Patterns in scientific abstracts

Constantin Orasan
School of Humanities, Languages and Social Sciences
University of Wolverhampton
C.Orasan@wlv.ac.uk
http://www.wlv.ac.uk/~in6093

1. Introduction

In today's world, large amounts of information have to be dealt with, regardless of the field involved. Most of this information comes in written format. Computers seem the right choice for making life easier, by processing text automatically, but in many cases at least partial understanding (if not full understanding) is necessary in order to automate a process. Luhn (1958) proposed a method for producing abstracts which works regardless of the type of the document. Although further research carried out in this area is based on his work, it has become apparent that general methods are not a solution. Instead, more and more research has been done into restricted domains, where certain particularities are used for "understanding". A well known case is that of DeJong (1982) where the structure of newspaper articles was used in order to get the gist of the articles and then generate summaries.

However, all these methods, due to their inherent specificity, require prior knowledge about the characteristics of the genre to which they are applied. Whether they refer to the distribution of words throughout a document or to the overall structure of the document, such textual features have to be identified in a corpus, by applying methodologies from corpus linguistics. For example, (Biber, 1998) investigated four different genres: conversation, public speeches, news reports and academic prose, at lexical, grammatical and discourse level. The findings showed that each genre had its own characteristics which distinguish it from the other genres. One can use these findings for the automatic identification of a genre, or for improving the results of an automatic method for one of these genres.

In this paper, the characteristics of a very narrow genre, that of scientific abstracts, are explored on three different levels: lexical, syntactic and discourse. The hypothesis is that it is possible to find patterns, which could be used at a later stage to find similarities between abstracts. It is hoped that these patterns will be useful for improving the results of automatic summarisation methods when applied to scientific texts.

The patterns identified in this paper are not only useful for automatic abstracting or computational linguistics in general, but they can also be used in order to teach students how to write abstracts. As is explained in the next section, both reading and writing an abstract are not a trivial task, and many students experience difficulties. Those students who are learning English as a second language have even greater problems with such tasks. The patterns which are identified in this paper could help them to write abstracts.

2. What is an abstract and why is it useful?

The notion of an abstract is part of everyday language, but there is more than one definition accepted for it. According to (Cleveland, 1983, p. 104) "an abstract summarises the essential contents of a particular knowledge record and is a true surrogate of the document". A similar definition is given by (Graetz, 1985): "the abstract is a time saving device that can be used to find particular parts of the article without reading it; … knowing the structure in advance will help the reader to get into the article; … if comprehensive enough, it might replace the article" These two definitions emphasise the most important function of an abstract (i.e. its role as a replacement for an entire document). However, these definitions refer to ideal abstracts produced by professional summarisers. This paper argues that it is very unlikely that an abstract produced by the author(s) of a paper (as in the case for most of the abstracts in the corpus used for this paper) is to be used as a replacement for the whole document. Therefore, a simpler definition of an abstract is considered more appropriate in this context: "a concise representation of a document's contents to enable the reader to determine its relevance to a specific information" (Johnson, 1995). So, the abstract is no longer a "mirror" of the document, but instead draws attention to the most important information contained within the document. Moreover, the main

purpose of this definition (i.e. to highlight what is important in the document) can be applied to automatic abstracting.

Regardless of the differences between the definitions mentioned above, they all highlight the use of abstracts as filtering devices. In the present days, people are constantly being bombarded by large amounts of information. Scientists and academics use abstracts to filter the existing literature when conducting research into a certain topic or when trying to keep up-to-date with the latest advances in their field of interest.[1] On the basis of the abstract, they can decide if an entire document is worth reading or not.

Swales (1990) considers the process of writing an abstract to be a "rite de passage" for gaining entry into the scientific community via a demonstration of increasing mastery of the academic dialect. This is true, writing an abstract is not a trivial task given that it does not allow redundancies and forces the writers to use a lot of compound words. As Halliday (1993) points out, in scientific texts, which include scientific abstracts, lexical density is very high, which makes it difficult to both read and write such texts.

There are several ways of classifying abstracts. One way is to divide them, according to their usage. The categories used in this case are: indicative, informative and critical (Lancaster, 1991). An *indicative* abstract provides a brief description to help the reader understand the general nature and scope of the original document without going into a detailed step by step account of what the source text is about. An *informative* abstract is more substantial than an indicative abstract and is often used as an alternative to the original document, following the main ideas presented in the document. Usually, these two categories are combined, an abstract performing both an indicative and informative function. A *critical* abstract gives not only a description of the contents, but also a critical evaluation of the original document. Usually, in this case, it is not the author of the paper who writes the abstract.

After analysing the corpus of abstracts used in this paper, it became evident that the vast majority of the abstracts are indicative. Some of them also have an informative function, but cannot be classified as informative because they cannot replace the original document. No critical abstract was included in the corpus.

There are also other ways of classifying an abstract: by the way they are used, and by their author, but these classifications are not relevant to this paper. Most of the abstracts in the corpus present the disadvantages of abstracts which are written by the authors of the papers, instead of trained summarisers. As is shown is section 6, most of the abstracts do not always follow the Problem-Solution-Evaluation-Conclusion structure (Swales, 1990). One reason for this could be because the authors do not consider the abstract to be particularly important, in many cases it is written just before the paper is submitted. Moreover, in some cases the content of the abstract does not necessarily reflect the content of the paper, as (Cleveland, 1983 p.110) has remarked "authors as abstractors have been known to use their abstracts to promote the paper; this can create a misleading abstract and is unfair to the user".[2]

2. The corpus

In order to analyse the structure of scientific abstracts, a corpus of abstracts has been built. It consists of 917 abstracts with 146,489 words. For research in corpus linguistics this may seem a 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. Moreover, as Sinclair (2000) has pointed out, small corpora are not necessarily bad; in some cases a small corpus is the right choice. The research presented in this paper required a lot of human input, and therefore its size had to be kept down to make the analysis possible. However, whenever possible, automatic processing was used.

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

[1] Given the purpose of this paper, the use of news abstracts for reporters, politicians or businesspersons is not considered. However, it can be argued that, in such a context, summaries of a single document are not so useful, digests (multidocument summaries) being more appropriate

[2] An example is the following sentence from an abstract of an article from conference proceedings: "By mastering the fundamental issues discussed in this paper, you will increase the return of your organisation's investment in data warehouses"

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. Six topics have been selected and the first few abstracts have been included in the corpus. Table 1 presents some details about each topic.

Topic	No. of words	No. of files	In proceedings	In journals
Artificial intelligence	82,141	512	230	282
Computer science	21,467	137	117	20
Biology	16,081	100	50	50
Linguistics	6413	50	26	24
Chemistry	12,096	68	43	25
Anthropology	7717	50	24	26
Total	146,489	917	490	427

Table 1: The characteristics of the corpus

Several remarks have to be made regarding the corpus. In this research, articles from artificial intelligence are of particular interest; therefore, most of the abstracts are from this field and other related areas (machine learning, information retrieval, etc.). However, other areas, like anthropology, chemistry and biology have been included for making comparisons between different genres. When computer searches were carried out in order to obtain abstracts which could be analysed in this paper, those abstracts found were from both journal articles and conference proceedings. In one case, that of the biology abstracts, the first fifty abstracts returned by the search came from conference proceedings. As a result, a second search has been performed in this area, this time restricting the search to abstracts of documents published in journals. These two categories of abstracts allow the author to see if there is a difference between the abstracts of papers published in conference proceedings and those published in journal articles. In some cases, an abstract can belong to more that one category. For example, it has been noticed that some abstracts considered as belonging to the field of linguistics, are concerned with some computational aspect of linguistics, and therefore could also be considered as belonging to the field of artificial intelligence. For each topic, the number of abstracts published in conference proceedings and journals is shown in Table 1.

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. However, it was considered that this would better reflect the use of English in the research community.

3. Length

The length of abstracts was considered to be the first way of comparing them. Given that usually a conference paper is shorter than one published in a journal, the author expected to find that the abstracts of journal articles were longer than those of conference papers. Also, in the case of the former, the editors impose a strict control on the quality of the article, and subsequently on their abstracts. The statistics showed that abstracts of the journal articles are noticeably longer, both in terms of sentences and words, than the ones belonging to the conference papers (Table 2). The shortest abstracts are the ones belonging to humanist disciplines (linguistics and anthropology). However, due to the small number of abstracts taken from these disciplines it is not possible to conclude that all the abstracts from the humanities are short.

Topic	Journal		Proceedings		Total	
	Sent/Abs	Words/Abs	Sent/Abs	Words/Abs	Sent/Abs	Words/Abs
Artificial Intelligence	8.20	165.67	6.05	152.44	7.24	159.74
Computer Science	9.58	232	5.94	163.30	6.40	159.32
Biology	7.9	196.18	5.65	130.02	6.78	163.43
Linguistics	5.78	149.52	5.92	108.65	5.85	127.83
Chemistry	8.58	215.08	6.34	163.30	7.14	181.85
Anthropology	6.23	157.88	6.08	154.43	6.16	156.26
Total	7.39	174.56	5.96	147.94	6.61	160.31

Table 2: The length of abstracts in sentences and words

The length of an abstract ranges from 1 sentence to 21 sentences in the case of abstracts taken from journal articles, and 1 to 16 sentences in the case of abstracts in conference proceedings. In both cases it is possible to find abstracts which have only one sentence, but usually this sentence just enumerates the topic covered by the article. An interesting result was obtained when the length in words was computed. The longest abstract has been published in conference proceedings. As a result of this, it is

possible to conclude that overall, the abstracts published in journals are longer than the ones published in conference proceedings, both in terms of words and sentences. However, there is no rule which states this.

4. Lexical level

There are several ways of analysing a corpus. The most basic form is by displaying and analysing lists of characteristics. The analyses can involve very simple wordlists or be more sophisticated, including the classic concordance format (Kennedy, 1998). In this paper the author starts by analysing word frequency lists and lists of n-grams. These lists are also compared with the same lists generated from a general purpose corpus, BNC (Burnard, 1995). In the next section, grammatical features of the texts are considered by analysing subject-predicate pair lists. Whenever it was necessary, specially designed programs were used to display the context.

4.1 Word frequency lists

Given the small size of the corpus it was thought unwise to make significant generalisations, but even by analysing the lists interesting features can be noticed. For this analysis and the ones which follow, the corpus was tagged using the FDG tagger (Tapanainen, 1997). Using this tagged version of the corpus, word frequency and lemma frequency lists were produced. For each case, two different lists were generated, with and without considering the part-of-speech of each word. These lists were compared with similar lists produced from BNC. It should be pointed out that the results of the comparison have to be treated with caution. The first reason is the huge difference in the size of the two corpora. Therefore, the decision was made not to draw a comparison between the two corpora in terms of frequencies, but using their position in the list instead. Of course the frequencies could have been normalised, but given the small size of the texts in the corpus, it was thought that the results would still be unreliable.

The second problem with this comparison comes from the language. BNC is a corpus of British English, whereas the corpus of abstracts was not filtered on the basis of the variety of English used. Moreover, it can be argued that most of the English used in the scientific domain is written in American English. For example, the word *summarisation* and its derived forms (e.g. summarising, summarise, etc.) appears 137 times. In 114 of the cases the American English spelling (AE) was used, whereas the British English spelling (BE) was used in only 23 cases.[3] The same problem was found with the word *generalise* (out of 83 occurrences, 78 used AE and only 5 BE) and *characterise* (56 occurrences, 46 used AE and 10 BE). However, it can be argued that there is no word with a high frequency of occurrence, which has different spellings in AE and BE.

Word freq. list from abs	Word freq. list with tags from abs	Lemma freq. list from abs.	Lemma freq. list with tags from abs.	Word freq. list from BNC
7132 the	7132 the DET	8442 the	8442 the DET	5538939 the
5908 of	5908 of PREP	5913 of	5913 of PREP	3086807 of
4156 and	4156 and CC	4543 be	4179 be V	2631593 to
3082 a	3069 a DET	4162 and	4162 and CC	2574912 and
2925 to	2466 in PREP	3293 a	3270 a DET	2091285 a
2485 in	1953 is V	3010 to	2870 in PREP	1824289 in
1953 is	1747 to TO	2890 in	1823 to TO	1088658 that
1575 for	1569 for PREP	1644 for	1635 for PREP	983593 is
1310 The	1310 The DET	1398 this	1269 this DET	917103 was
1204 that	1178 to PREP	1277 that	1187 to PREP	897690 I
1093 are	1093 are V	1165 we	1165 we PRON	849027 for
910 on	894 on PREP	940 use	917 on PREP	847109 it
886 with	886 with PREP	933 on	909 with PREP	802227 's
790 by	789 by PREP	909 with	906 system N	712502 on
717 an	717 an DET	906 system	822 by PREP	661109 be
698 be	698 be V	823 by	781 an DET	657574 with
607 this	597 that CS	781 an	718 have V	631554 The
585 as	575 that PRON	735 have	694 it PRON	619043 as
528 system	563 this DET	694 it	615 that PRON	596588 you
521 which	528 system N	636 as	597 that CS	501209 at

Figure 1 Different frequency lists

Figure 1 shows the different frequency lists for the first 20 entries in the lists. It is evident that the first 6 entries of the word frequency list, from the corpus of abstracts and from BNC, are almost identical.

[3] However, it should be pointed out that in BNC both spellings of the word *summarise* can be found, although the British spelling is more frequent that the American one (1220 BE, 751 AE)

However, further down in the list, differences appear. The word *was*, which in BNC occupies the 9th position, in the corpus of abstracts appears in the 51st position. This can be explained by the small number verbs which appear in the past tense. In the frequency list of the corpus the most frequent noun is *system*, appearing in the 19th position, but if the lemmatised version of the list is taken into consideration instead, it is in the 15th position (even higher if the different types of systems are taken into consideration e.g. *eco-system*, *geo-system*). After this position, the nouns are quite frequent; *paper* occupies the 21st position (24th in lemmatised list), *data* 23rd (25th), *information* 27th (26th). In BNC, the first noun on the list is *time*, which occupies the 71st position, and there are not many nouns in the first 200 words (e.g. *people* 94th, *years* 120th, etc.). It is also evident that the types of nouns are completely different.

Also, a quick check on the lists reveals that many of the most frequent words from BNC, belong to more than one grammatical category. This is not true with the most frequent words in the corpus of abstracts, especially the ones which are nouns. This may indicate that the abstracts focus more on abstract states, objects and processes. A similar result was obtained by (Biber, 1998) studying nominalization in scientific texts.

4.2 N-gram lists

N-grams are groups of consecutive N words in the corpus. Punctuation marks were not considered as being part of an n-gram, therefore all of those containing punctuation marks were removed. When sorted by their frequency, n-grams uncover frequent patterns in a corpus. N-grams (with N from 2 to 9) have been generated. Initially the idea was to compare them with the ones produced from BNC, but it became apparent that there is not much of a link between the two, except for the very frequent patterns *of the*, *in the*, which are not very useful. However, this is not surprising, given that the n-grams are an indicator of a document's contents. Figure 2 shows the first 29 entries of 2-grams, 3-grams and 4-grams lists. The lists with a higher number of words had much lower frequencies and for space reasons they are not displayed here.

2-grams		3-grams		4-grams	
1276	of the	143	in this paper	41	in this paper we
640	in the	115	be use to	26	can be use to
360	this paper	72	the use of	20	this paper present a
320	on the	61	base on the	20	in the context of
319	of a	58	be base on	17	the world wide web
311	and the	53	a set of	17	it be show that
306	to the	50	show that the	17	be one of the
273	have be	48	we show that	16	the size of the
258	in this	47	the problem of	16	a wide range of
256	for the	47	the development of	15	one of the much
250	can be	46	the number of	15	be base on a
242	base on	44	this paper present	14	this paper we present
215	in a	43	one of the	14	on the other hand
204	it be	43	be apply to	14	in the form of
201	be a	42	we present a	14	be base on the
198	be use	42	this paper we	13	this paper describe the
196	of this	41	a number of	13	of this paper be
192	with the	39	this paper describe	13	in the field of
174	to be	39	can be use	12	the performance of the
166	that the	37	a variety of	12	on the basis of
163	show that	35	in term of	12	in the size of
144	be the	35	be able to	11	with respect to the
143	use to	34	of the system	11	this paper describe a
142	the system	33	the performance of	11	this paper be to
141	number of	33	base on a	11	the use of a
139	by the	32	we propose a	11	can be apply to
138	as a	31	with respect to	10	the development of a
123	artificial intelligence	31	the result of	10	of a set of
121	such as	28	some of the	10	in the presence of

Figure 2: The lists of 2,3,4-grams from the lemmatised version of the corpus

It seems that the lists are not seriously influenced by the type of abstract. A comparison of the lists of n-grams produced from abstracts published in journals and the ones from proceedings did not reveal many differences. When the 3-grams are considered, in both cases the first element on the list is *in this paper*, followed by *be use to*. However, the third element from the first list, *we show that*, appears in the 67th position in the second list. In the list of 2-grams, *show that*, appears in the 14th position in the

first category and only 48th on the second category. Given that the size of the two subcorpora is almost the same, such a result is unexpected. However, it could be explained by the fact that in many cases, the conference papers present work in progress and, therefore the conclusion is not necessarily the strongest point of a paper and therefore no reference is made to the conclusion in the abstract. As is argued in section 6, there are cases when the abstracts do not have an evaluation section, but this happens less frequently with the abstracts belonging to journal papers.

The n-gram lists also uncover terms for specific domains (e.g. *information retrieval, neural networks, world wide web,* etc.). Although this is a possible use of them, this paper does not intend to investigate this aspect.

4.3 The case of the noun *paper*

The analysis of word lists and n-gram lists represent a very easy and powerful way to find patterns in texts. For example, if the word *paper* is taken, it is appears 499 times, in 473 abstracts, which means that more than half of the abstracts use it. In one abstract, it is used four times, its authors introducing each move using the following constructions: *this paper investigates*, *this paper introduces*, *this paper describes*, *this paper ends*. There is another abstract in which the word *paper* is used 3 times. In 24 abstracts it is used twice, although in the rest of the abstracts it appears only once. In addition to this the word *study* is used as a noun 170 times, *research* 154 times and *work* 111 times. Even though the nouns *study*, *research* and *work* are not always synonymous with the word *paper*, these three words together with *paper* strongly indicate that most abstracts make a reference to the paper from which they are derived using constructions like: *in this paper, in this study* etc.

The n-gram lists strengthen this aforementioned conclusion. The word *paper* usually appears in constructions like *this paper* (360 times) or *the paper* (115). Constructions such as *this study* (18), *this research* (14), *this work* (25), *this article* (27) are also found. In many cases, the word *paper* is used as the subject of verbs like: *present* (62 times), *describe* (50), *be* (45), *introduce* (15). Clearer patterns appear when more words are considered: *in this paper* (143 times), *this paper presents* (44), *this paper we* (42), *this paper describes* (39). By increasing the number of contextual words, the patterns become less frequent: *in this paper we* appears 41 times, *this paper presents a* (20 times), *this paper we present* (14 times), *this paper describes the* (13 times). Even when 5 words are considered, the first element on the list is *in this paper we present* (14 times).

As a result of this analysis, it is reasonable to conclude that the patterns found with regards to the word *paper*, are not accidental and cannot merely be explained by a high number of occurrences of the word *paper*. Instead, they represent patterns specific for abstracts. If the most frequent noun (i.e. *system*) is considered, such patterns do not appear.

5. Grammatical level

Analyses of the word frequency lists and n-gram lists proved to be a very useful way of discovering patterns. However, they can only reveal patterns between words which are adjacent. As a result of this, it is possible that many patterns were missed due to some modifiers or adverbs. In this section, grammatical structure is used for uncovering patterns in the abstracts. All the abstracts were tagged using the FDG tagger. This tagger, in addition to assigning part-of-speech tags to each word, also provides partial dependency relations between words. These dependency relations were then used for finding common noun-verb pairs. Consequently, two lists were generated. The first list represents subject-predicate pairs and the second one contains pairs of nouns, which are not the subject of the sentence (e.g. objects), and verbs. It should be pointed out that the process of generating these lists was completely automatic; therefore they contain some errors. However, the number of these errors is relatively low, and they do not influence the validity of the results. Figure 3 presents the first 20 entries from the two lists.

120	it be	65	be system
106	we present	53	be problem
88	we show	38	be information
86	there be	37	present paper
84	that be	37	be data
66	we propose	36	present system
63	paper present	36	be knowledge
58	which be	33	present approach
56	we describe	29	be it
50	paper describe	28	be model
35	we discuss	27	be agent
31	result show	25	be research
30	it show	25	be first
28	we introduce	24	play role
26	they be	24	be number
25	approach be	24	be method
24	we use	23	be science
24	we develop	23	be process
24	this be	23	solve problem

Figure 3: Pairs of subject-predicate and verb-noun

As expected, the patterns found using n-grams also appear in these lists, but with an increased frequency. This is normal given the fact that intervening adverbs do not affect the patterns. For example, the pair *we-present* is found 106 times, an increase of 5 from the list of 2-grams. This is because of groups like *we also present*. The increase is greater in the case of *we-show*, from 71 to 88.

In the list of subject-predicate pairs some interesting pairs can be noticed. The first pair in the list is *it-be*. After manually checking all the appearances in the corpus, it was noticed that in only 8 cases *it* was used as an anaphoric pronoun. This finding is not surprising given that it has been shown by previous research that the pronoun *it* is frequently used in the scientific domain as non-anaphoric. In addition to the 112 pairs of *it-be*, where *it* is used non-anaphorically, there are 86 appearances of existential *there* as subject for the verb *be*. All these cases suggest that existential sentences are frequently used in abstracts.

The subject *we* appears with a closed set of verbs (e.g. *present, show, describe, discuss, etc.*). This set includes verbal processes, in Halliday's terms (Halliday, 1994), and *presentational* processes; they can be used to determine the different types of moves (as is shown in the next section). The subject is usually the one which realises an action. Eight out of the first 20 pairs contain the subject *we* in them, which suggests that the author is present in the abstract as the one who *presents, shows*, etc.

The large number of *be* predicates in the second list (14 times in the first 20 most frequent pairs) reiterates the fact that existential sentences are quite frequent. The pair *present-paper* appears usually because passive voice is used (e.g. *…is presented in this paper)*, and therefore it is an instance of the subject-predicate pair *paper-present*. The other pairs suggest that *systems* and *approaches* are presented and *problems* are solved.

6. The structure of abstracts

The most distinctive feature of abstracts is their rhetorical structure. Gopnik (1972) has identified three basic types of scientific paper: the 'controlled experiment', the 'hypothesis testing' and the 'technique description'. Each type has its own structure, but according to Hutchins (1977) they can be reduced, either by degradation or by amelioration, to a problem-solution structure. However, this structure is too general for the purposes of this paper. A more detailed organisation can be identified in the scientific papers: background information about the domain, the problem, the solution to the problem, evaluation of the solution and conclusion. Sometimes, they are referred to as *moves*. Graetz (1985) and Swales (1990) claim that an abstract should have the following structure: problem-solution-results-conclusion. However, Salanger-Meyer (1990a) analysed a corpus of 77 abstracts from the medical domain for this structure and found that only 52% of the abstracts followed the structure.

In each abstract, the moves have been manually identified. Given the large number of abstracts in the corpus and the difficulty of annotation, only 67 abstracts have been selected and annotated. Therefore the results presented in this section are just preliminary, in the future a semiautomatic procedure will be used. Out of the 67 annotated abstracts, 35 were published in journals and 32 in conference proceedings. The patterns found in each move are similar to the ones reported in (Salanger-Meyer, 1990b) for abstracts from the medical domain.

During the annotation, five moves have been considered: *Introduction, Problem, Solution, Evaluation, Conclusion*. Ideally, an abstract should contain all five moves, but an abstract with only *Problem, Solution,* and *Evaluation* moves, has been considered to be perfectly acceptable and correct. It should be pointed out that the annotation process is difficult and in many cases highly subjective. At present, the corpus has been annotated by only one person, the author, but in the future at least some of the abstracts have to be annotated by another person in order to compute the reliability of the annotation using interannotator agreement measures. Out of 67 abstracts, it was found that only 39 of them (58%) could be considered to follow the expected structure. In the rest of the cases, either an important move was missing or the moves were in logical order. In quite a few cases, it was noticed that the evaluation was given before the method was presented. The abstracts of journal papers proved to be slightly better than the ones of conference papers in terms of organisation (21 abstracts from journal papers and 18 abstracts from conference papers), but more data have to be investigated.

6.1 The *Introduction* section

The introduction section is meant to provide the reader with some background information, to explain what has been done in the field, etc. Given the constraints on the size of an abstract, this move is not compulsory. Moreover, abstracts are written for relatively informed readers and therefore they should not provide too many background details. However, by analysing the abstracts in the corpus, the author has noticed that there are cases when the introduction is quite long, in some cases almost half of the abstract. Usually this section makes references to previous work using expressions like *existing approaches, prior work, previous work*. The sentences in this section contain general truths (e.g. *Storing and accessing texts … has a number of advantages over traditional document retrieval methods* or *Discourse analysis plays an important role in natural language understanding*) usually expressed through the present simple tense. In addition to the present simple tense, the present perfect is used quite often for stating generic truths, but is usually used for emphasising the weaknesses of the previous work (e.g. *It is usually expected … but …* or *The standard approach … has been to … Such an approach becomes problematic).*

The introduction also gives hints about the problem which is going to be solved, highlighting in an appropriate way the weaknesses of previous approaches. In some cases it was quite difficult to make a clear distinction between the *Introduction* and the *Problem*.

6.2. The *Problem* section

The introduction section, prepares the reader for the problem section. In this section, the problem with which the article deals, is expressed. There are cases when the problem is not clearly stated, but the reader can usually infer it from the introduction and the solution sections. Even though the reader can guess the problem, it is not desirable to have an abstract which does not state the problem explicitly.

Given the fact that there are often some very frequent patterns, this section can be identified relatively easily. Usually it is explicitly signalled through phrases like: *we describe, we present, a formalism is presented, we outline,* etc. The preferred tense for stating the problem is the present simple. In some cases comparison with previous work is used for stating the problem: *because of …existing methods can erase …we propose ….* In these cases the *Problem* section also serves also as introduction.

6.3. The *Method* section

In this section of the abstract, the author(s) should explain how the problem is resolved. This section is very important for the reader because it enables him/her to understand the kind of approach to solve the problem that was used. Some sentences from this section are marked overtly using phrases like: *an alternative solution, the approach described here uses ….* Some of the patterns used for stating the problem, also appear in the method section (e.g. *This paper reports on a Japanese information extraction system that merges information using a pattern matcher and discourse processor*) In this example, as well as in the other cases where a pattern which would be usually found in the *Problem* section appears, the phrase has a double role. On one hand, it reiterates the problem, or state the problem if it has not been stated already, and on the other hand, it explains the method.

The verb tense seems to be the present tense, although (Biber, 1998) found that the past test is more frequently used in this move. Normally, this move should describe each step taken in the research for solving the problem, and therefore the past simple should be preferred. However, in the abstracts of the corpus used for this paper, the method is described in general terms, and not as a sequence of steps.

6.4 The *Evaluation* section

Another important section is the one which summarises the evaluation. An abstract without an evaluation section is not considered to be correct. This is because, the evaluation proves the validity of the method proposed. If a researcher is trying to find a solutions for a problem similar to the one discussed in the abstract, he or she needs this section in order to judge the usefulness of the solution. If the evaluation suggests that the solution is appropriate for the given problem, one can read the whole article in order to obtain a full explanation of the method and a detailed evaluation.

It has been noticed that whenever the verb *show* is used, it appears in the evaluation move in phrases like: *we show, it is shown*. In addition to the verb *show*, other phrases are also used for explicitly marking this move (e.g. *this work provides, reveal, yield, investigate, find* etc.) Besides these phrases, words which either refer to measures or ways of measuring are used (e.g. *limited, compare, quantify, are tuned, shortcomings, this allows us to measure* etc.)

There are cases when the evaluation is not presented. Instead, the authors indicate that an evaluation was performed (e.g. *we briefly describe some experiments, we present three case studies,* etc.) On the basis of such an abstract, one cannot decide if the article is relevant or not, and therefore it is not really useful.

The preferred verb tense is the present simple tense, but a large number of verbs in conditional tense have been noticed. This could indicate that the authors do not want to overestimate their results (e.g. *can be adapted, can use,* etc.) However, in addition to being part of the evaluation, these sentences can be also used as a conclusion. Connectors, which hardly appear in the other moves, are quite frequent in this one (e.g. *however, in addition,* etc.), being used for justifying the evaluation.

6.5 The *Conclusion* section

Many research papers have a conclusion section in which the results of the method are placed in a broader context. However, it is not absolutely necessary to have this section separate in the abstracts, but its presence makes an abstract more valuable in a broader context. Therefore, an abstract that does not contain such a section is not considered to be incorrectly structured. In many cases the evaluation section also plays the role of conclusion.

In the majority of cases, this move contains an explicit reference to the abstract (e.g. *this work provides, these observations suggest,* etc.) In addition, phrases like *this paper concludes, as a conclusion,* or adverbs (e.g. *therefore, as a result,* etc.) are used to mark the beginning of the conclusion move.

7. Conclusion and future work

In this paper it has been shown that regardless of the level of analysis, lexical, syntactic or discourse, it is possible to extract patterns from scientific abstracts. The simplest way of analysing, the lexical level, uncovered groups of words which usually go together. As an example the word *paper* has been analysed, noticing that it appears in frequent patterns. By using dependency relations between subject and predicate, pairs were generated. Many of these pairs were also found in the lists of 2-grams, but given that in this case the intervening adverbs do not matter anymore, the frequency of the pairs is higher, so giving more reliable figures.

By analysing the discoursal structure of abstracts, it became evident that the scientific abstracts, written by the authors of the papers, do not necessarily follow the structure which the literature predicts. In each abstract, the moves were manually annotated and those groups of words which signal the type of a move were identified. Some of the words seem to indicate reliably a certain type of move (e.g. the verb *show* appears usually in the evaluation section).

The question which arises at this point is how these patterns can be used in computational linguistics and especially for automatic summarisation. For the beginning it has been noticed that the word *paper* appears usually only once in the abstract, in constructions like *in this paper we*. These constructions are very similar to the findings reported in (Paice, 1981), where common patterns from full-length papers, called *indicating phrases*, were identified in scientific papers and used for producing a summary. Therefore, a sentence from a document which contains a pattern similar with one previously identified, is more likely to be important, and consequently worth including in the abstract. However the usefulness of each pattern has to be assessed in a corpus of scientific papers.

The patterns, which have been identified, are not only frequent in the abstracts, but in many cases, they also indicate a certain move. This suggests that it could be possible to design an automatic procedure

for identifying each move in the text. However, it has been shown that more than one pattern is used to introduce a move, therefore for each move it is possible to find more that one way of introducing it. This suggests that it is possible to find general templates for each move. Such an approach would not be new, (Paice and Jones, 1993) proposing a similar method for building abstracts. However, in their case the templates are very specific for a certain domain.

Patterns in abstracts are useful not only for automatic abstracting. They are also useful for helping researchers to produce their own abstracts. Narita (2000) proposed a system for Japanese, which can be help writing abstracts by displaying sentences and collocations from a corpus of annotated abstracts.

8. Acknowledgements

The author of this paper would like to thank Ramesh Krishnamurthy, Dr. Andrew Caink and Richard Evans for their comments provided at different stages of this paper, and IEE and Journal of Artificial Intelligence Research for the permission to use their abstracts for this research

9. Bibliography

Biber, D, Conrad, S and Rippen R., 1998, *Corpus Linguistics: Investigating Language Structure and Use*, in Cambridge Approaches to Linguistics Series, Cambridge University Press

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

Cleveland, D.B., 1983, *Introduction to Indexing and Abstracting*, Libraries Unlimited Inc.

DeJong, G. 1982, An overview of the FRUMP system. In W. G. Lehnert and M. H. Ringle (eds), *Strategies for natural language processing*. Hillsdale, NJ: Lawrence Erlbaum, pp. 149 – 176

Graetz, N, 1985, Teaching EFL students to extract structural information from abstracts. In Ulign J. M and Pugh A. K. (eds) *Reading for Professional Purposes: Methods and Materials in Teaching Languages*, Leuven: Acco, pp. 123 – 135

Halliday M.A.K and Martin J.R., 1993, *Writing Science: Literacy and Discursive Power*, The Falmer Press

Hutchins, W. J., 1977, On the structure of scientific texts, *UEA Papers in Linguistics*, 5(3) pp. 18 – 39

Johnson, F, 1995, Automatic abstracting research, *Library review* 44(8)

Kennedy, G, 1998, *An Introduction to Corpus Linguistics*, Longman

Lancaster F. W., 1991, *Indexing and abstracting in theory and practice*, Library Association Publishing Ltd.

Luhn, H. P., 1958, The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2): 159 – 165

Narita, M, 2000, Constructing a Tagged E-J Parallel Corpus for Assisting Japanese Software Engineers in Writing English Abstracts, in *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'2000)*, Athens, Greece

Paice, C. D., 1981, The automatic generation of literature abstracts: an approach based on the identification of self-indicating phrases. In Oddy, R. N., Rijsbergen, C. J. and Williams, P.W. (eds.) *Information Retrieval Research*, London: Butterworths, pp. 172 – 191

Paice, C. D. and Jones, P. A., 1993, The identification of important concepts in highly structured technical papers. In *Proceedings of ACM-SIGIR'93*, pp. 123 – 135

Salanger-Meyer, F., 1990a, Discoursal flaws in Medical English abstracts: A genre analysis per research- and text-type, *Text*, 10(4), pp. 365 – 384

Salanger-Meyer, F, 1990b, Discoursal movements in medical English abstracts and their linguistic exponents: a genre analysis study, *INTERFACE: Journal of Applied Linguistics* 4(2) pp. 107 – 124

Sinclair, J.M., 2001, Preface. In Ghadessy, M., Henry, A. and Roseberry, R. L. (eds) *Small Corpus Studies and ELT: Theory and Practice*, John Benjamins

Swales, J. M., 1990, *Genre Analysis: English in academic and research settings*, Cambridge University Press

Tapanaine, P and Jarvinen, P., 1997, A Non-Projective Dependency Parser. In *Proceedings of the 5th Conference of Applied Natural Language*, pp. 64 – 71