A Study of Anaphoric Expressions in Human Produced Scientific Abstracts

Constantin Orasan

Computational Linguistics Group
School of Humanities, Languages and Social Sciences
University of Wolverhampton
Stafford Street, Wolverhampton, WV1 1SB
United Kingdom
C.Orasan@wlv.ac.uk

Abstract

One of the main reasons for having low quality automatic extracts is the presence of dangling anaphors. This paper analyses the referential expressions in a corpus of human written scientific summaries and tries to identify ways for improving the quality of automatic extracts. By recording the distance between the anaphoric expressions and their referents we noticed that humans do not use an aggregation-like process to avoid the dangling anaphors. The small number of anaphoric pronouns noticed in the summaries suggests that inclusion of a pronominal anaphora resolution module in a summarisation system is not necessary, but one which resolves referential noun phrases should be included given the large number of anaphoric noun phrases. These ideas were reiterated by investigating computer produced extracts.

1. Introduction

1.1. The problem of information overload in our society

With the current explosion of information available, people find increasingly difficult to keep up-to-date with the latest advances in a certain field, or to find a relevant piece of information. One way to deal with this problem is by using abstracts instead of full texts, reducing the time necessary to achieve a task. Unfortunately, only a relatively small number of texts have abstracts and sometimes their quality is quite low. This problem can be alleviated using computer programs which produce summaries automatically.

It is generally accepted that there are two main approaches to produce summaries. The first one is called extract and rearrange because it extracts the most important sentences from a text and tries to arrange them in a coherent way. The first method which produced such summaries was proposed by Luhn (1958) and similar methods are still widely used. The second approach tries to understand the text and, then, generates an abstract, for this reason being referred as understand and generate. The best-known method that uses such an approach is described in (DeJong, 1982).

Given that the methods which “understand” a text are too domain dependent, whenever robust methods are required, extraction methods should be preferred. Even though, the extraction methods currently used are more advanced than the one proposed by Luhn (1958), some of them taking into account even the discourse structure, many of these methods still produce summaries which are not very coherent. This happens because the extraction methods do not take into consideration the links between different units in texts, the resulting extracts consisting of disjoint sentences.

One of the main reasons of incoherent abstracts is the presence of dangling anaphors, which are anaphoric expressions deprived of their antecedent. Dangling anaphors can appear in human produced texts (both in written and spoken texts) due to negligence or lack of concentration, when the person who produces the text forgets to explain an entity or considers it known, or they can be found in computer produced texts, such as automatic extracts. In automatic extracts, this situation appears when a sentence containing an anaphoric reference is extracted, without extracting the one that contains the referred entity. Given that automatic anaphora resolvers are not accurate enough to replace these expressions with their antecedents, at present, no action is taken to improve the quality of the extracts.

1.2. The research questions

In this paper, we study the use anaphoric expressions in human produced scientific summaries in order to identify ways to improve the quality of automatic abstracts. We also try to identify sentence in the documents similar to the ones in the abstracts, and learn if there is a difference between the ways the contained anaphoric expressions are used. From this comparison, we try to learn if human abstractors replace the anaphoric expressions with their antecedents in abstracts.

In order to improve the cohesion, and as a direct result the coherence, of automatic abstracts, Paice (1981) proposed the aggregation process as a way of avoiding dangling anaphors. In this process, whenever a sentence containing anaphoric expressions is extracted, one or more sentences preceding the extracted sentence are included in the summary, so that it is likely that the referred entity is included. While investigating the human produced summaries, we try to find if a similar process can be identified.

Johnson et. al (1993) attempt to avoid dangling anaphors in computer produced abstracts by extracting only the important sentence which do not contain anaphoric expressions or rhetorical connectives. The presence of such expressions was recorded during the analysis to decide if humans use a similar process.

The main goal of this paper is to find ways to improve the cohesion of automatic extracts, but its findings can also be used in order to teach students how to write abstracts.
1.3. Previous use of abstracts for natural language processing

The abstracts have been used before to improve the quality of the automatic abstracts in (Orasan, 2001). There, the choice of lexis and grammatical constructs was recorded in order to identify special characteristics of the abstracts. In addition to this, a small study of the abstracts’ structure was pursued highlighting the requirements for automatic summaries.

In (Liddy et. al, 1987) and (Liddy, 1990), the anaphoric expressions from human produced abstracts are investigated, and rules to resolve them are proposed. Their working hypothesis is that the accuracy of an abstract’s retrieval system could improve if the anaphors are resolved in abstracts. Even though the reported accuracy of their rules is very high, the precision of the retrieval process does not improve conclusively.

Oakes and Paice (1998) used abstracts from journal articles in the field of crop protection to extract a set of templates which captured recurring patterns and expressions. Afterwards, these templates were used to produce automatic abstracts. Evaluation on a small sample set suggested that the method is effective.

The structure of this paper is as follows: section 2, the corpus of abstracts used for this research is described. Section 3 presents the parameters recorded for each anaphoric expression in the abstracts and the kind of insights that are likely to gain from these parameters. Sections 4 and 5 present the results of the analysis on human produced abstracts and automatic extracts. The obtained results are discussed in section 6 and conclusions are drawn.

2. The corpus

In this paper, the corpus described in (Orasan, 2000) was used. It consists of 917 abstracts totalling 146,489 words. For research in corpus linguistics this may seem as very small corpus, but building a corpus of abstracts which has a large number of words is a tremendous task, given the small size of one abstract. Two sources were used for building the corpus. The first one was the Journal of Artificial Intelligence Research (thereafter JAIR), from which 141 abstracts, with 24,509 words, were extracted. As the name suggests, this journal publishes articles in the field of artificial intelligence. Due to the fact that the size of this corpus was too small and the author wanted to compare abstracts from different areas, the INSPEC database was used as a second source of abstracts. The INSPEC database contains abstracts of papers from more than 4,200 journals and 1,000 conferences. Table 1 presents some details about each topic.

<table>
<thead>
<tr>
<th>Topic</th>
<th>No. of words</th>
<th>No. of files</th>
<th>In proceedings</th>
<th>In journals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Artificial intelligence</td>
<td>82,141</td>
<td>512</td>
<td>230</td>
<td>282</td>
</tr>
<tr>
<td>Computer science</td>
<td>21,467</td>
<td>137</td>
<td>117</td>
<td>20</td>
</tr>
<tr>
<td>Biology</td>
<td>16,081</td>
<td>100</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>Linguistics</td>
<td>6413</td>
<td>50</td>
<td>26</td>
<td>24</td>
</tr>
<tr>
<td>Chemistry</td>
<td>12,096</td>
<td>68</td>
<td>43</td>
<td>25</td>
</tr>
<tr>
<td>Anthropology</td>
<td>7737</td>
<td>50</td>
<td>24</td>
<td>26</td>
</tr>
<tr>
<td>Total</td>
<td>146,489</td>
<td>917</td>
<td>490</td>
<td>427</td>
</tr>
</tbody>
</table>

Table 1: The characteristics of the corpus

Given the amount of human input required by our research, at present only a part of the corpus was investigated. In the future, we intend to extend our analysis in order to obtain more reliable results.

As mentioned in the introduction, we also wanted to compare the use of anaphoric expression in the abstracts with the anaphoric expressions in full texts. In order to perform this comparison, all the articles from the Journal of Artificial Intelligence Research have been downloaded and converted to plain text. It was not possible to find an electronic version for the articles corresponding to the abstracts extracted from INSPEC, therefore the comparison was performed only on the texts from JAIR. Even though the number of texts used for this comparison seems small, the amount of work necessary to identify pairs of similar sentences from the abstract and full text makes almost impossible a large-scale comparison.

We should emphasise that no conditions were imposed on abstracts’ place of publication or the author’s(s’) mother tongue. Therefore, not all of them are written in perfect English. Moreover, as later shown, some abstracts have unexpected mistakes for a human produced summary.

3. Parameters investigated in the abstracts

Mitkov (2002) defines anaphora as the linguistic phenomenon of pointing back to a previously mentioned item in the text. The word which is referred to is called antecedent, and if both the anaphoric expression and the antecedent have the same referent in the real world they are called coreferential. Several entities that are coreferential to each other form a coreferential chain.

The most common types of anaphoric expressions are pronouns and noun phrases, but it is also possible to find verb anaphora, adverb anaphora and zero anaphora. Given that we try to find ways to improve computer produced abstracts, and the current technology is not able to deal reliably with other types of anaphora than pronouns and noun phrases, we recorded only their use.

During our investigation, we recorded the following parameters for each referential expressions:

- The type of reference: anaphoric, exophoric or inferable
- The kind of relation between the coreferential elements: pronoun, synonymy, repetition
- For pronominal anaphora, the distance in sentences between the pronoun and its antecedent
- The type of referent: noun phrase, clause, discourse segment

As can be noticed these parameters are very similar with the ones used by (Biber et. al., 1998) to investigate the characteristics of referring expressions in four different genres: conversation, public speeches, written news
reportage and academic prose. We used similar parameters because we wanted to see if we could identify similarities between the academic prose and summaries of scientific papers. Even though, they constitute a slightly different genre, it was expected that similarities would be identified.

The type of the reference was recorded in order to find out if certain expressions are mainly use in one way. The information about the relations between coreferential entities could be used to design algorithms which target only certain types of relations, whereas the distance between the pronominal anaphoric expression and the antecedent was recorded in order to test the use of “aggregation-like” rules in human produced summaries. A positive evidence of this fact would be a large number of sentences containing the antecedents for pronouns in sentences preceding the one with the anaphoric expression.

For most of the anaphoric expressions the antecedent is a noun, but in some cases it can be a clause, a full sentence or even a bigger discourse segment. During our analysis, we recorded those anaphoric expressions for which the antecedent is not a noun phrase.

In addition to these parameters we counted the number of non-anaphoric noun phrases in summaries and how many of these are definite noun phrases. This information could prove useful to decide if it is practical to include in a summarization system a module which decides if a definite noun phrase is referential.

4. Anaphoric expressions in human produced abstracts

In this section we investigate the use of referential expressions in human produced abstracts. For each referential expression, the parameters described in Section 3 are recorded.

It should be pointed out that the described analysis was not always straightforward because of the technical jargon used in the abstracts and small context available. There are cases where without domain knowledge it is difficult to know exactly which is the referred entity. We also noticed that not all the abstracts are carefully written, some of them containing errors and unconnected sentences, making the task more difficult.

4.1. Pronominal expressions

It is well known that the number of pronouns used in scientific texts is much lower than the one in other genres, and Orașan (2001) shows that the use of pronouns in scientific abstracts is also very low. Actually, the average number of anaphoric pronouns found in the part of the corpus which was annotated is around 1 per abstract. However, it was also noticed that number of anaphoric pronouns used in abstracts is unevenly distributed, so there are abstracts with a large number of anaphoric pronouns whereas other do not have any. This suggests that the use of anaphoric pronouns is more a matter of style and not necessary a constraint of the genre.

Statistics of the distance between the pronouns and their antecedents show that usually the antecedents are in the same sentence with the pronoun, in most of the cases as subject of the main clause.

During the analysis of our corpus, we noticed a large number of exophoric pronouns such as we and our, but as expected no you pronouns.

A surprising result was very small number of pleonastic pronouns found in the investigated summaries. In comparison with the number of anaphoric it, the number of pleonastic pronouns is almost negligible. Also, very small numbers of reflexive pronouns, one pronouns and surface count anaphors were identified in the abstracts.

A pattern which seem to repeat quite frequent is the anaphoric pronoun this followed by a verb other than to be, referring to a discourse segment. More data has to be analysed in order to produce a reliable rule that decided when the pronoun refers to a segment.

4.2. Noun phrases

The other category of referential expressions investigated in this paper is referential noun phrases. As in the case of pronouns, for each referential noun phrase we recorded the type of reference and the type of relation between it and the referred expression. Given that referential noun phrases are used more frequently in scientific texts than the pronouns, they are more important for the cohesion of abstracts.

While analysing pronouns we noticed that quite often they were not part of a coreferential chain, but just being part of an antecedent, pronoun pair. With noun phrases we noticed the opposite. They tend to form coreferential chains which run throughout the document. It was also noticed that the widest spread chains contains abbreviations referring to systems’ names or other terms common to the field of the abstract. The large number of abbreviations identified in the abstracts and their distribution throughout the abstract suggest that any summarization method should include a module which identifies these abbreviations. In addition to the fact that these abbreviations can be easily identified and they constitute important terms of the paper, in most of the cases they are not ambiguous, the identity being enough for saying that they are coreferential. However, the abbreviations which have a modifier should be treated with caution because there are cases when they refer to different entities as in first order HMM and second order HMM, found in one of the abstracts.

Another very frequent type of relation between coreferential noun phrases is identity of the head, and in many cases even repetition of the full noun phrase. As in the case of abbreviations, modified noun phrases have to be treated carefully because sometimes they can refer to different entities. During our investigation, we noticed long coreferential chains referring to the proposed method or system (e.g. our method, the method, the system, our system).

Less frequent than identity relations between noun phrases, we noticed cases where inference has to be applied in order to identify the referred entity (e.g. the computational model corefers with a mathematical model). In certain cases field knowledge is necessary to decide if two entities are related, as in the case of a stochastic finite-state model which is referred by the model, and when it is implemented, by the system. Because of lack of field knowledge, there were cases
when we could not decide if two entities corefer only on the basis of the summary.

A noun phrase which was constantly exophoric is \textit{this paper}. In some cases it is referred by noun phrases such as \textit{the paper}, but in all the investigated cases it referred to the paper from which the summary was produced, and not a different one.

Orasan (2001) showed that the number of nouns in abstracts it much higher than in other genres. Given this finding, we wanted to find out how many of the noun phrases appearing in summaries are not referential, and moreover, how many of these referential noun phrases are definite, and therefore are potentially referential. We found that on average, each summary contains 23.3 non-referential noun phrases, of which more than a quarter, around 6.8, are definite. The large number of definite noun phrases which were identified in summaries suggests that it is necessary to include in any summarization system a module which identifies non-referential definite noun phrases.

4.3. The use of anaphoric expressions in scientific texts

In this section we present a small investigation of the ways the referred entities in the summaries appear in the full papers. In order to do this, for each summary sentence containing an anaphoric expression we tried to find one or more sentences which cover the same message in the full text. Given the amount of work necessary to identify these equivalent sentences, at present the investigation was carried out only on 10 files from the Journal of Artificial Intelligence Research. In the future we intend to investigate more files in order to be able to draw more reliable conclusions.

Examination of the way pronouns are used in scientific texts revealed similar results with the ones obtained by Biber et al. (1998). The number of anaphoric pronouns used in texts is very small in comparison with the number of referential noun phrases. We also noticed that the majority of anaphoric pronouns have the antecedent in the same sentence, making superfluous to have a pronominal anaphora resolver in a summarisation system.

The process of pairing sentences from the full document with sentences from the summary proved more difficult that expected. For a large number of sentences, we could not find any sentence or small set of sentences which contain the same information. However, there were few summary sentences containing referential expressions for which we could find equivalents. In most of the cases where equivalents were identified, a sentence from the summary had several corresponding sentences in the document. We also noticed sentences from the summary which appeared almost identical in the conclusions of the full paper.

The hypothesis of the research presented in this section was that we would be able to identify sentences containing anaphoric expressions in the document which are used in the summary with the anaphoric expressions replaced by their antecedent. Instead of confirming our hypothesis we noticed exactly the opposite phenomenon with the anaphoric pronouns: wherever it is possible noun phrases are replaced by pronouns as in the following example:

\textbf{Abstract:} \textit{...this paper describes an interactive note-taking system for pen-based computers with two distinctive features. First, it actively predicts what the user is going to write. Second, it automatically constructs a custom, button-box user interface on request.}

\textbf{Document:} \textit{Our software has two distinctive features: first, it actively predicts what the user is going to write and provides a default that the user may select; second, the software automatically constructs a graphical interface at the user's request.}

An explanation of this phenomenon could be that pronouns are used instead of noun phrases because the abstracts’ authors try to keep their length as short as possible. In addition to this, the distances between the coreferential expression in abstracts are small, so pronouns can be used without any danger of introducing ambiguity.

5. Anaphoric expressions in automatic abstracts

In this section, we investigate automatic abstracts produced using a TF/IDF-based summarisation method (Zechner, 1996). One might argue that such a method is far too simple to be able to produce high quality extract. This is true, but many summarisation systems use TF/IDF in conjunction with other methods to extract sentences, so our findings could be used for a wide range of summarisation systems.

Due to time restrictions, the results reported in this section are rather preliminary, more analysis being necessary to draw more reliable conclusions. Summaries of 10 papers from the Journal of Artificial Intelligence Research were produced and the referential expressions from them were analysed. Given that the texts were quite long, we decided to produce 5\% summaries.

Investigation of the pronouns occurring in the automatic summaries revealed some interesting features. Out of 28 \textit{it} pronouns present in the automatic abstracts only one of them did not have its antecedent in the abstract. For the rest, the antecedent was either in the same sentence, 24 cases, or in the previous one in 3 cases. The other anaphoric pronouns which appeared in the extracts were \textit{their} and \textit{his} with their antecedents in the same sentence as the pronoun. Given that only some of the sentences were extracted, in several anaphoric pronouns became from cataphoric, but without losing their meaning, as shown in the following example:

\textit{Although it is not completely satisfactory in all practical cases (Baader, 1990b; Nebel, 1991), the descriptive semantics has been considered to be the most appropriate one for general cyclic statements in powerful concept languages.}

In comparison with the number of pleonastic pronouns identified in the human produced abstracts, it is much higher in automatic extracts, on average almost one in each extract. However, this difference can be also due to the fact that the extracts are longer than the summaries.

For the same reason the statistics about the noun phrases in extracts have to be treated with caution. It was noticed that almost a third of the non-referential noun phrases were definite which reiterates the necessity to identify the non-referential definite noun phrases. If the
number of dangling pronouns found in the extracts was negligible, about 20% of the definite noun phrases were dangling. As noticed in the case of pronouns, because only some of the sentences are included in the extract, some noun phrases which had the referent before then, now have their referent after as in *The MDL principle has been used… followed by According to the minimum description length (MDL) principle…*. Therefore, a program which produces extracts should change the order of the sentences so that the sentence which contains the antecedent appears first in the extract.

6. Discussion and conclusions

As shown in (Salanger-Meyer, 1990) and (Orasan, 2001) the summaries produced by humans follow more or less the introduction-problem-solution-conclusion structure. Each of these moves contains very few sentences and constitutes a discourse segment on its own. Given that pronominal anaphoric expressions rarely refer to entities form another discourse segment, it is normal that their number is quite small.

The small number of pronouns present in human produced summaries, and the fact that we did not find strong indications that human abstractors extract sentences from the full document and replace anaphoric pronouns with the referred entities suggest that inclusion of a pronominal anaphora resolver in a summarisation system is not absolutely necessary. Moreover, given the small number of pronouns present in academic papers it is likely that an approach which ignores all the sentences containing anaphoric pronouns, like the one proposed by (Johnson et. al., 1993), would not eliminate very important sentences.

The analysis of the noun phrases showed that their number is much larger and they have to be taken into consideration when a summary is extracted. Statistics from human produced summaries and computer extracts showed that a large number of definite noun phrases are not referential, and therefore a module which determine if a definite noun phrase is referential should be included in any summarisation system.

If the dangling pronoun did not seem to be a problem in automatic extracts, the dangling noun phrases are much more frequent in them, and therefore they should be tackled by summarisation systems.

The links between different segments constituting an abstract are realised by rhetorical connectives, such as indicating phrases (e.g. *we propose a method, we show*) which usually also indicate the type of the move and by repeated references to the same entity throughout the summary. In most of the cases repetition of noun phrases or abbreviations are used to refer to these entities.

The large number of exophoric pronouns (e.g. *we, our*) found in the abstracts can be explained by high frequency of expressions specific to the academic language such as *we show that, we present a method, our method, our system* etc. As a result of the same fact, *you* or *your* pronouns can be found in the abstracts.

During our analysis of the human produced abstracts we noticed a large number of exophoric noun phrases. Some of them are part of expressions which are commonly used in academic language, such as *this paper or our results*, whereas the others refer to concepts specific to the domain of the summary.

Comparison between the use of referential expressions in summaries and in documents did not suggest that the humans use a process similar with *aggregation*. On contrary, the humans seem to use more pronouns than in the source sentences in order to keep the length of the summary short.

In this paper we investigated a corpus of abstracts trying to find way to improve the results of automatic summarisation. Even though the size of the analysis was small because of time constraints, we identified several modules which should be included in any summarisation system which tries to improve the cohesion of produced abstracts.

7. References


