

Annotation for Summarisation: marking important information in similar texts

Laura Hasler

University of Wolverhampton

Whilst much is made of the importance of multiple annotations of documents for summarisation allowing interannotator agreement to be computed, there is no research regarding what happens when an annotator is faced with several versions of a text which is primarily “about” the same topic. This paper details how human annotators mark important information in a corpus containing sets of texts that are concerned with the same story, but are written from a different perspective, and the impact this has on those sentences which are considered important. The analysis comprises two levels: lexical and discourse, and proves the hypothesis that features relating to both content and structure do indeed affect the importance associated with certain textual units.

1 INTRODUCTION

This work originated from some interesting observations made when analysing a corpus of texts built for automatic summarisation as part of the AHRB-funded project “CAST: a Computer Aided Summarisation Tool”, which aims at developing a computer-aided summarisation tool for newswire and scientific texts (see Orasan et al., 2003 and <http://www.clg.wlv.ac.uk/projects/CAST/>). During the analysis, it became apparent that a number of texts were “about” the same series of events, that is, that we had several “similar” versions of several different texts annotated for important information by the same people. This led to the observation that there was no existing research which looked at this aspect of annotating texts for summarisation, and what is described here is the investigation into the impact of features relating to content and structure on how annotators mark important information in similar newswire texts. For a description of the original corpus, including the annotation scheme, annotation guidelines and analysis, see Hasler et al. (2003).

1.1 Annotated Corpora and Automatic Summarisation

Automatic summarisation is an area within the fields of Computational Linguistics and Natural Language Processing (NLP) which develops the process of using computers to create summaries of texts, as opposed to employing humans to do the same task. There are several reasons for the emergence of such a research area, including the bias, cost and other problems associated with human-written summaries. Annotated corpora have proved essential in many areas of NLP and Computational Linguistics. They have been particularly useful in automatic summarisation, where they are mainly used for machine learning, to learn patterns for the extraction of important (and other) information from texts, as well as for the evaluation of summarisation methods (for example, Edmundson, 1969; Marcu, 1997).

When annotating corpora, one accurate method is to employ humans to indicate those parts of text to be annotated with whatever information necessary (in our case, important sentences). These human-selected units of text can then be used as a gold standard by which to measure the performance of a system, as well as for discerning which types of units are chosen or discarded by humans during the summarisation process. In most cases, more than

one annotator is used for the annotation process, as this allows the interannotator agreement to be measured, helping to better validate the data.

1.2 Motivation

Whilst much is made of the importance of such multiple annotations of documents for summarisation, there is no research regarding what happens when an annotator is faced with several versions of the same text, i.e. texts which are primarily “about” the same topic. Although perhaps not as pertinent to automatic summarisation as a whole as different texts annotated by more than one person, it is still interesting to investigate patterns of important information and reasons for those patterns within groups of similar texts. This could be useful in an area such as multi-document summarisation, where several similar documents can provide the source for a single summary. This paper details how human annotators mark important information in a corpus containing groups of texts that are concerned with the same story but are written from a different perspective. The impact this perspective has on those sentences which are considered important is also described.

The analysis comprises two levels: lexical and discourse, and proves the hypothesis that stylistic factors do indeed affect the importance associated with certain textual units. The results of this analysis show that discourse features relating to both content and structure have a strong impact on the information that human annotators mark as important in similar texts.

1.3 Structure of Paper

The paper is structured as follows: section 2 describes the corpus and the annotation applied to it. The results obtained from analysing the corpus, and a discussion of these results, including average similarity, and lexical and discourse level analyses, are detailed in section 3. The paper finishes with conclusions drawn in section 4.

2. CORPUS AND ANNOTATION

This section describes the composition of the small corpus used for the analysis of similar texts, and gives more information about the guidelines, scheme and tool which were used during the annotation process.

2.1 The Corpus

The corpus contains 30 different newswire texts with a total of 40 annotations, and 11 groups of similar texts (containing between two and five versions of a story). The texts were taken from Hasler et al.’s (2003) corpus, the main (newswire) section of which was taken from the Reuters Corpus (Rose et. al., 2002). The fact that some of the texts were annotated by more than one person is not important here, because we are not concerned with interannotator agreement but with the same person’s annotation of similar texts. As this is the case, the same texts annotated by different people are considered as different annotations (and are therefore counted separately; hence the need for the distinction between the number of texts and the number of annotations in the corpus), because they have been marked by a different annotator, and are completely independent of any marking carried out by anyone else.

2.2 Annotation

The guidelines, scheme and tool used for the annotation are those used in Hasler et al. (2003), as the texts used were taken from this annotated corpus. The sections of the guidelines relevant to the research presented in this paper contain information about the compression rate for the marked sentences (30% total: 15% essential and 15% important sentences), as well as which types of sentences and information are considered worthy of marking (as essential and important) and what types of information should not be marked (i.e. are considered unimportant). Only parts of the guidelines are relevant here, as we are only concerned with the way important information is marked in texts which are similar. The guidelines help to ensure more consistent and reliable annotation. The complete set of guidelines can be viewed at <http://www.clg.wlv.ac.uk/projects/CAST/guidelines.pdf>.

The annotation scheme accounts for several different types of information, but again, here we are only concerned with sentences marked as essential and important. In this analysis, the two categories are collapsed into one, as the concern is with information which is considered to be generally important, and not the distinction between the labels.

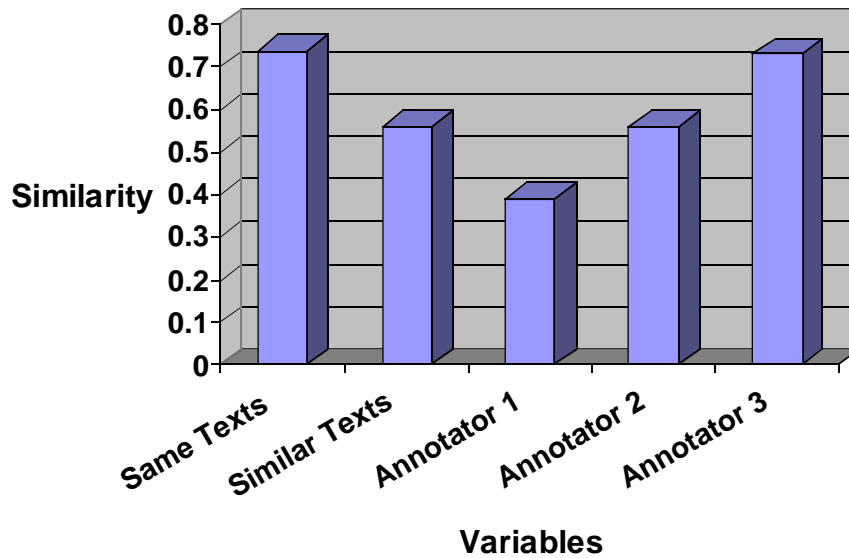
3. RESULTS AND DISCUSSION

In this section, the results obtained from an analysis of the annotated similar texts in terms of the effect of features regarding both content and structure at the lexical and discourse levels are discussed. There are also some observations about the sentences themselves, but these are mainly concerned with the information within them and their position in the different versions of similar texts, and so are not classed as a separate “syntactic” level analysis, as they fit better into the lexical and discourse categories.

3.1 Average Similarity

During the analysis of the larger corpus, Hasler et al. (2003) measured the similarity of the same texts annotated by different people, using the cosine distance (Salton and McGill, 1983). The average similarity of these texts was 0.74. However, when this figure was computed for similar texts annotated by the same person, it fell to 0.56, which is considerably lower. Although this figure is much lower, it is still possible to term the texts “similar”, as on reading them one can judge whether they are essentially about the same thing (and these are). If aspects such as synonymy were also taken into account, then the figure may be higher. There was some variation in the scores for the individual annotators, but these scores were always less than the figure for the same texts. Annotator 1 averaged 0.39, with Annotator 2 at 0.56 and Annotator 3 averaging 0.73. The graph below (Graph 1) displays these figures for the average similarity of marked sentences in the texts. The fact that the similarity was significantly lower for these types of text than for the other texts (same text, different annotators) proves that it must be certain aspects of the text (and not the annotator) which is responsible for the difference. The following sections discuss the reasons for this lower similarity in more detail.

The similarity of the full texts also affects the similarity of the marked sentences. In general, the higher the full text similarity, the higher the similarity between the extracts. For example, in one instance the full text similarity was 0.99, with a corresponding extract similarity of 0.81, whereas a full text similarity of 0.50 gave an extract similarity of 0.27 (figures taken from Annotator 3). Section 3.3 provides more details on this phenomenon. However, this is not always the case, as the position of the information within the source text has some impact on the sentences which are considered important enough to be included in the extract.

Graph 1: Average similarity of marked sentences

3.2 Lexical Level Analysis

The similarity measure (see above) can be seen as a lexically-based measure as it is computed by taking into account the overlap in vocabulary between two texts, in this case, between similar texts dealing with what is essentially the same story. It is especially important to use a measure such as this here (as opposed to a measure which computes the similarity of marked sentences by the sentence number; for example, S2 is marked in both texts), as discourse features such as text structure could have an impact on this; different versions of a text can and often do, quite feasibly, present the same information in different positions or sentences within the texts. Therefore a measure which is based on the similarity of the sentences (but not the words/information in them) marked will be inaccurate. In the texts in this corpus, a lack of shared content demonstrated by low vocabulary overlap equals a below average similarity between texts.

The title/headline (also important at the discourse level; see section 3.3 for more details) is generally a good indicator of the content of the text. It can be seen as the starting point for the selection of important information. So, if the titles of the texts are very different, the similarity of the full texts, and therefore of the marked sentences, tends to be much lower than if they are very similar or the same. For example, two texts with the titles *War hero Colin Powell hits road with Dole campaign* and *War hero Colin Powell boosts Dole campaign* have an extract similarity of 0.63, whilst for the texts *Aboriginies burn flag in Canberra budget protests* and *Australian govt pushes budget, despite protests* it is 0.40 (figures taken from Annotator 2). However, it is important to point out that not all texts which have the same (or very similar) titles share exactly the same information with the same lexical representation in the same position, and because of this, the extract similarity does not always depend directly on this type of lexical measure (it is also affected by the structure of the text; see section 3.3).

3.3 Discourse Level Analysis

As mentioned above, the position of the information in the full text is important in determining whether it is considered important. This is due to the fact that in news texts, the most important aspects of the most important information are usually presented first (van Dijk, 1988 expands on this further), and often follow on from the title or headline of the piece (the headline and lead sentence or paragraph are often considered to summarise a news article). In the majority of the annotated texts (70%), the first sentence of the text was marked as important. Annotator 3 only marked the first sentence (S1) in 2 out of 13 cases, but she also marked the title (headline) in many cases, which was unnecessary as it will be automatically extracted to appear with the selected sentences from the text it belongs to. In any future annotations, it will be made clearer that the annotators should not mark the headline of the texts. It is also interesting that Annotator 3 felt that the S1 was generally too similar to the headline to be included and usually marked the second sentence (S2), whilst the other annotators were aware that the headline would be automatically included and still marked the S1 nearly all of the time. One reason for this is that the headline alone is often very short, whereas S1 contains more (and often valuable) information which it is necessary to include in the set of marked sentences despite the fact that a small part of that information has already been given. However, the title still proved to be important in the discourse level analysis as it reflected not only the most important content, but also the structure of the text, i.e. the text is structured so that the most important content appears in the position most appropriate to it.

The structure of the text has a strong impact on the sentences that are marked as important; the position and order in which the information appears in the text really matters, as does the space allotted to that information. Position and order tell us what information the author considered necessary to introduce first, and therefore (generally) considered most important. The space allotted to information indicates the main ideas of the text, as well as its importance. The same information appearing in the same position and in the same order in two texts indicates an above average similarity between those texts. All three of these aspects are intertwined with the notion of discourse structure, in that they both affect and are affected by how the sentences work together to give the desired effect of a newsworthy story. This overall structure (which incorporates, among other things, position, order and space) obviously affects (and is in turn affected by) what is considered important.

To illustrate the influence of these three structural features (for information in the full text) on sentences marked as important, let us consider two extracts with a very low similarity of 0.22. The full texts were essentially “about” the same thing; that is, they contained information about the same series of events involving the same participants (Italian and Albanian actions and reactions surrounding the sinking of a ship with Albanian refugees on board). However, this information was presented very differently in the two texts (the full text similarity was 0.40). Text 1 focuses on Albania as a “rebel-held wasteland”, whereas Text 2 is more concerned with the security force Italy was sending to Albania. Text 1 is concerned with the state of Vlore, containing many different descriptions and examples of the town in fairly short sentences. There is no mention of Italy or the ship sinking until near the end of the text, and then only 6 or 7 short sentences (out of a total of 38), suggesting that this information is not of crucial importance. This is reflected in the annotation, as Italy is only mentioned in one marked sentence, as a place where people fled to because of the state of Vlore. In contrast, Text 2 does not mention Vlore at all until the end of the text, and again very briefly (5 or 6 sentences), and only as somewhere that the refugees fleeing to Italy came from and that the people from this place want revenge against the Italians. Again, this is reflected in the marked sentences, as only two contain any reference to Vlore. The differences between the texts are also reflected in the two titles: *Albania’s second port now a rebel-held wasteland* versus *Italy presses on with security force for Albania*. When the sets of marked sentences are read as extracts, they are completely different:

Albania’s second port now a rebel-held wasteland (total sentences: 38)

- S1 - Shops, bars and cafes have been looted on almost every street.
 S2 - The main bank has been burned
 S3 - At least one person is killed every day
 S4 - Vlore, once Albania's bustling second city, is now a violent wasteland with the recently installed rebel Committee for Public Salvation struggling to keep the place afloat.
 S5 - An Italian-led multinational security force is due to be deployed in Vlore in the coming weeks to protect humanitarian supplies destined for its port.
 S7 - Vlore spearheaded a nationwide revolt against president Sali Berisha after a string of fraudulent investment schemes failed in January, wiping out millions of dollars in savings.
 S10 - "We have nothing left."
 S11 - All the structures and institutions of a town have collapsed," said Thanas Laluci, an economist on the rebel council.
 S13 - Mounds of dirt, block of concrete and piles of boulders sit in pot-holed roads.
 S14 - Streams of garbage fester everywhere.
 S16 - Berisha's three-storey presidential villa overlooking the sea has been cleaned out and trashed.
 S25 - "For two months we have had no supplies of food, equipment or medicine by land or sea."
 S27 - No-one here has been paid, no pensions given, no money, nothing." Said Laluci.
 S28 - Residents blame Berisha for their fate, which forced more than 13,000 Albanians to flee to Italy in unsafe boats in March alone.

Italy presses on with security force for Albania (total sentences: 25)

- S1 - Italy pressed on with preparing a multinational security force to deploy in lawless Albania and tried to ease strained relations by calling for a joint Rome-Tirana enquiry into the sinking of a refugee boat.
 S2 - "The government wants to clear this up completely...
 S3 - So that no shadow of a doubt remains," Prime Minister Romano Prodi told reporters.
 S6 - One key point of deployment is Vlore, the Adriatic port where the ill-fated refugees set out on Friday.
 S7 - Residents in the town have vowed revenge if Italian troops arrive.
 S8 - More than 80 Albanians were feared drowned in a collision between their boat and an Italian navy ship.
 S11 - Survivors accuse the Italians of ramming their vessel to stem a flood of refugees after about 13,000 fled across the Adriatic in March.
 S13 - Prodi said that the incident must not be allowed to taint previously close ties and his government insists the tragedy makes the mission ever more urgent to stem a tide of refugees.

The distribution of marked sentences was relatively even throughout the texts; there were no striking differences in the numbers of sentences taken from the beginning, middle or end of a text. Having said this, most of the time the first sentence was marked, but there were no other such distinguishing patterns; for example, there was no instance of, say, sentence 11 or the final sentence usually being marked. This even distribution can be partly explained by van Dijk's (1988) claim that news articles have 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 (the quote here describes a terrorist attack, but this can easily be adapted to other sub-genres of news text):

"...main acts and participants 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."¹

¹ Teun A. van Dijk (1988) *News as Discourse*. Hillsdale, NJ: Lawrence Erlbaum Associates, p.48

4. CONCLUSIONS AND FUTURE WORK

In this paper, I have discussed the impact of lexical and discourse features on the selection of important information in similar newswire texts. The average similarity of information marked by the same annotator as important in these texts was noticeably lower than the same type of information marked by different annotators in texts describing the same events.

Both the lexical and discourse levels of analysis indicated a definite impact on the information marked as important in similar texts, especially with regards to vocabulary overlap (the basis of the similarity measure), title, and the position and space allotted to information within the text.

This analysis is also strongly dependent on the similarity of the full texts; texts with very different perspectives (and therefore low document similarity) generally displayed a very low similarity for the information marked, whereas texts containing (near) identical sentences (and very high document similarity) displayed a much higher similarity. However, it is also important to remember that there are exceptions to these general findings, and that not all texts with similar lexical features demonstrate high similarity due to certain discourse aspects within them, and vice versa.

On a more practical level, the results presented here emphasise the difficulty of annotation for summarisation, as even in texts which are “similar”, annotators do not select the same information as important. This is due to a number of factors, as discussed in the paper. The results and discussion also suggest that in areas such as multi-document summarisation, where several similar texts are used as the basis for one summary, a measure such as similarity alone is not sufficient to select the information, as it can give figures which do not truly represent the similarity as recognised by humans. Future work will consider features such as synonymy in order to improve the accuracy of the similarity measure.

REFERENCES

- Edmundson, H. P. (1969) New methods in automatic abstracting. *Journal of the Association for Computing Machinery*, 16 (2): 264-285.
- Hasler, L., C. Orasan and R. Mitkov (2003) ‘Building better corpora for summarisation’. In *Proceedings of Corpus Linguistics 2003*, pages 309-319, Lancaster, UK, 28-31 March.
- Marcu, D. (1997) ‘From discourse structures to text summaries’. In *Proceedings of the ACL/EACL '97 Workshop on Intelligent Scalable Text Summarization*, Madrid, pp 82-88.
- Orasan, C., R. Mitkov and L. Hasler (2003) ‘CAST: a Computer Aided Summarisation Tool’. In *Proceedings of the 10th Conference of the European Chapter for Computational Linguistics (EACL2003)*, pages 135-138, Budapest, Hungary, 12-17 April.
- 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 Third International Conference on Language Resources and Evaluation (LREC-2002)*, Las Palmas de Gran Canaria, pp 827-833.
- Salton, G. and McGill, M.J. (1983) *Introduction to Modern Information Retrieval*. New York: McGraw-Hill.
- van Dijk, T. A. (1988) *News as Discourse*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Laura Hasler

Research Group in Computational Linguistics
School of Humanities, Languages and Social Sciences
University of Wolverhampton
Stafford Street
Wolverhampton
WV1 1SB
United Kingdom

L.Hasler@wlv.ac.uk
<http://clg.wlv.ac.uk>