Pronominal anaphora resolution for text summarisation

Constantin Orăsan
Research Group in Computational Linguistics
University of Wolverhampton
Stafford St.
Wolverhampton, WV1 1SB, UK
C.Orasan@ulw.ac.uk

Abstract
The assumption of term-based summarisation method is that the importance of a sentence can be determined by the importance of the words it contains. One drawback of these methods is that they usually consider the words in isolation, ignoring relations such as anaphoric links between them. This paper investigates to what extent the integration of pronominal anaphora resolution into the summarisation process can improve the informativeness of the automatically produced summaries. Evaluation of three anaphora resolution methods plus three baselines on a corpus of journal articles shows that anaphora resolution can have a beneficial effect on informativeness. In addition, one experiment which uses a simulated anaphora resolver with predefined accuracy is performed in order to demonstrate that term-based summarisation can benefit from anaphora resolution, but that high accuracy methods are necessary.

Keywords
automatic summarisation, evaluation, anaphora resolution

1 Introduction
With the current information overload experienced by researchers, it is increasingly difficult to keep up-to-date with all current developments in a field, and it has become necessary to look for a piece of information only when it is needed. Automatic summarisation can help people deal with this abundance of information by extracting the gist of it. Term-based summarisation is one of the most common components of text summarisation systems and is the focus of this paper. First proposed by Luhn [27], it is still widely used today in combination with other methods [23, 44, 25, 43] due to its lack of complexity and its high speed. The assumption of term-based summarisation is that it is possible to determine the importance of a sentence on the basis of the words it contains. The most common way of achieving this is to weigh all the words in a text and calculate the score of a sentence by adding together the weights of the words within it. In this way, a summary can be produced by extracting the sentences with the highest scores until the desired length is reached. One drawback of most implementations of term-based summarisers is that they score words in isolation, ignoring links between words such anaphoric relations.

This paper investigates the extent to which pronominal anaphora resolution can improve the results of a term-based summariser by resolving pronouns to their antecedents and incorporating this information in the summariser. It has to be pointed out that the purpose of this paper is not to produce a new summarisation method, but to assess whether information from an anaphora resolver can be beneficial for the summarisation process. To this end, term-based summarisation is very appropriate for this task as it depends on a limited number of parameters and any change in its performance can be justified by the additional information from the anaphora resolver.

The paper is structured as follows: Section 2 briefly presents background information about automatic summarisation, pronominal anaphora resolution and previous attempts to combine the two. Section 3 presents the term-based summarisation method employed in this paper, as well as the anaphora resolvers used to enhance the summarisation method. The corpus used in our experiments is described in Section 4, which is followed by a section on evaluation. The evaluation focuses on both the accuracy of anaphora resolution and on the performance of term-based summarisation methods, and tries to establish whether there is any correlation between the accuracy of an anaphora resolver and the increase in the accuracy of the summariser when it incorporates the resolver. In order to further assess the influence of anaphora resolution for automatic summarisation, the evaluation also presents the effects of a resolver with controlled accuracy on term-based summarisation. The paper finishes with conclusions.

2 Background
Both automatic summarisation and anaphora resolution have received extensive attention from the research community. This section briefly describes the two areas with an emphasis on the aspects relevant to this paper. Due to space restrictions no attempt will be made to present a comprehensive overview of the two fields. Such an overview for automatic

---

1 At this stage the accuracy and performance are used as general terms. Section 5 explains what they mean in the context of this paper.
summarisation can be found in [28] and for anaphora resolution in [31].

2.1 Automatic summarisation

Automatic summarisation is a field in computational linguistics concerned with the development of systems which can produce summaries automatically. These systems take one or several related documents, and summarise the most important information from them or information related to aspects chosen by a user. These systems could prove very useful for example to researchers who need to quickly find out the content of an article. Unfortunately, with the current technology, it is difficult to produce automatic summaries which replace the whole document. Instead, automatic summarisation can be used to produce summaries which indicate whether a document is relevant to one’s interests, allowing researchers to quickly browse through large masses of information.

The first attempt to produce automatic summaries is presented in [27], and relies on the distribution of words in a text to identify the important sentences. The promising results obtained by Luhn encouraged other researchers to apply similar approaches, in most cases in combination with other methods [14, 8, 23, 44, 43]. Alternative summarisation methods rely on presence of certain words [14] or phrases [39], discourse structure [34, 29, 12], anaphoric or coreferential links [9, 2] or lexical repetition [7, 11] to name but a few.

2.2 Anaphora resolution

Halliday and Hasan [17] describe anaphora as ‘cohesion which points back to some previous item’. Anaphora resolution is the process of resolving an anaphor, the pointing back expression, to the word or phrase it points back to, to an antecedent. If the antecedent and the anaphor have the same referent in the real world they are coreferential [31].

One of the most widespread type of anaphor and usually dealt with by computational linguists is pronominal anaphora with NPs as antecedent, a type of nominal anaphora [31]. In this case, an anaphoric expression is represented by a personal, possessive or reflexive pronoun, and the antecedent consists of one or several NPs. Due to their characteristics and high frequency of use, such anaphoric expressions pose great challenges to most of the fields in computational linguistics. The main reason for this is that these expressions do not carry much information on their own, and therefore have to be processed before they can be used.

Researchers in pronominal anaphora resolution have dedicated a great amount of effort to developing automatic resolution methods. Some of the methods rely mainly on one type of information, whereas others try to combine information from different sources. A syntax-based method is proposed in [20], whilst [21] relies mainly on semantic information for resolving personal pronouns. Centering Theory [16], a discourse theory of local coherence, is used as the only method to resolve pronouns in [10]. Methods which combine several types of information include methods employing preferences, indicators and constraints [24, 22, 3, 30] and machine learning based methods [1, 15, 38, 40, 5]. A comprehensive discussion of these methods can be found in [31], and the methods investigated here are briefly described in Section 3.2.

2.3 Pronominal anaphora resolution and automatic summarisation

Even though it was hypothesised that pronominal anaphora resolution could have a beneficial influence on the summarisation process, very few researchers have employed it to produce summaries. Often, pronominal anaphora resolution is part of a larger system which employs coreference resolution or lexical relations to produce summaries [4, 7, 9, 2]. However, these approaches do not try explicitly to assess the influence of pronominal anaphora resolution on the summarisation process. A small study on how pronominal anaphora resolution can influence a Swedish summarisation system is discussed in [19]. Manual evaluation on 10 newswire texts indicates that both the average important information in a summary and summaries’ coherence improve when a pronominal anaphora resolver is used.

Orašan [36] performed a series of experiments similar to the ones presented in this paper. The results reported there suggest that anaphora resolution can help the summarisation process, but due to the size of the corpus used in the investigation, it is difficult to make any generalisation. Steinberger et. al. [41] show how anaphora resolution can be used to improve the accuracy of a summariser based on latent semantic analysis, but they do not focus only on pronominal anaphora. As in the case of term-based summarisation, this method also uses the frequency of words to identify the important terms in a text, and uses them to extract important sentences. Evaluation on the CAST corpus [18] shows that anaphora resolution improves the results of the summariser significantly at both 15% and 30% compression rates.

3 Method

The method employed to produce summaries in this paper relies on terms and how they occur in sentences. The way this method works is described in Section 3.1. As already mentioned, the purpose of this paper is to assess whether the accuracy of the term-based summariser used here can be improved when it integrates an anaphora resolver. The anaphora resolvers employed in this research are described in Section 3.2, followed by the approach used to enhance the term-based summariser in Section 3.3.

3.1 Term-based summarisation

Term-based summarisation assumes that the importance of a sentence can be determined on the basis of the importance of the words it contains. To achieve this, each word is scored using term-weighting measures and then used to determine importance of sentences. The most common measures used to score each word are term frequency and TF*IDF. Moreover,
evaluation of several term-weighting measures on the
corpus used in this paper reveals that term frequency
and TF*IDF are the most appropriate ones [37].

Term frequency (TF) assigns to each word a score
equal to its frequency in order to indicate the topicality
of the concept represented by it. The main drawback
of this method is that it wrongly assigns high scores to
frequent tokens such as prepositions and articles. For
this reason, a stoplist is used to filter out such words.

$$\text{TF}(w) = \text{the frequency of word } w$$ (1)

The words awarded high scores by term frequency
are not necessary the most indicative of the
importance of a sentence. There are open class words
which appear frequently in a document but are not
good indicators of the topicality of a sentence. This
normally happens with words that occur frequently
not only in the document, but also in a collection
of documents. *Inverse document frequency* addresses
this problem by measuring the importance of a
word in report to how many documents from a
collection contain it, and assigning it a score inversely
proportionate to the number of documents which
include it. This means that words appearing in many
documents will not be awarded a high score. Because
document frequency is too weak to be used on its
own as a scoring method, it is usually combined with term
frequency. The formula used in this paper is:

$$\text{TF} * \text{IDF}(w) = \text{TF}(w) * \log \frac{N}{n_w}$$ (2)

where $N$ is the number of documents in the collection,
and $n_w$ is the number of documents in the collection
which contain the word $w$. As in the case of term
frequency, it was noticed that the performance of a
term-based summariser increases when a stoplist is
used to filter out stopwords, even though these words
obtain low scores.

3.2 Anaphora resolution

For the experiments described in this section the
Anaphora Resolution Workbench “a parameter-driven
environment for consistent evaluation of anaphora
resolution” [6] was used. This environment
implements several well-known anaphora resolution
algorithms and enables comparison between them on
the basis of the same preprocessing tools and data.
In this section, the knowledge-poor methods and the
three baselines implemented in this environment
were used. These methods are briefly explained next.

Kennedy & Boguraev (K&B): The anaphora
resolution method proposed by Kennedy and
Boguraev [22] adapts the method proposed by Lappin
and Leass [24] so it can be run without a parser,
and extends it with several other factors. The K&B
algorithm resolves third person pronouns with noun
phrase antecedents by employing a set of ten salience
preferences which rank candidates for antecedents.
Each preference has an initial weight which is used
to build coreference classes that contain pronouns
and their antecedents. Kennedy and Boguraev [22]
reports that the algorithm was evaluated on a corpus
containing 306 pronouns and the observed accuracy
was around 75%.

CogNIAC: is a high precision anaphora resolution
algorithm which can resolve a subset of anaphors
that do not require world knowledge or sophisticated
linguistic processing [3]. The algorithm relies on six
highly accurate rules to select the antecedent of a
pronoun. Because the rules apply to only some of
the pronouns, the original version of the algorithm
was extended to include two more rules which allow
it to operate in robust mode (i.e. it attempts to solve
every single anaphor). The robust algorithm achieved
77.9% accuracy on the MUC-6 corpus, whilst the high
accuracy non-robust algorithm achieves 92% precision
and 64% recall, but it resolves only some pronouns.

MARS: is a robust anaphora resolution method
which relies on a set of boosting and impeding
indicators to select the antecedent [30]. The algorithm
assigns scores to each candidate using the indicators,
and the candidate with the highest aggregate score
is selected as the antecedent for a pronoun. The
method was evaluated on technical manuals and a
hand-simulated evaluation reported results over 80%.
The method used in this paper implements the original
algorithm which does not include the extensions
proposed in [32] or a pleonastic pronoun recogniser.²

Baselines: In order to have a clear idea of how
effective the anaphora resolution methods are, three
baseline methods were used: BLAST selects the
closest candidate which agrees in gender and number
with the anaphor; BLASTSUBJ selects the most
recent subject which agrees in gender and number
with the anaphor; and BRAND randomly selects an
antecedent which agrees in gender and number with
the anaphor from the list of candidates.

3.3 Enhanced term-based sum-
marisation

The term-based summariser described in Section 3.1
relies on word frequencies to calculate the score of
a word. Because some of these words are referred
to by pronouns, the frequencies of the concepts they
represent are not correctly calculated. The enhanced
term-based summarisation method takes the output
of the anaphora resolver and increases the frequencies
of words referred to by pronouns, thereby producing
more accurate frequency counts. Section 5.2 evaluates
the improved term-based summarisation method.

4 Corpus

For the experiments described in this paper, a corpus
of journal articles published in the Journal of Artificial
Intelligence Research (JAIR) was built. The corpus
used here contains 65 texts with over 600,000 words
in total. In order to assemble this corpus, electronic
versions of the texts have been downloaded and

² This is due to the way the Anaphora Resolution Workbench
implements MARS.
converted to plain text. As the conversion was not perfect due to the presence of equations, formulae and other types of special formatting in the source, the resulting files were passed through a series of filters, which cleaned wrongly converted parts of the text and marked special information such as equations, tables, figures, footnotes and headers.

For the purpose of automatic summarisation, the corpus was automatically annotated with sentence boundaries, token boundaries and part-of-speech information using the FDG tagger [42]. In order to evaluate the performance of the automatic summarisation methods the author produced abstract was identified and extracted from the article.

A third of the corpus was also annotated with coreference information in order to evaluate the anaphora resolution methods used in this paper and for the experiment presented in Section 5.3. The difficulty of the annotation task and amount of time required to annotate a text made it impossible to apply the annotation to a larger part of the corpus. The annotation guidelines used for this purpose were derived from those proposed in [33], but instead of marking full coreferential chains, only parts of the coreference chains which contain nominal anaphoric pronouns were annotated. Therefore, if a chain did not contain a pronoun it was completely ignored. The annotation was applied using PALinkA [35], a multi-purpose annotation tool.

The annotation process first involved the automatic identification of all personal, reflexive and possessive pronouns, and annotation of these pronouns as potentially anaphoric. After this, each annotated pronoun was manually checked to see whether it was really referential, and that its antecedent was one or several NPs. For referential pronouns with NP antecedents, all the antecedents from the current paragraph and the most recent heading were identified. The reason for restricting the annotation only to these antecedents was due to the fact that all the anaphora resolution methods used here identify antecedents only from the current paragraph or the most recent heading, and therefore for the current investigation annotation of full coreferential chains would have been unnecessary. The corpus contains a total of 1873 referential pronouns, the vast majority are personal pronouns (1324), followed by possessive pronouns (502), with only a negligible number of pronouns being reflexive (47). The majority of referential pronouns are represented by different forms of the *it* pronoun, followed by different forms of *they* pronouns. As expected, the pronouns *he* and *she* have a very low frequency in the corpus.

## 5 Evaluation and discussion

In order to evaluate the effectiveness of anaphora resolution for automatic summarisation, the author produced summaries were considered the gold standard and the automatic summaries were compared to them. The measure used for computing the informativeness of an automatic summary is the cosine similarity between it and the author produced summary as proposed by Donaway et. al. [13]. Before computing the similarity, stopwords were eliminated.

For each of the 65 texts in our corpus summaries of 2%, 3%, 5%, 6% and 10% compression rates were produced. The reason for producing summaries of so many compression rates was to determine whether anaphora resolution influences the term-based summarisation method differently when it produces summaries of different lengths. Moreover, as can be seen in Section 5.2, the two term weighting methods investigated here lead to different results depending on the compression rate used.

In the rest of this section, the accuracy of the anaphora resolution methods employed here is first assessed to find out which one leads to the best results. After that, Section 5.2 investigates whether anaphora resolution can help term-based summarisation. The section finishes with an experiment where a simulated anaphora resolution system with predefined accuracy is used. The purpose of this experiment is to get further insights into how anaphora resolution can help term-based summarisation.

### 5.1 Evaluation of anaphora resolution

All anaphora resolution methods used in this paper are robust. For this reason, the only measure used in the evaluation is *success rate*, computed as the number of correctly resolved anaphors divided by the number of anaphors identified by the system [31]. Table 1 contains the average success rate obtained by different anaphora resolution methods on the coreferentially annotated corpus.

<table>
<thead>
<tr>
<th>Method</th>
<th>Average</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>MARS</td>
<td>0.512</td>
<td>0.080</td>
</tr>
<tr>
<td>K&amp;B</td>
<td>0.435</td>
<td>0.088</td>
</tr>
<tr>
<td>BLAST</td>
<td>0.507</td>
<td>0.077</td>
</tr>
<tr>
<td>BRAND</td>
<td>0.166</td>
<td>0.066</td>
</tr>
<tr>
<td>BLASTSUBJ</td>
<td>0.115</td>
<td>0.061</td>
</tr>
<tr>
<td>CogNIAC</td>
<td>0.084</td>
<td>0.058</td>
</tr>
</tbody>
</table>

Table 1: The average success rate obtained by different anaphora resolution methods

For each of the 65 texts in our corpus summaries of 2%, 3%, 5%, 6% and 10% compression rates were produced. The reason for producing summaries of so many compression rates was to determine whether anaphora resolution influences the term-based summarisation method differently when it produces summaries of different lengths. Moreover, as can be seen in Section 5.2, the two term weighting methods investigated here lead to different results depending on the compression rate used.

In the rest of this section, the accuracy of the anaphora resolution methods employed here is first assessed to find out which one leads to the best results. After that, Section 5.2 investigates whether anaphora resolution can help term-based summarisation. The section finishes with an experiment where a simulated anaphora resolution system with predefined accuracy is used. The purpose of this experiment is to get further insights into how anaphora resolution can help term-based summarisation.

For this reason, the only measure used in the evaluation is *success rate*, computed as the number of correctly resolved anaphors divided by the number of anaphors identified by the system [31]. Table 1 contains the average success rate obtained by different anaphora resolution methods on the coreferentially annotated corpus.

The success rate of all the methods evaluated here is much lower than that reported by their authors. There are two justifications for this. First, the evaluation performed by the authors was an evaluation of the algorithm and not an evaluation of a practical system. This means that the algorithms were either hand-simulated or they processed manually prepared data. In contrast, the evaluation presented here was fully automatic and the systems had to deal with errors introduced by preprocessing steps such as part-of-speech tagging and NP extraction. The second reason for obtaining lower results is that the anaphora resolution methods used here were not designed to deal with texts from the scientific domain: MARS was developed for the technical domain and includes indicators specific for this domain, CogNIAC was tested on the MUC-6 texts, and K&B was evaluated

---

3 A system is considered robust if it tries to resolve all the pronouns which are anaphoric. Some systems such as the non-robust version of CogNIAC resolve only a part of these pronouns because of the way they were designed.
Table 2: The average informativeness of summaries produced by the improved summarisation method

<table>
<thead>
<tr>
<th>Method</th>
<th>2%</th>
<th>3%</th>
<th>5%</th>
<th>6%</th>
<th>10%</th>
</tr>
</thead>
<tbody>
<tr>
<td>No anaphora</td>
<td>0.315</td>
<td>0.449</td>
<td>0.461</td>
<td>0.467</td>
<td>0.484</td>
</tr>
<tr>
<td>Blast</td>
<td>0.351</td>
<td>0.476</td>
<td>0.495</td>
<td>0.498</td>
<td>0.511</td>
</tr>
<tr>
<td>Blastsubj</td>
<td>0.358</td>
<td>0.481</td>
<td>0.493</td>
<td>0.501</td>
<td>0.514</td>
</tr>
<tr>
<td>Brand</td>
<td>0.357</td>
<td>0.478</td>
<td>0.494</td>
<td>0.499</td>
<td>0.512</td>
</tr>
<tr>
<td>CogNIAC</td>
<td>0.354</td>
<td>0.478</td>
<td>0.493</td>
<td>0.500</td>
<td>0.512</td>
</tr>
<tr>
<td>K&amp;B</td>
<td>0.357</td>
<td>0.481</td>
<td>0.493</td>
<td>0.501</td>
<td>0.511</td>
</tr>
<tr>
<td>MARS</td>
<td>0.355</td>
<td>0.480</td>
<td>0.494</td>
<td>0.500</td>
<td>0.513</td>
</tr>
<tr>
<td>Perfect</td>
<td>0.352</td>
<td>0.495</td>
<td>0.511</td>
<td>0.517</td>
<td>0.521</td>
</tr>
<tr>
<td>TF*IDF</td>
<td>0.314</td>
<td>0.439</td>
<td>0.457</td>
<td>0.472</td>
<td>0.498</td>
</tr>
<tr>
<td>Blast</td>
<td>0.350</td>
<td>0.475</td>
<td>0.494</td>
<td>0.500</td>
<td>0.514</td>
</tr>
<tr>
<td>Blastsubj</td>
<td>0.351</td>
<td>0.476</td>
<td>0.495</td>
<td>0.498</td>
<td>0.511</td>
</tr>
<tr>
<td>Brand</td>
<td>0.357</td>
<td>0.478</td>
<td>0.494</td>
<td>0.499</td>
<td>0.512</td>
</tr>
<tr>
<td>CogNIAC</td>
<td>0.354</td>
<td>0.478</td>
<td>0.493</td>
<td>0.500</td>
<td>0.512</td>
</tr>
<tr>
<td>K&amp;B</td>
<td>0.357</td>
<td>0.481</td>
<td>0.493</td>
<td>0.501</td>
<td>0.511</td>
</tr>
<tr>
<td>MARS</td>
<td>0.355</td>
<td>0.480</td>
<td>0.494</td>
<td>0.500</td>
<td>0.513</td>
</tr>
<tr>
<td>Perfect</td>
<td>0.352</td>
<td>0.495</td>
<td>0.511</td>
<td>0.517</td>
<td>0.521</td>
</tr>
</tbody>
</table>

5.2 Evaluation of enhanced summarisation method

As already mentioned, it is assumed that term-based summarisers do not achieve a very high performance because they ignore the fact that some words are referred to by pronouns, and therefore their frequency is not accurately computed. In this section, the three anaphora resolution methods and three baselines evaluated in the previous section are incorporated into the term-based summariser.

The results of the term-based summarisation methods augmented with information from anaphora resolvers are presented in Table 2. The row labeled No anaphora indicates the informativeness of summaries when no information from an anaphora resolver is incorporated in the term-based summariser, whilst the row Perfect corresponds to an anaphora resolver with a success rate of 100%. The results in the Perfect row were obtained only for the texts annotated with coreference information because the manual annotation is considered to be the output of a perfect anaphora resolver.\(^4\) The rest of the rows indicate the informativeness of the summary when an automatic anaphora resolver was integrated into the system and were calculated on the whole corpus.

5.3 Robust anaphora resolver with predefined accuracy

Evaluation of the anaphora resolvers presented in Section 3.2 showed that they often resolve pronouns to the wrong antecedent. As a result, some concepts have their scores wrongly increased. The anaphora resolver simulated in this section tries to perform in the same manner as the automatic anaphora resolvers investigated in Section 3.2, by boosting the frequency scores of both correct and incorrect antecedents, but it is designed in such a way that its success rate can be controlled. For this experiment, the success rate of this resolver was increased from 10% to 100% in 10% increments. In order to achieve this, a predefined percentage of correct (pronoun, antecedent) pairs were selected from each text. For the rest of the pronouns, wrong antecedents were selected in order to introduce errors. This process was repeated 100 times for each text and for each success rate value to ensure fairness and reliability of the experiment. The manual annotation was used to simulate this anaphora resolver, and so the experiment was carried out only on the coreferentially annotated texts. Figures 1 and 2 present the results of the experiment.

The results of these experiments are in line with the results reported in Section 5.2, but still contain some unexpected features because they show that even if only 10% of the pronouns are correctly resolved, the

\(^4\) We acknowledge the fact that errors in the annotation can limit the degree of ‘perfectness’ of the output.
results of the automatic summariser are significantly better. The next significant improvement is obtained only for an anaphora resolver which achieves at least 60% success rate and is used with term frequency. For the automatic summariser which uses TF*IDF, it is necessary to have an anaphora resolver which achieves around 80%-90% success rate to have a noticeable improvement.

6 Conclusions

This paper has investigated the influence of pronominal anaphora resolution on term-based summarisation. The underlying hypothesis was that by incorporating an anaphora resolver into the term-weighting process it is possible to obtain more accurate frequency counts of concepts referred to by pronouns. To this end, three robust anaphora resolvers and three baselines were incorporated into two term-weighting measures, which were in turn used by a term-based summariser. Comparison of the informativeness of summaries produced by this improved term-based summariser revealed that there is no correlation between the informativeness of a summary and the performance of the anaphora resolver used to improve the frequency counts. Despite this, the results clearly indicate that the summarisation process benefits from anaphora resolution.

The beneficial influence of anaphora resolution on term-based summarisation was further investigated by performing an experiment with a simulated anaphora resolver with controlled success rate. The results of the experiment show that due to the increase of scores for both correct and incorrect antecedents, a significant improvement of the summaries’ informativeness is noticed only when accuracy of the resolver is between 60% and 80%, depending on the term-weighting method. This explains why no difference was observed for the relatively poor performance of anaphora resolvers investigated here.

The integration of an anaphora resolver into the term-based summariser also reveals some interesting results. Without anaphora resolution, term frequency leads to the best results only for 2% and 3% compression rates. Once an automatic anaphora resolution is integrated into the term-based summariser the differences between summaries produced using term frequency and those produced using TF*IDF at 5%, 6% and 10% become negligible. Moreover, if a perfect anaphora resolver is used, the summariser which uses term frequency always performs significantly better than the summariser which uses TF*IDF.

This paper has focused only on how pronominal anaphora resolvers can be used in the summarisation process. For the future it would be interesting to extend this research to other types of anaphoric expressions such as definite descriptions. Another interesting development of this paper would be to use other evaluation methods such as ROUGE [26] for measuring the informativeness of summaries in order to find out whether the findings change. The summarisation methods employed in this paper are rather simple and do not necessary reflect the state of the art of summarisation methods. In the future it is planned to evaluate the influence of pronominal anaphora resolution on other summarisation methods which could benefit from this information.

References


