Building better corpora for summarisation

Laura Hasler, Constantin Orasan and Ruslan Mitkov
Research Group in Computational Linguistics
School of Humanities, Languages and Social Sciences
University of Wolverhampton
{L.Hasler, C.Orasan, R.Mitkov}@wlv.ac.uk

1. Introduction

Annotated corpora have proved essential in many areas of NLP, and over the years have been successfully exploited for a wide range of tasks in Computational Linguistics, including part-of-speech tagging, parsing and information extraction. One field in which they have been particularly useful is automatic summarisation. Within this field, annotated corpora are mainly used for machine learning, to learn patterns for the extraction of important (and other) information from texts, as well as for the more complex task of evaluation of summarisation methods (Edmundson, 1969; Kupiec, Pederson and Chen, 1995; Zechner, 1996; 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. 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. There are semi-automatic (Orasan, 2002) and automatic (Jing and McKeown, 1999; Marcu, 1999) ways to annotate corpora, but given that we are investigating new types of information to be marked, manual annotation is most appropriate here. Despite the fact that they are vital to the field, corpora annotated for summarisation are relatively sparse, and those resources which do exist do not contain as much information as they could.

This paper presents an enhanced annotated corpus which differs from the majority of available resources in that it contains more information. In addition to containing information about the importance of the sentences, we also indicate parts which can be removed from sentences marked as *essential* or *important*, and provide a different label for those sentences which are not significant enough to be marked as important in their own right, but which have to be considered as they contain information essential for the understanding of the content of other sentences marked as *essential* / *important*. These last two types of information can give us an insight into the conciseness and coherence of summaries, respectively. Our corpus also contains annotations for *linked* sentences, where both sentences are considered *important* or *essential*, but one relies on the other to be completely understood. The building of the corpus is part of the "CAST – A Computer-Aided Summarisation Tool" project, which develops a tool to aid humans in creating summaries.

The rest of the paper is structured as follows: Section 2 details previously annotated corpora for summarisation. The next section discusses our corpus. The annotation guidelines are described in Section 4. Section 5 comprises our results and discussion, followed by our conclusions and future work.

2. Previous attempts at annotating corpora for summarisation

Edmundson (1969) was the first to use corpora for summarisation. Human judges annotated the most important sentences from a corpus consisting of 200 documents in the areas of physics, life science, information science and humanities. Guidelines were used to ensure consistency, and the annotators were advised to minimise redundancy and maximise coherence.

Kupiec, Pederson and Chen (1995) took a different approach in order to make the task of annotation slightly easier. Still using human annotators and providing guidelines for the annotation, sentences from summaries were aligned with sentences from the corresponding full texts. 79% of sentences appearing in the summaries could be matched with sentences in the full texts in their corpus of 188 scientific and technical documents. Teufel and Moens (1997) took a similar approach but achieved a much lower percentage of alignment (31.7%), using a corpus of 202 Computational Linguistics articles. This highlights the effect of text type, for both the alignment method and for annotation with regards to summarisation in general.

Although work has been done using clauses or elementary discourse units (Marcu, 1997) as the standard unit of extraction, as well as allowing the human judge to determine the most appropriate

units (Tsou, Lin and Lan, 1997), the corpora described above are all concerned with sentences, and sentences may not be the best textual units on which to solely base a summarisation system. We decided to select sentences due to the simplicity of defining the unit, but we also removed parts from these sentences so that only the most important information is present in the extract.

3. The corpus

The first step in building any corpus is to select the texts for inclusion in it. A description of these texts and the properties of the corpus we built can be found in Section 3.1. To encode the added information present in our corpus, a suitable annotation scheme was developed which accounts for all the types of information that it was necessary to include. This annotation scheme is presented in Section 3.2. Given the complexity of this scheme, the annotation cannot be applied using a simple text processor, but a specially developed tool has to be used instead. The tool is briefly described in Section 3.3.

3.1 Texts used in the corpus

One purpose of the CAST project is to develop summarisation methods for newswire texts. In order to tune and evaluate these methods, a corpus containing these types of texts had to be built. The texts included in our corpus were taken from the Reuters Corpus (Rose et. al., 2002). In addition to these, we also included a few popular science texts from the BNC (Burnard, 1995). We incorporated this latter type of text because it contains features from both the newswire and scientific genres, and the next step in our project is to extend our corpus to include scientific documents. By annotating these texts now, we hope to identify potential problems which could arise during the annotation of scientific documents later in our project.

	Newswire	*Popular science*	*Total*
No. of texts selected	147	16	163
No. of texts annotated by at least 2 annotators	31	12	43
No. of texts annotated by at least 3 annotators	7	-	7
Total words annotated	117378	28095	145473
Total sentences annotated	5214	1370	6584

Table 1: Corpus statistics

Table 1 summarises the statistics of our corpus. The texts with multiple annotations can be used to measure the interannotator agreement, and as a direct result of this, the difficulty of the task. The results of the comparison are presented in Section 5.3.

For the annotation process we used four annotators; three graduate students and the first author of this paper. Three of the annotators were native English speakers, and the fourth had advanced knowledge of English. Before starting the annotation process, the annotation guidelines were explained to the annotators (see Section 4.2 for the guidelines).

During the annotation, it became apparent that a selection of the texts were concerned with the same story, but they were written from a different perspective. We annotated some of these texts to give us an insight into whether the angle/order/authorship of a news text has any bearing on those sentences which are considered important. Some preliminary findings are presented in Section 5.7.

The fact that the Reuters Corpus is freely available for research makes it possible for us to apply the same policy to the part of that corpus annotated by us. The newswire part of the corpus can be freely downloaded from the web page of the CAST project (http://clg.wlv.ac.uk/projects/CAST/). Given that the science texts are taken from the BNC it is not possible to distribute this part of the corpus.

When we started developing the corpus, we wanted to produce a resource from which the whole research community could benefit. In light of this, we decided to use XML, an encoding widely used by researchers in Computational Linguistics. Given that an XML encoded file can be quite difficult to read and annotate, we had to use a special tool (see Section 3.3).

3.2 Annotation scheme

Section 1 states that our corpus is annotated with more information than the just the importance of the sentences. For this reason, it was not enough to indicate the importance of each sentence in a simplistic way (e.g. a character/symbol before each sentence); we had to develop a more complex annotation scheme. In this section we present the annotation scheme used in our corpus.

```
<P ID="P3"><S ID="S3"><EXTRACT COMMENT="" ID="0" IMP="ESSENTIAL"><W ID="W19">Impl
ied</W><W ID="W20">volatilities</W><W ID="W21">for</W><W ID="W22">U.S.</W><W ID="
W23">interest</W><W ID="W24">rate</W><W ID="W25">options</W><W ID="W26">were</W><
W ID="W27">steady</W><W ID="W28">to</W><W ID="W29">higher</W><W ID="W30">in</W><W
ID="W31">quiet </W><W ID="W32">dealings</W><PUNCT>,</PUNCT><W ID="W33">with</W><W
ID="W34">the</W><W ID="W35">underlying</W><W ID="W36">futures</W><W ID="W37">post
ing</W><W ID="W38">mild</W><W ID="W39">declines</W><W ID="W40">on</W><W ID="W41">
the</W><W ID="W42">day</W><PUNCT>.</PUNCT></EXTRACT></S></P> <P ID="P4"><S ID="S4
"><EXTRACT COMMENT="" ID="1" IMP="ESSENTIAL"><W ID="W43">Volatilities</W><W ID="W
44">at</W><W ID="W45">the</W><W ID="W46">short</W><W ID="W47">end</W><W ID="W48">
of</W><W ID="W49">the</W><W ID="W50">yield</W><W ID="W51">curve</W><W ID="W52">we
re</W><W ID="W53">bid</W><W ID="W54">for</W><W ID="W55">a</W><W ID="W56">second</
W><W ID="W57">day</W><PUNCT>,</PUNCT><W ID="W58">as</W><W ID="W59">the</W><W ID="
W60">market</W><W ID="W61">continues</W><W ID="W62">to</W><W ID="W63">adjust</W><
W ID="W64">for</W><W ID="W65">a</W><W ID="W66">higher</W><W ID="W67">interest</W>
<W ID="W68">rate</W><W ID="W69">environment</W><W ID="W70">after</W><W ID="W71">a
</W><W ID="W72">prolonged</W><W ID="W73">period</W><W ID="W74">of</W><W ID="W75">
stability</W><PUNCT>,</PUNCT><REMOVE COMMENT="" ID="6"><W ID="W76">Eurodollar</W>
<W ID="W77">analysts</W><WID="W78">said</W><PUNCT>.</PUNCT></REMOVE></EXTRACT></S
></P>
```

Figure 1: An extract from our corpus

We marked three types of information: the importance of the sentence (*essential* or *important)*, *links* between sentences, and sections which can be *removed* from the marked sentences. The importance of the sentence was marked using the `<EXTRACT ID="XXX">` tag, where the `ID` attribute identifies the tag uniquely. This tag has another attribute, `IMP`, which indicates the importance of the sentence and can take the following values: `ESSENTIAL` for sentences which are considered *essential* by the annotators, and `IMPORTANT` for those deemed *important*. In addition there is a third value for the `IMP` attribute, `REFERRED`, which indicates that a sentence is neither *essential* nor *important*, but required for the comprehension of a marked sentence. If a sentence is not marked in any way it is considered to be unimportant. The links between the sentences are marked by an empty tag `<LINK REFERRED="XXX"/>` which indicates that the `EXTRACT` tag which contains it is linked to the `EXTRACT` tag with the `ID` indicated by the attribute `REFERRED`. The sections from the selected sentences which are redundant or irrelevant are indicated by the `<REMOVE>` tag. All these tags have an optional attribute `COMMENT`, where the annotators can provide comments on the annotation process.

Figure 1 presents a short extract from our corpus.

3.3 Annotation tool

In order to achieve the annotation described in the previous section, a multi-purpose annotation tool was used. In addition to helping with the marking process, the tool indicates throughout the annotation process what proportion of the file has been selected, and registers the time taken to annotate a file.

The tool used to perform the annotation, *PALinkA*, offers a user-friendly graphical interface with different colours for the different types of information, which makes it easier to distinguish between them, as well as speeding up the annotation process. *PALinkA* is easy to use, even for non-computer experts. To mark a unit of text, the annotator uses the mouse to indicate the boundaries of the unit, the tag assigned to the unit, and whatever attributes are required by the tag. To avoid errors, some attributes such as unique IDs and references are determined automatically by the tool.

PALinkA is a multi-purpose annotation tool which can be used to annotate a variety of phenomena. The set of tags which can be used in the annotation is specified by a preferences file loaded in the tool before annotation starts. For our project, a preferences file containing the annotation scheme described above was used.[1]

4. Annotation guidelines

Guidelines are essential for consistent and reliable annotation of texts. Annotation guidelines are a set of rules that indicate to the annotator which parts of a text should and should not marked, and for what types of information. They contain examples to aid the annotator in this process, and should be simple to follow but include enough detailed information to ensure strict adherence to them. Without these guidelines, the human annotator would have the freedom to indicate whatever parts of the text they wanted, regardless of their actual relevance to the task in hand. By developing guidelines to which annotators can adhere, we can reduce the amount of discrepancies between annotators who work on the

[1] The tool and the preferences file can be downloaded from the project's web page.

same text, as well as across a range of texts annotated by the same person, therefore ensuring consistent and valid annotation. In the field of summarisation these guidelines are especially important, given the notorious subjectivity of what is considered "important" enough to be included in a summary of a text.

4.1 Other examples of guidelines

Mitkov et al. (2000) suggest a "master" strategy for annotation of coreference and anaphora, which comes from a combination of different annotators' approaches, and this can be adapted to summarisation as it is a fairly general strategy (changing those parts pertaining to the particular field) that could be extended to any area. These guidelines again seem to be based on general common sense, and provide a sound base for any type of annotation. These general guidelines, adapted for summarisation, are:

- 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).
- Ensure that the annotation is done in one intensive period (relatively easy in our case, as the texts are news articles and not too long), as sporadically annotating a file can lead to the annotator having to re-read the document for familiarisation several times.
- Comment on troublesome cases of annotation and then discuss them with other annotators to decide upon the best solution to tackle them in future.

Cremmins (1982) provides a number of pointers for writing abstracts. Although these are meant for human-written abstracts, they can also be valuable when we need a human annotator to mark those relevant sections in a text which contain information that could be used in a summary. Amongst other things, Cremmins advocates that the abstractor scan the whole text before deciding on important information (similar to the first "master" strategy point above), as well as comparing different versions of the same information to ensure that there is no over-emphasis of relevant information at the expense of other equally (or possibly more) relevant information through careless or inadequate reading of the text. As is evident from examining the guidelines below, this last point regarding non-repetition of information is central to our approach.

Mitkov, Le Roux and Descles (1994) propose a set of rejection rules for knowledge-based automatic abstracting in the sub-language of elementary geometry. Some of these rules (such as elimination of text in brackets and text in quotation marks) may be appropriate for our corpus, but they need expanding on (see Section 4.2). Other rules they present which are similar to ours here concern the elimination of subordinate clauses, although we argue that not all subordinate clauses need to be removed, and text containing examples.

4.2 Our guidelines for annotation

In addition to the general strategies mentioned in Section 4.1, more specific guidelines were formulated based on an analysis of the type of texts we wanted to annotate for our corpus. As we work with news texts, the guidelines are slightly different to guidelines for texts in other areas, due to the features of newswire as a genre. It is important to remember that there may always be cases which prove to be the exception to the rule, which is why we provide room for comments and explanations as part of the annotation tool. A length restriction of 30% was imposed on the amount of sentences that could be marked; 15% to be marked as *essential* and 15% as *important*. The outline of our annotation guidelines, and the motivations behind them are described below.

The annotators were instructed to identify the main topic of the text and mark sentences (using the `<EXTRACT IMP="ESSENTIAL">` tag in *PALinkA*) which gave *essential* information about the topic, keeping as close to 15% of the full text as possible. A newspaper text is usually "about" one main topic, that is, it tends to concentrate on one main focus throughout the length of the text. This means that the information which is considered suitable for inclusion in a summary (i.e. marked as *essential* / *important*; Section 5.1 elaborates more on the need for this distinction) should generally relate to this topic.

Sub-headings generally summarise the text which follows and are subsequently 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 instructed to mark sub-headings as long as they are relevant to the main topic of the text.

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

(S1) For film lovers the Festival's the place to be in September.

(S2) It grows from strength to strength each year.

In our annotation process, if S2 is marked, the sentence containing the referent for "it", should be marked in some way. If S1 is important enough, it is already in the list of selected sentences, otherwise, it has to be highlighted with the <EXTRACT IMP="REFERRED"> tag. The links between sentences need to be indicated by the <LINK> tag.

Tables, figures and examples (including constructions starting with "e.g.", "for example", "such as", "like", "for instance" etc.) are not necessary in a summary, and therefore should not be marked, 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. Reported speech should not be included unless it presents vital and new information concerning the main topic of the text which is not presented elsewhere. Reported speech in news texts tends to provide opinions or statements to emphasise points already made in the article and does not usually warrant marking. However, it is important to distinguish between this speech and other text in quotation marks which may be important, for example:

(S3) But its "killer app" is reinventing how the software industry works.

Sentences which contain 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, we can look at the following pair, where (S4) is preferable (in most cases) to (S5):

(S4) Inflow of export proceeds picking up: $300 million likely by February 15.

(S5) 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.

Within these marked sentences, there may be parts which are not relevant to the overall importance. 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. This was to be done using the <REMOVE> tag in *PALinkA*. So within sentences already marked as *essential*, irrelevant subordinate clauses should be marked as *remove*. The "which" clause could be removed from the following sentence if it is not important within the context:

(S6) 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. For example, bracketed text would be considered important if it is an abbreviation which will replace a noun phrase later in the text, as in:

(S7) A Poverty Reduction and Growth Facility (PRGF)…

(S8) It is intended that PRGF-supported programmes…

Adjuncts (not single words) which specify dates, times, places etc. should be marked for removal, unless they are vital to the main topic, likewise examples (see above) should be removed, as should phrases such as "in addition to…", due to…" and "compared to…" which elaborate on information, and constructions like "a spokesman said", "it was claimed", "she told *The Guardian*" and "he explains".

Once the annotators had completed this selection process of *essential* and *remove*, they were advised that if the total amount of marked text (the <REMOVE> tag subtracts that part of the sentence marked

as *remove* from the whole percentage of the marked text) is substantially below 15% of the full text, 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, both as the annotation proceeded as well as more general comments at the end of the document.

5. Results and discussion

This section details the main findings resulting from an analysis of our corpus. Here we discuss the reason for maintaining a distinction between *essential* and *important* tags to mark sentences, correlations between the overall text length and the time taken to annotate that text, interannotator agreement for sentence classification, the distribution of marked sentences, a comparison of *removed* parts of sentences, sentences marked as *linked* or *referred*, the annotation of similar texts and an assessment of our annotation guidelines.

5.1 Distinction between tags

We felt it was important to distinguish between those sentences which were *essential* to the general understanding of a text, and those which were *important*. This distinction enables us to produce two types of summary – a shorter one (*essential* sentences only) and a longer one comprising *essential* and *important* sentences. It has been shown (Marcu, 1997) that human judges agree more consistently on those units in a text which they consider to be either *very important* or *unimportant* than when the same judges are asked to identify textual units which they consider to be *important*. As we keep this distinction in our annotation, we will also be able to test for ourselves whether this is the case in our corpus, and to pave the way for further investigation of the subjectivity of the notion of importance within summarisation. Marcu (1997) asks judges to assign a label to each unit in the text, whereas we are concerned with the selection of those units considered *essential* and *important* within a certain compression rate from a text.

5.2 Text length / time

During the annotation process, we recorded the time take to annotate the text, and measured the text length in words. Our hypothesis was that we would be able to identify a link between the length of the text and the time taken to annotate it. The time measurement was especially accurate as the annotator could stop the clock on the tool whenever they paused during the annotation. We computed the Pearson correlation between text length and time and found a strong correlation. For three out of the four annotators, the correlation was found to be significant at the 0.01 level. For the fourth annotator, the correlation was slightly lower, being significant at the 0.05 level. This correlation indicates that the texts annotated were of similar complexity, as the processing time depended on the length.

5.3 Interannotator agreement for classifying the sentences

As shown in Section 4.2, the annotators had to identify those sentences which contained information worth including in a summary. They were asked to label 15% of the whole text as *essential*, and another 15% as *important*. In order to assess the quality of the annotation we computed the interannotator agreement using the Kappa statistic (Siegel and Castellan, 1988). Kappa indicates whether there is any agreement among annotators, and takes into account not only the classification provided by the annotators, but also the possibility of their agreeing by chance. Kappa takes values between 0 and 1, and it is considered that a value over 0.8 indicates high agreement between annotators, whereas values between 0.68 and 0.8 indicate moderate agreement. Usually values below 0.68 are taken to indicate little or no agreement.

Part of our corpus was annotated by two or three annotators, so it was possible to compute their agreement on these files. In order to be able to compute the Kappa statistic, we considered the annotation process as a 3-class taxonomy of *essential*, *important* and *unimportant* sentences.[2] We computed the agreement for each text, as well as on the whole set of texts annotated by two or three annotators. The agreement between both two and three annotators was very low, and varies

[2] The annotators were not asked to identify the unimportant sentences, a sentence was considered unimportant if it was not marked in any way.

considerably from one text to another, suggesting that there is little agreement among our annotators. The overall agreement for two annotators was 0.24, and for three annotators 0.23**.**

We analysed the files in an attempt to discover the reason for the low agreement and noticed that sometimes the annotators focused on different sub-topics which they saw as relevant to the main topic. More importantly, we could see that a high proportion of the disagreement was because the annotators did not agree, in a large number of cases, as to whether a sentence was *important* or *essential*. The same sentences were marked as to be included, but with a different tag by different annotators. For this reason, we collapsed the *important* and *essential* classes into a single class and recomputed the Kappa statistic. In this case the value increased to 0.35 for texts marked by both two and three annotators, but the agreement is still very low.

One possible reason for this low agreement is the fact that the annotators were not experts or professionals in summarisation, and so may not have been as accurate or confident in their annotations as they could have been otherwise. Another explanation is that some of the annotators found parts of the guidelines not explicit enough to ensure consistent annotation (see Section 5.8 for an assessment of the annotation guidelines). Section 6 discusses future avenues for investigating and improving these issues.

Further investigation, which considered the extracted sentences as summaries, revealed that the annotators selected different sentences, but covered more or less the same information. In order to see how much the annotators agreed on the information (as opposed to the sentences) they selected, we computed the similarity between the sentences marked by the annotators. The similarity was computed using the cosine distance (Salton and McGill, 1983). The values for this distance indicate higher agreement between annotators, the average for *essential* and *important* sentences was 0.73. We also computed the similarity for *essential* sentences only, given that there are two possible extracts from each text. When we did this, the agreement decreased to 0.57. These figures prove that it may not be 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. The similarity measure relies on words, not on their senses. In future, we plan to improve its accuracy by taking into account synonymy relations between words. As a result of this change we expect an increase in the similarity.

5.4 Distribution of marked sentences in texts

It is a view widely held that the important sentences in news texts are located at the beginning of the text. An analysis of the distribution of marked sentences in our corpus proved this not to be the case. We divided the text into three equal parts (beginning third, middle third and end third) to see where the majority of sentences were chosen from. In the whole corpus, the highest proportion of sentences selected came from the first two parts of the text, with a slightly higher proportion coming from the middle third (38.7%) than the beginning third (35.4%). A lower proportion of sentences were marked in the end third of the texts (25.9%), and although this was not as high as the other two thirds, it was still high considering the general claim that such sentences can usually be found at the start of a text. Although the distribution of our annotators' sentences does not match with the generally held view, they do match up with each other. The texts marked by all four annotators strongly displayed these patterns, despite the fact that annotation was carried out independently.

5.5 Comparison of removed sentence segments

Given the low agreement between the sentences marked for extraction and the fact that we did not provide a strict definition for the units to be removed, we could not apply a similarity measure like the one used in Section 5.3. Instead, we tried to identify recurring patterns manually. An analysis of the removed sentence segments demonstrates definite patterns with regards to the types of constructions that are likely to contain irrelevant information within important sentences and from which we can learn what kind of information we would not want to include 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 serve to qualify previous information given in a text. For this reason, we chose not to include qualifiers in general as a category as this would subsume other important groups.

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 adverbial phrases ("during…") acting as temporal adjuncts. Phrases introduced by "elsewhere" and "meanwhile" also appeared with this kind of information. The second

category contains removed information concerning the coverage of events via reported speech. Whilst the actual speech or the gist of what was said may be important, who reported this was not usually important. Verbs of reporting, such as “said”, “writes”, “revealed”, “reported”, “pointed out”, “alleged”, “added that”, “according to” either preceded or followed by a nominal group (generally person, organisation, type of report) are characteristic of this information.

General information irrelevant to the main topic of the text was mainly contained in subordinate clauses. This provided us with another group of removed sentence segments. These clauses contained all kinds of different information, but were generally exemplified by starting with “where”, “when”, “with”, “as”, “after”, “due to”. Interestingly, there were a number of co-ordinate clauses starting with “and” and “but” which were also removed. This occurred when the second co-ordinate clause elaborated on 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 which qualified nominal phrases made up a fourth category. The main types of relative clause began with “which”, “who” and “that”, as well as the verb “including”, and served to give further information about an entity in the text, not needed in a summary. Examples typically started with “usually”, “such as”, “like”, “especially”, and gave us a fifth group of removed text. Another group was text in between or following dashes, which again qualified preceding information or entities. The seventh and final category of removed segments of sentences comprised other ways of elaborating textual units, such as apposition of nominal phrases, and the use of the prepositions “to”, “from”, “by” and “for” to give more information about a change that had occurred. Words introducing counter-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.

5.6 Linked and referred sentences

As explained in Section 4.2, the reasoning behind marking linked and referred sentences in our corpus is that they will give us an insight into how we can improve the coherence of summaries by examining the types of links that occur between sentences.

Preliminary results show that the number of sentences which need to be extracted only because they contain entities referred to in other sentences is very low. In our corpus, we identify 42 such sentences, 0.64% of the total number of sentences. This suggests that in general the set of important sentences contains enough information to be understandable on its own. One explanation for this low number could have been that there were no links between sentences. After counting the number of links marked between tagged sentences, such an explanation was quickly ruled out as we found that over 18% of the selected sentences contained links to one or more other sentences. As expected, most of the links were needed because of pronouns and noun phrases, a much lower number were due to anaphoric verbs. Given that the annotators were not asked to mark the relations explicitly, it is not possible to obtain these statistics automatically. In the future, we intend to detail these relations to improve the investigation.

5.7 Annotation of similar texts

As mentioned in Section 3.1, some of the texts in our corpus were based on the same sequence of events, but were slightly different versions of these events. One annotator marked important sentences in 5 versions of one text, 4 of another, and 2 of a third text in order to assess the impact of different versions of the same story on marking important information. To do this, we computed the similarity between the different versions of the texts using the cosine distance, as we did for the other texts (see Section 5.3).

Preliminary results show that when the same annotator is instructed to mark the important information in different versions of the same texts, the similarity between them is not very high. When this is compared to the average similarity for the same text annotated by different people, we can see that it is considerably less: 0.44 compared to 0.73. This shows that differences in texts primarily “about” the same thing have a considerable effect on the information that is considered important. This is due to the different locations of certain information within the texts, how much space is devoted to that information, and what subsidiary information is introduced to flesh out the main points of the story. The title is a good indicator as to the main focus of each text, including both the structure and the content.

5.8 Assessment of guidelines

A discussion with the annotators generated several points with regards to the suitability of the guidelines developed in order to help them in their annotation. As we can see, although the guidelines seemed generally helpful and the comments positive, we recognise that there are ways in which they could be improved to ensure more consistent annotation in the future.

In general, the annotation guidelines were helpful for the task. They 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 may be faced when annotating and are concise and clear. They are strict with regards to the amount of sentences which could be extracted, however, this may have been slightly too long considering a summary was meant to be made from it. According to the guidelines for summarisation the length restriction is 15% *essential* sentences and 15% *important* sentences. In some cases, it was difficult to adhere to the 15% restriction as it was not easy to distinguish between the *important* and *essential* sentences in a text, and more than 15% could have been considered *essential* or *important*.

An analysis of the corpus showed that there were many exceptions to the rules stated by the guidelines; this had, however, been pre-empted and was the reason for the inclusion of "unless vital to the main topic" on the end of many of the instructions. Dates and times were important in some contexts, such as articles where time was an important factor, as was reported speech, and sometimes even the speaker, especially where there were conflicting arguments from two different organisations or people, and examples were occasionally necessary to emphasise an important point. Sub-headings were not very relevant when they did occur, which was much more rarely than predicted, as they tended to be subjective "asides" on a particular situation and made no sense without the whole of the original text.

One obvious problem was a general misunderstanding of the term "figures", from the instruction not to include tables and figures. Some of the annotators took this to mean numerical figures, whereas the meaning in the guidelines was illustrations such as graphs. In future, instructions such as this should be explained more clearly.

6. Conclusions and future work

In this paper, we discussed issues involved in building a corpus for summarisation. We believe that our corpus is better than existing corpora for summarisation because it includes features which are not accounted for in currently available resources. In addition to the possibility of computing the performance of summarisation methods in terms of the information extracted, our corpus can be used to account for phenomena such as coherence. The former is obtained by comparing the information extracted by an automatic method against the information selected by a human annotator as important. The latter can be computed by identifying sentences in the automatic extract which were marked by the human annotators as being linked to one or several sentences. As well as this, the corpus can be exploited in little investigated fields such as the identification of sub-sentential units for removal in order to produce concise summaries.

So that we could assess the interannotator agreement, some texts were annotated by more than one annotator. The results of this comparison showed that there is little agreement between the sentences selected for extraction, but this is similar to the findings of other researchers (Lin and Hovy, 2002). When comparing the information content of the extracts, we noticed that the agreement was much higher, which proves that it may not be sentences which are the most important elements of the text to mark, but the actual information that one considers important.

The guidelines provided for the annotators also have an influence on the interannotator agreement. A discussion of the guidelines with the annotators revealed certain directions for future improvements. 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. In the future, we are considering employing professional summarisers, because they have more experience in dealing with unfamiliar topics.

Acknowledgements

This paper was written as part of an AHRB-funded project, "CAST – A Computer-Aided Summarisation Tool".

References

Burnard, L. 1995 *Users Reference Guide: British National Corpus Version 1.0*. Oxford: Oxford University Computing Services.

Cremmins, E.T. 1982 *The Art of Abstracting*. Philadelphia: ISI Press.

Edmundson, H.P. 1969 New methods in automatic abstracting. *Journal of the Association for Computing Machinery,* 16 (2): 264-285.

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)*, Berkeley, pp129-136.

Kupiec, J., Pederson, J. and Chen, F. 1995 A trainable document summarizer. In *Proceedings of the 18th ACM/SIGIR Annual Conference on Research and Development in Information Retrieval,* Seattle, pp 68-73.

Lin, C-Y. and Hovy, E. 2002 Manual and Automatic Evaluation of Summaries. In *Proceedings of the ACL 2002 Workshop on Text Summarization*, Pennsylvania, pp 45-51.

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.

Marcu, D. 1999 The automatic construction of large-scale corpora for summarization research. In *Proceedings of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99*), Berkeley, pp 137-144.

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.) *Mathematical Linguistics*. The Netherlands: North-Holland, pp 415-421.

Mitkov, R., Evans, R., Orasan, C., Barbu, C., Jones, L. and Sotirova, V. 2000 Coreference and anaphora: developing annotating tools, annotated resources and annotation strategies. In *Proceedings of the Discourse Anaphora and Anaphora Resolution Colloquium (DAARC'2000)*, Lancaster, pp 49-58.

Orasan, C. 2002 Building annotated resources for automatic text summarisation. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002)*, Las Palmas de Gran Canaria, pp 1780-1786.

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.

Siegel, S. and Castellan, N.J. 1988 *Nonparametric Statistics for the Behavioral Sciences*. 2nd Edition New York: McGraw-Hill.

Teufel, S. and Moens, M. 1997 Sentence extraction as a classification task. In *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scallable Text Summarization*, Madrid, pp 58-59.

Tsou, B.K., Lin, H-L. and Lan, T.B.Y. 1997 A Comparative Study of Human Efforts in Textual Summarization. In *Proceedings of the 3rd Pacific Association for Computational Linguistics Conference (PACLING '97)*, Tokyo.

Zechner, K. 1996 Fast Generation of Abstracts from General Domain Text Corpora by Extracting Relevant Sentences. In *Proceedings of COLING - 96, The International Conference on Computational Linguistics*, Copenhagen, pp 986-989.