included in an abstract only because information contained in them is necessary to understand another unit in the extract.

Example:
~~*Prawns are not small fry.*~~ *According to the Shellfish Association of Great Britain,* ~~*they*~~ [***prawns***] *are our most popular seafood.*

Verb phrases
Prefer the simple present and simple past tenses where possible – this makes the abstract easier to read as well as usually shortening the text.

Replace verbs with alternative forms after other operations are applied to ensure grammaticality.

Simplify certain structures involving verbs to make the abstract easier to read and to save space.

Example:
Poor roads means that machinery, tools and other inputs which the settlers need ~~*are often not getting*~~ [***often do not get***] *through.*

Passivisation
Passivise sentences in order to delete irrelevant subjects or subjects not mentioned elsewhere in the text, and to shorten the text.

Passivise sentences in order to depersonalise them, for example, if they contain a pronoun such as *we* which is not easy to resolve.

Example:
~~*…McElroy is preparing*~~ *for private companies to take over the reigns of the weather craft;* --- [***Preparations are being made***] *for private companies to take over the reigns of the weather craft;*

Abbreviations
Abbreviate full versions of words which have standard abbreviations, such as *kilometres* (*km*).

Replace words with standard symbols where possible, for example per cent with %.
Retain common acronyms and initialisms, such as WHO and UK in the abstract, and replace full versions with them.

Reordering

Emphasising information

Place sentences containing more important information earlier in the text, and relegate those containing less important information to later. What the text is about is very important in this case.

Place units containing the most important/relevant information at the beginning of a sentence to emphasise its importance.

Example:
THE MOVE by Britain's Department of Industry (DoI) to increase greatly its research spending is in danger of misfiring. The department has admitted to a substantial underspend in two important areas or research in the financial year that has just ended. [*Organisations that apply for the DoI's research cash say that bureaucratic procedures and shortage of staff are to blame.*] ***If spread over the whole of the department's research budget, the shortfall would mean that between 25 and 40% of the cash allocated for 1982-83 has remained unspent.***

THE MOVE by Britain's Department of Industry (DoI) to increase greatly its research spending is in danger of misfiring. The department has admitted to a substantial underspend in two important areas or research in the financial year that has just ended. ***If spread over the whole of the department's research budget, the shortfall would mean that between 25 and 40% of the cash allocated for 1982-83 has remained unspent.*** [*Organisations that apply for the DoI's research cash say that bureaucratic procedures and shortage of staff are to blame.*]

Improving coherence and readability

Place sentences and other units about the same topic, or containing the same kind of information about the main topic, together so that the text is easier to read and does not seem 'choppy'. However, bear in mind that there may be introductory and concluding sentences which do not fit this overall pattern.

Place sentences which introduce or conclude some aspect of the extract at the beginning and end of the abstract, respectively.

Example:
Faced with these problems, computer scientists began in the mid 1970s to experiment with using many processors in a single machine to work in parallel on a single problem. There are practical difficulties with parallel computers. [*Meiko was set up in 1985 by members of the group which developed the transputer chip for Inmos. The transputer contains a processor, memory, and communications links on a single chip, which makes it an ideal building block for multi-processor computers.*] ***When users want to improve the performance of their machine, they do not have to throw it away and buy a newer model. Instead, they can buy more processors and add them to the existing machine.***

Faced with these problems, computer scientists began in the mid 1970s to experiment with using many processors in a single machine to work in parallel on a single problem. ~~*There are practical difficulties with parallel computers.*~~ ***When users want to improve the performance of their machine, they do not have to throw it away and buy a newer model. Instead, they can buy more processors and add them to the existing machine.*** [*Meiko was set up in 1985 by members of the group which developed the transputer chip for Inmos. The transputer contains a processor, memory, and communications links on a single chip, which makes it an ideal building block for multi-processor computers.*]

Merging

Restructuring of clauses and sentences

Restructure clauses and sentences to reflect the importance of the information contained within them. This can include transforming a sentence into a relative clause modifying the object of the previous sentence, and a PP or an adjective modifying an NP in a different sentence.

Restructure sentence and clauses to save space and avoid repetition. This can include transforming a sentence into a relative clause modifying the object of the previous sentence, and a PP or an adjective modifying an NP in a different sentence.

Merge units to make the abstract more concise.

Example:

~~*The first cyclone, named*~~ *Larry*~~*,*~~ *reached maximum (Category 5) intensity* ~~*at about the time of landfall.*~~ [*It hit* ~~*the coast at the town of*~~ *Innisfail at 0700* ~~*Eastern Standard Time (AEST) on*~~ *Monday*]*. …* [*It could turn out to be the most destructive cyclone ever to hit Queensland*]*.*

[***Possibly***] ***the most destructive cyclone to ever hit Queensland****, Larry reached maximum (Category 5) intensity* [***when***] ***it hit Innisfail at 0700 Monday****.*

Punctuation

Replace the full stop at the end of the first merged sentence with a comma or a hyphen.

Example:

Diesel engines are the most fuel efficient vehicle power source available~~*.*~~ [***,***] ~~***B***~~ *but diesel engines also have very dirty exhaust emissions.*

Connectives

Replace the full stop at the end of the first merged sentence with a connective, or insert a connective after punctuation has been replaced with punctuation, to create a grammatical sentence which flows well.

Example:

Jones is also determined to increase funding for basic research~~*.*~~[***, and***] ~~*A*~~*another of* ~~*Jones's*~~ ***his*** *major policy drives is information.*

Appendix IV: Previously published work

Some of the work described in the thesis has been previously published in conference proceedings during the period of PhD study. This work has, for the main part, been extended or recontextualised before its inclusion in the final version of this thesis. A brief description of these papers and their relation to the thesis is offered below, in chronological order.

Laura Hasler, Constantin Orasan and Ruslan Mitkov (2003) Building Better Corpora for Summarisation. In *Proceedings of Corpus Linguistics 2003*, 309-319. Lancaster, UK, 28-31 March.

This paper describes the annotation of the CAST corpus, the guidelines used to do this and an analysis of the corpus. The guidelines used in the CAST annotation and parts of the analysis discussion are presented in Chapter 4. Their presentation is similar to the original paper. However, the guidelines were modified for the annotation of extracts for this thesis and related to the three stages of human summarisation, and the analysis is compared with the deletion operations identified in Chapter 5. The modified guidelines can be found in Appendix I.

Laura Hasler (2004a) An Investigation into the Use of Centering Transitions for Summarisation. In *Proceedings of the 7th Annual Colloquium of the UK Special Interest Group in Computational Linguistics (CLUK'04)*, 100-107. Birmingham, UK, 6-7 January.

This paper discusses the possibility of using Centering Theory for summarisation. A small-scale and basic experiment from the paper gave the basic idea for the development of CT as an evaluation method in Chapter 7. CT as an evaluation method is developed further, using a modified CT analysis and developing a metric which reflects the transitions present in summaries. The evaluation is also done on a much larger scale.

Laura Hasler (2004b) "Why Do You Ignore Me?" - Proof that not all Direct Speech is Bad. In *Proceedings of the 4th International Conference on Languages Resources and Evaluation (LREC2004)*, 1041-1044. Lisbon, Portugal, 26-28 May.

This paper discusses cases of direct speech from the CAST corpus, identified during the annotation (Hasler, Orasan and Mitkov, 2003). The findings are summarised in Chapter 4 and Chapter 5 during the discussion of guidelines, and the observations informed some of the modifications made to the guidelines in order to improve them for the annotation of extracts in this thesis.

Bibliography

Alonso i Alemany, L. and Fuentes Fort, M. (2003) Integrating Coherence and Cohesion for Automatic Summarisation. In *Proceedings of the 10th Conference of The European Chapter of the Association for Computational Linguistics (EACL'03)*, 1-8. Budapest, Hungary, 12-17 April.

American National Standards Institute (1997) Guidelines for Abstracts. ANSI/NISO Z39.14-1997, National Information Standards Organization (NISO) Press, Bethesda, MD, USA.

Banko, M., Mittal, V. O. and Witbrock, M. J. (2000) Headline Generation Based on Statistical Translation. In *Proceedings of the 38th Annual Meeting of the Association of Computational Linguistics (ACL2000)*, 318-325. Hong Kong, 1-8 October.

Barzilay, R. and Elhadad, M. (1997) Using Lexical Chains for Text Summarization. In *Proceedings of the ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization (ISTS'97)*, 10-17. Madrid, Spain, 11 July.

Barzilay, R. (2003) *Information Fusion for Multi-document Summarisation: Paraphrasing and Generation*. PhD thesis, Columbia University, USA.

Barzilay, R. and Lapata, M. (2005) Modeling Local Coherence: An Entity Based Approach. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, 141-148. Ann Arbor, MI, USA, 25-30 June.

Baxendale, P. B. (1958) Man-made Index for Technical Literature: An Experiment. *IBM Journal of Research and Development* 2(5): 354-361.

Benbrahim, M. and Ahmad, K. (1994) Computer-aided Lexical Cohesion Analysis and Text Abridgement. Computing Sciences Report CS-94-11, University of Surrey, UK.

Benbrahim, M. and Ahmad, K. (1995) Text Summarisation: The Role of Lexical Cohesion Analysis. *The New Review of Document and Text Management* 1: 321 - 335.

Boguraev, B. and Kennedy, C. (1999) Salience-based Content Characterization of Text Documents. In Mani, I. and Maybury, M. T. (eds.). *Advances in Automatic Text Summarization*. Cambridge, MA: The MIT Press. 99-110.

Borko, H. and Bernier, C. L. (1975) *Abstracting Concepts and Methods*. San Diego, CA: Academic Press.

Brandow, R., Mitze, K. and Rau, L. (1995) Automatic Condensation of Electronic Publications by Sentence Selection. *Information Processing and Management* 31(5): 675-685.

Brennan, S. E., Friedman, M. A. and Pollard, C. J. (1987) A Centering Approach to Pronouns. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics (ACL'87)*, 155-162. Stanford, CA, USA, 6-9 July.

Burnard, L. (1995) Users Reference Guide: British National Corpus Version 1.0. Oxford University Computing Services.

Chuah, C.-K. (2001a) Just What May be Deleted or Compressed in Abstracting? In *Proceedings of Traitement Automatique du Langage Naturel (TALN2001)*, 339-344. Tours, France, 2-5 July.

Chuah, C.-K. (2001b) Aggregation by Conflation of Quasi-Synonymous Units in Author Abstracting. In *Proceedings of Traitement Automatique du Langage Naturel (TALN2001)*, 143-152. Tours, France, 2-5 July.

Cleveland, D. B. (1983) *Introduction to Indexing and Abstracting*. Littleton, CO: Libraries Unlimited, Inc.

Craven, T. C. (1988) Text Network Display Editing with Special Reference to the Production of Customized Abstracts. *Canadian Journal of Information Science* 13(1/2): 59-68.

Craven, T. C. (1993) A Computer-aided Abstracting Tool Kit. *Canadian Journal of Information and Library Science* 18(2): 19-31.

Craven, T. C. (1996) An Experiment in the Use of Tools for Computer-assisted Abstracting. In *Proceedings of The 59th American Society for Information Science Annual Meeting (ASIS'96)*, 203-208. Baltimore, MD, 21-24 October.

Craven, T. C. (1998) Human Creation of Abstracts with Selected Computer Assistance Tools. *Information Research* 3(4): paper 47.

Craven, T. C. (2000) Abstracts Produced Using Computer Assistance. *Journal of the American Society for Information Science* 51(8): 745-756.

Cremmins, E. T. (1996) *The Art of Abstracting*. Arlington, Virginia: Information Resources Press.

Crystal, D. (2003) *The Cambridge Encyclopedia of the English Language*. Cambridge: Cambridge University Press.

Cunningham, H., Maynard, D., Bontcheva, K. and Tablan, V. (2002) A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 168-175. Philadelphia, PA, USA, 6-12 July.

DeJong, G. (1982) An Overview of the FRUMP System. In Lehnert, W. G. and Ringle, M. H. (eds.). *Strategies for Natural Language Processing*. Hillsdale, New Jersey: Lawrence Erlbaum Associates. 149-176.

Donaway, R., Drummey, K. and Mather, L. (2000) A Comparison of Rankings Produced by Summarization Evaluation Measures. In *Proceedings of the NAACL-ANLP2000 Workshop on Automatic Summarization*, 69-78. Seattle, WA, USA, 30 April.

Edmundson, H. P. (1969) New Methods in Automatic Abstracting. *Journal of the Association for Computing Machinery* 16(2): 264-285.

Endres-Niggemeyer, B. (1998) *Summarizing Information*. Berlin: Springer.

Fellbaum, C., Ed. (1998) *WordNet: An Electronic Lexical Database*. Cambridge, MA, The MIT Press.

Fum, D., Guida, G. and Tasso, C. (1982) Forward and Backward Reasoning in Automatic Abstracting. In *Proceedings of the 9th International Conference on Computational Linguistics (COLING'82)*, 83-88. Prague, Czechoslovakia, 5-10 July.

Goldstein, J., Mittal, V. O., Carbonell, J. G. and Kantrowitz, M. (2000) Multi-document Summarization by Sentence Extraction. In *Proceedings of the NAACL-ANLP2000 Workshop on Automatic Summarization*, 40-48. Seattle, WA, USA, 30 April.

Gordon, P. C., Grosz, B. J. and Gilliom, L. A. (1993) Pronouns, Names, and the Centering of Attention in Discourse. *Cognitive Science* 17(3): 311-347.

Grosz, B. J., Joshi, A. K. and Weinstein, S. (1995) Centering: a Framework for Modelling the Local Coherence of Discourse. *Computational Linguistics* 21(2): 203-225.

Gunning, R. (1952) *The Technique of Clear Writing*. New York, NY: McGraw-Hill.

Hahn, U. and Reimer, U. (1999) Knowledge-based Text Summarization: Salience and Generalization Operators for Knowledge Base Abstraction. In Mani, I. and Maybury, M. T. (eds.). *Advances in Automatic Text Summarization*. Cambridge, MA: The MIT Press. 215-232.

Harnly, A., Nenkova, A., Passonneau, R. and Rambow, O. (2005) Automation of Summary Evaluation by the Pyramid Method. In *Proceedings of Recent Advances in Natural Language Processing 2005 (RANLP'05)*, 226-232. Borovets, Bulgaria, 21-23 September.

Hasler, L. (2003) Annotation for Summarisation: Marking Important Information in Similar Texts. In *Proceedings of the 1st Cambridge Postgraduate Conference in Linguistics (CamLing 2003)*, 100-106. Cambridge, UK, 26 April.

Hasler, L., Orasan, C. and Mitkov, R. (2003) Building Better Corpora for Summarisation. In *Proceedings of Corpus Linguistics 2003*, 309-319. Lancaster, UK, 28-31 March.

Hasler, L. (2004a) An Investigation into the Use of Centering Transitions for Summarisation. In *Proceedings of the 7th Annual Colloquium of the UK Special Interest Group in Computational Linguistics (CLUK'04)*, 100-107. Birmingham, UK, 6-7 January.

Hasler, L. (2004b) "Why Do You Ignore Me?" - Proof that not all Direct Speech is Bad. In *Proceedings of the 4th International Conference on Languages Resources and Evaluation (LREC2004)*, 1041-1044. Lisbon, Portugal, 26-28 May.

Hirschman, L. and Mani, I. (2003) Evaluation. In Mitkov, R. (ed.). *The Oxford Handbook of Computational Linguistics*. Oxford: Oxford University Press. 414 - 429.

Hoey, M. (1991) *Patterns of Lexis in Text*. Oxford: Oxford University Press.

Hovy, E. and Lin, C.-Y. (1999) Automated Text Summarization in SUMMARIST. In Mani, I. and Maybury, M. T. (eds.). *Advances in Automatic Text Summarization*. Cambridge, MA: The MIT Press. 81-94.

Hovy, E. (2003) Text Summarization. In Mitkov, R. (ed.). *The Oxford Handbook of Computational Linguistics*. Oxford: Oxford University Press. 583 - 598.

Hutchins, W. J. and Somers, H. L. (1992) *An Introduction to Machine Translation*. London: Academic Press Limited.

Jing, H. and McKeown, K. (1999) The Decomposition of Human-Written Summary Sentences. In *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval (SIGIR-99)*, 129-136. Berkeley, CA, USA, 15-19 August.

Jing, H. and McKeown, K. (2000) Cut and Paste Based Text Summarization. In *Proceedings of the 1st Conference of the North American Chapter of the Association of Computational Linguistics (NAACL2000)*, 178-185. Seattle, WA, USA, 29 April-3 May.

Jing, H. (2001) *Cut-and-Paste Text Summarization*. PhD thesis, Columbia University, USA.

Johnson, F. C., Paice, C. D., Black, W. J. and Neal, A. P. (1993) The Application of Linguistic Processing to Automatic Abstract Generation. *Journal of Document and Text Management* 1(3): 215-239.

Kameyama, M. (1998) Intrasentential Centering: A Case Study. In Walker, M. A., Joshi, A. K. and Prince, E. (eds.). *Centering Theory in Discourse*. Oxford: Clarendon Press. 89-112.

Karamanis, N. and Manurung, H. M. (2002) Stochastic Text Structuring Using the Principle of Continuity. In *Proceedings of the 2nd International Natural Language Generation Conference (INLG'02)*, 81-88. New York, NY, USA, 1-3 July.

Karamanis, N. (2003) *Entity Coherence for Descriptive Text Structuring*. PhD thesis, University of Edinburgh, UK.

Karamanis, N., Poesio, M., Mellish, C. and Oberlander, J. (2004) Evaluating Centering-based Metrics of Coherence Using a Reliably Annotated Corpus. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL'04)*, 391-398. Barcelona, Spain, 21-26 July.

Kay, M. (1980) The Proper Place of Men and Machines in Language Translation. Research Report CSL-80-11, Xerox Palo Alto Research Center, Palo Alto, CA, USA.

Kay, M. (1997) The Proper Place of Men and Machines in Language Translation. *Machine Translation* 12(1): 3-23.

Kibble, R. (1999) Cb or not Cb: Centering Theory Applied to NLG. In *Proceedings of the ACL'99 Workshop on the Relation between Discourse/Dialogue Structure and Reference*, 72-81. College Park, MD, USA, 21 June.

Kibble, R. and Power, R. (1999) Using Centering Theory to Plan Coherent Texts. In *Proceedings of The 12th Amsterdam Colloquium*, 187-192. Amsterdam, Netherlands, 18-21 December.

Kibble, R. and Power, R. (2000) An Integrated Framework for Text Planning and Pronominalisation. In *Proceedings of the 1st International Natural Language Generation Conference (INLG2000)*, 77-84. Mitzpe Ramon, Israel, 12-16 June.

Kibble, R. (2001) A Reformulation of Rule 2 of Centering Theory. *Computational Linguistics* 27(4): 579-587.

Kincaid, J. P., Fishburne, R. P., Rogers, R. L. and Chissom, B. S. (1975) Derivation of new readability formulas (automated readability index, Fog count and Flesch reading ease formula) for navy enlisted personnel. Research Branch Report 8-75, Naval Air Station, Memphis, TN, USA.

Kintsch, W. (1974) *The Representation of Meaning in Memory*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Kintsch, W. and van Dijk, T. A. (1978) Toward a Model of Text Comprehension and Production. *Psychological Review* 85(5): 363-394.

Kupiec, J., Pederson, J. and Chen, F. (1995) A Trainable Document Summarizer. In *Proceedings of the 18th International ACM/SIGIR Conference on Research and Development in Information Retrieval*, 69-73. Seattle, WA, USA, 9-13 July.

Lapata, M. and Barzilay, R. (2005) Automatic Evaluation of Text Coherence: Models and Representations. In *Proceedings of the 19th International Conference on Artificial Intelligence (IJCAI'05)*, 1085-1090. Edinburgh, UK, 31 July-5 August.

Liddy, E. D. (1991) The Discourse-level Structure of Empirical Abstracts: An Exploratory Study. *Information Processing and Management* 27(1): 55-81.

Lin, C.-Y. and Hovy, E. (1997) Identifying Topics by Position. In *Proceedings of the 5th Conference on Applied Natural language Processing (ANLP'97)*, 283-290. Washington D.C., 31 March-3 April.

Lin, C.-Y. (2001) SEE - Summary Evaluation Environment. http://haydn.isi.edu/SEE/.

Lin, C.-Y. and Hovy, E. (2003) Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In *Proceedings of the Human Language Technology Conference/North American Chapter of the Association for Computational Linguistics (HLT/NAACL2003)*, 71-78. Edmonton, Canada, 27 May-1 June.

Lin, C.-Y. (2004) ROUGE: A Package for Automatic Evaluation of Summaries. In *Proceedings of The ACL2004 Workshop Text Summarization Branches Out, Barcelona, Spain, 25-26 July*, 74-81.

Luhn, H. P. (1958) The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development* 2(2): 159-165.

Mani, I., Firmin, T., House, D., Chrzanowski, M., Klein, G., Hirschman, L., Sundheim, B. and Obrst, L. (1998) The TIPSTER SUMMAC Text Summarization Evaluation: Final Report. MITRE Technical Report MTR 98W0000138, The MITRE Corporation, McLean, VA, USA.

Mani, I. and Bloedorn, E. (1999) Summarizing Similarities and Differences Among Related Documents. *Information Retrieval* 1(1): 35-67.

Mani, I., Gates, B. and Bloedorn, E. (1999) Improving Summaries by Revising Them. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, 558-565. College Park, MD, USA, 20-26 June.

Mani, I. (2001) *Automatic Summarization*. Amsterdam/Philadelphia: John Benjamins.

Mani, I. and Maybury, M. T., Eds. (1999) *Advances in Automatic Text Summarization*. Cambridge, MA, The MIT Press.

Mann, W. C. and Thompson, S. A. (1988) Rhetorical Structure Theory: Towards a Functional Theory of Text Organization. *Text* 8(3): 243-281.

Marcu, D. (1997) *The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts*. PhD thesis, University of Toronto, Canada.

Marcu, D. (1999) Discourse Trees are Good Indicators of Importance in Text. In Mani, I. and Maybury, M. T. (eds.). *Advances in Automatic Text Summarization*. Cambridge, MA: The MIT Press. 123-136.

McKeown, K., Robin, J. and Kukich, K. (1995) Generating Concise Natural Language Summaries. *Information Processing and Management* 31(5): 703-733.

Miike, S., Itoh, E., Ono, K. and Sumita, K. (1994) A Full-text Retrieval System with a Dynamic Abstract Generation Function. In *Proceedings of the 17th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, 152-161. 3-6 July.

Minel, J.-L., Nugier, S. and Piat, G. (1997) How to Appreciate the Quality of Automatic Text Summarization? Examples of FAN and MLUCE Protocols and their Results on SERAPHIN. In *Proceedings of the ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization (ISTS'97)*, 25-31. Madrid, Spain, 11 July.

Mitkov, R., Le Roux, D. and Descles, J. P. (1994) Knowledge-based automatic abstracting: Experiments in the sublanguage of elementary geometry. In Martin Vide, C. (ed.). *Current Issues in Mathematical Linguistics*. Amsterdam: North Holland. 415-421.

Mitkov, R. (1995) A Breakthrough in Automatic Abstracting: the Corpus-based Approach. Working paper, University of Wolverhampton, UK.

Mitkov, R. (2003) Anaphora Resolution. In Mitkov, R. (ed.). *The Oxford Handbook of Computational Linguistics*. Oxford: Oxford University Press. 267-283.

Mitkov, R. and Ha, L. A. (2003) Computer-aided Generation of Multiple-choice Tests. In *Proceedings of the HLT-NAACL 2003 Workshop on Building Educational Applications Using Natural Language Processing*, 17 - 22. Edmonton, Canada, 31 May.

Mitkov, R. and Orasan, C. (2004) Discourse and Coherence: Revising the Claims and Conventions in Centering Theory. In *Proceedings of the 5th Discourse and Anaphora Resolution Colloquium (DAARC2004)*, 109-114. Sao Miguel, Azores, Portugal, 23-24 September.

Morris, A., Kaspar, G. and Adams, D. (1992) The Effects and Limitations of Automated Text Condensing on Reading Comprehension Performance. *Information Systems Research* 3(1): 17-35.

Nanba, H. and Okumura, M. (2000) Producing More Readable Extracts by Revising Them. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING2000)*, 1071-1075. Saarbruecken, Germany, 31 July-4 August.

Narita, M. (2000) Constructing a Tagged E-J Parallel Corpus for Assisting Japanese Software Engineers in Writing English Abstracts. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC2000)*, 1187-1191. Athens, Greece, 31 May-2 June.

Narita, M., Kurokawa, K. and Utsuro, T. (2002) A Web-based English Abstract Writing Tool Using a Tagged E-J Parallel Corpus. In *Proceedings of the 3rd*

International Conference on Language Resources and Evaluation (LREC2002), 2115-2119. Las Palmas de Gran Canaria, Spain, 29-31 May.

Nenkova, A. and Passonneau, R. (2004) Evaluating Content Selection in Summarization: The Pyramid Method. In *Proceedings of the Human Language Technology Conference/North American Chapter of the Association for Computational Linguistics (HLT/NAACL2004)*, 145-152. Boston, MA, USA, 2-7 May.

Ono, K., Sumita, K. and Miike, S. (1994) Abstract Generation Based on Rhetorical Structure Extraction. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING'94)*, 344-348. Kyoto, Japan, 5-9 August.

Orasan, C. (2001) Patterns in Scientific Abstracts. In *Proceedings of Corpus Linguistics 2001*, 423-432. Lancaster, UK, 30 March-2 April.

Orasan, C., Mitkov, R. and Hasler, L. (2003) CAST: a Computer-Aided Summarisation Tool. In *Proceedings of the 11th Conference of The European Chapter of the Association for Computational Linguistics (EACL'03)*, 135-138. Budapest, Hungary, 12-17 April.

Orasan, C. (2003a) PALinkA: A Highly Customizable Tool for Discourse Annotation. In *Proceedings of the 4th SIGDial Workshop on Discourse and Dialog*, 39-43. Sapporo, Japan, 5-6 July.

Orasan, C. (2003b) An Evolutionary Approach for Improving the Quality of Automatic Summaries. In *Proceedings of the ACL'03 Workshop on Multilingual Summarization and Question Answering - Machine Learning and Beyond*, 37-45. Sapporo, Japan, 11 July.

Orasan, C., Pekar, V. and Hasler, L. (2004) A Comparison of Summarisation Methods Based on Term Specificity Estimation. In *Proceedings of the 4th*

International Conference on Language Resources and Evaluation (LREC2004), 1037-1041. Lisbon, Portugal, 26-28 May.

Orasan, C. (2006) *Comparative Evaluation of Modular Summarisation Systems Using CAST*. PhD thesis, University of Wolverhampton, UK.

Orasan, C. and Hasler, L. (2006) Computer-aided Summarisation: What the User Really Wants. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC2006)*. Genoa, Italy, 24-26 May.

Orasan, C. and Hasler, L. (forthcoming) Computer-aided summarisation: How much does it really help? In *Proceedings of Recent Advances in Natural Language Processing 2007 (RANLP'07)*. Borovets, Bulgaria, 27-29 September.

Orasan, C., Hasler, L. and Mitkov, R. (forthcoming) Corpora for Text Summarisation. In Ludeling, A. and Kyto, M. (eds.). *Corpus Linguistics - An International Handbook*. Berlin: Mouton de Gruyter.

Paice, C. D. (1981) The Automatic Generation of Literature Abstracts: An Approach Based on Self-indicating Phrases. In Oddy, R. N., Rijsbergen, C. J. and Williams, P. W. (eds.). *Information Retrieval Research*. London: Butterworths. 173-191.

Paice, C. D. and Jones, P. A. (1993) The Identification of Concepts in Highly Structured Technical Papers. In *Proceedings of the 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, 69-78. Pittsburgh, PA, USA, 27 June-1 July.

Pinto Molina, M. (1995) Documentary Abstracting: Toward a Methodological Method. *Journal of the American Society for Information Science* 46(3): 225-234.

Poesio, M., Stevenson, R., di Eugenio, B. and Hitzeman, J. (2004) Centering: A Parametric Theory and its Instantiations. NLE Technical Note TN-02-01/CS Technical Report CSM-369, University of Essex, UK.

Pollock, J. J. and Zamora, A. (1975) Automatic Abstracting Research at Chemical Abstracts Service. *Journal of Chemical Information and Computer Sciences* 15(4): 226-232.

Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. (1985) *A Comprehensive Grammar of the English Language*. Harlow: Longman.

Radev, D. R. and McKeown, K. R. (1998) Generating Natural Language Summaries from Multiple On-line Sources. *Computational Linguistics* 24(3): 469-500.

Rau, L. F., Jacobs, P. S. and Zernik, U. (1989) Information Extraction and Text Summarization Using Linguistic Knowledge Acquisition. *Information Processing and Management* 25(4): 419-428.

Reimer, U. and Hahn, U. (1988) Text Condensation as a Knowledge Base Abstraction. In *Proceedings of the 4th Conference on Artificial Intelligence Applications (CAIA'88)*, 338-344. San Diego, CA, USA, 14-18 March.

Reiter, E. and Dale, R. (2000) *Building Natural Language Generation Systems*. Cambridge: Cambridge University Press.

Rose, T. G., Stevenson, M. and Whitehead, M. (2002) The Reuters Corpus Volume 1 - from Yesterday's News to Tomorrow's Language Resources. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC2002)*, 827-833. Las Palmas de Gran Canaria, Spain, 29-31 May.

Rowley, J. E. (1988) *Abstracting and Indexing*. London: Clive Bingley.

Saggion, H. and Lapalme, G. (2000) Concept Identification and Presentation in the Context of Technical Text Summarization. In *Proceedings of The NAACL-ANLP 2000 Workshop on Automatic Summarization*, 1-10. Seattle, WA, USA, 30 April.

Salager-Meyer, F. (1990) Discoursal Movements in Medical English Abstracts and their Linguistic Exponents: A Genre Analysis Study. *INTERFACE: Journal of Applied Linguistics* 4(2): 107-124.

Salton, G. and McGill, M. J. (1983) *Introduction to Modern Information Retrieval*. New York: McGraw-Hill.

Salton, G., Singhal, A., Mitra, M. and Buckley, C. (1997) Automatic Text Structuring and Summarization. *Information Processing and Management* 33(2): 193-207.

Siegel, S. and Castellan, N. J. (1988) *Nonparametric Statistics for the Behavioral Sciences*. New York: McGraw-Hill.

Skorochod'ko, E. F. (1971) Adaptive Method of Automatic Abstracting and Indexing. In *Proceedings of the 5th International Federation for Information Processing Congress (IFIP 71)*, 1179-1182. Ljubljana, Yugoslavia, 23-28 August.

Sparck Jones, K. and Galliers, J. (1996) *Evaluating Natural Language Processing Systems: An Analysis and Review*. Berlin: Springer Verlag.

Sparck Jones, K. (1999) Automatic Summarizing: Factors and Directions. In Mani, I. and Maybury, M. T. (eds.). *Advances in Automatic Text Summarization*. Cambridge, MA: MIT Press. 1-12.

Sparck Jones, K. (2001) Factorial Summary Evaluation. In *Proceedings of the DUC 2001 Workshop on Text Summarization*. New Orleans, LA, USA, 13-14 September.

Strube, M. and Hahn, U. (1999) Functional Centering - Grounding Referential Coherence in Information Structure. *Computational Linguistics* 25(3): 309-344.

Strzalkowski, T., Stein, G., Wang, J. and Wise, B. (1999) A Robust Practical Text Summarizer. In Mani, I. and Maybury, M. T. (eds.). *Advances in Automatic Text Summarization*. Cambridge, MA: The MIT Press. 137-154.

Suri, L. A. and McCoy, K. F. (1994) RAFT/RAPR and Centering: A Comparison and Discussion of Problems Related to Processing Complex Sentences. *Computational Linguistics* 20(2): 307-317.

Swales, J. M. (1990) *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press.

Teufel, S. and Moens, M. (1997) Sentence Extraction as a Classification Task. In *Proceedings of the ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization (ISTS'97)*, 58-59. Madrid, Spain, 11 July.

Teufel, S. and Moens, M. (1999) Argumentative Classification of Extracted Sentences as a First Step towards Flexible Abstracting. In Mani, I. and Maybury, M. T. (eds.). *Advances in Automatic Text Summarization*. Cambridge, MA: The MIT Press. 155-171.

The Oxford English Dictionary Second Edition (1989, Volume I) Oxford: Clarendon Press.

The Oxford English Dictionary Second Edition (1989, Volume IV) Oxford: Clarendon Press.

The Oxford English Dictionary Second Edition (1989, Volume XI) Oxford: Clarendon Press.

Tucker, R. (1999) *Automatic Summarising and the CLASP System*. PhD thesis, University of Cambridge, UK.

van Deemter, K. and Kibble, R. (2000) On Coreferring: Coreference in MUC and Related Annotation Schemes. *Computational Linguistics* 26(4): 615-623.

van Dijk, T. A. (1979) Recalling and Summarizing Complex Discourse. In Burchart, W. and Hulker, K. (eds.). *Textverarbeitung/Text Processing*. Berlin: de Gruyter. 49-93.

van Dijk, T. A. (1988) *News as Discourse*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Walker, M. A. (1998) Centering, Anaphora Resolution and Discourse Structure. In Walker, M. A., Joshi, A. K. and Prince, E. F. (eds.). *Centering Theory in Discourse*. Oxford: Oxford University Press. 401-435.

Walker, M. A., Joshi, A. K. and Prince, E. F. (1998) Centering in Naturally Occurring Discourse: An Overview. In Walker, M. A., Joshi, A. K. and Prince, E. F. (eds.). *Centering Theory in Discourse*. Oxford: Oxford University Press. 1-28.