

An Empirical Comparison of Predicates Used in Document Abstracting

Horacio Saggion¹

¹Department of Information and Communication Technologies
Universitat Pompeu Fabra
C/Tanger 122 - Campus de la Comunicació
Barcelona - 08018
Spain
horacio.saggion@upf.edu

Abstract. *We present an empirical study of the use of rhetorical predicates in abstracts written by professional abstractors. We are particularly interested in finding whether abstracts have some common structures and whether these structures can be predicted using features computed from actual text abstracts. We analyse a corpus of abstracts produced by three different abstractors and measure their relatedness in terms of rhetorical predicates. We also use a Support Vector Machines learning algorithm to investigate to what degree data from a human abstractor can be used to predict the structure of abstracts by a different abstractor. This study has implications for text abstracting systems aiming at simulating the way humans produce abstracts.*

1. Introduction

In recent years there has been an increased interest in the production of abstract and as a consequence research in areas such as sentence compression [Soricut and Marcu 2007, Banko et al. 2000], paraphrase [Barzilay, R. and Lee, L. 2004], and text abstracting operations [Oakes and Paice 2001, Jing and McKeown 2000, Saggion 2009] has intensified. If automatic summarization systems are to be based on text abstracts produced by humans, then the content and organization of abstracts should be assessed to identify regularities which could be simulated by machines. In this work we are particularly interested in examining the structure of abstracts produced by experienced abstractors in order to see if they share characteristics of text organization. Text organization refers to information types and how they are arranged in the text. Figure 1 shows abstracts where descriptions, results, explanations, etc. are some of the information types suggested by the highlighted predicates. We defined a set of metrics to compare the structures of abstracts based on direct comparison of the predicates used to signal information. This is an obvious limitation of this work, but allows us to establish a working methodology for further research. The rest of this short communication is organised in the following way: the next Section reports on related work on document structure for abstracting. In Section 3 we describe the dataset used for the experiments and next in Section 4 we detail the linguistic analysis of the text. In Section 5 we describe the adopted methodology while Section 6 discusses the results. Finally, Section 7 closes the paper.

2. Related Work

The work presented here relates to the problem of the conceptual and rhetorical structure of abstracts and to work aiming at generating abstracts automatically. [Liddy 1991] was

ABSTRACTOR I:

Describes the addition of a collection management module to an advanced reference course in the arts and humanities, and **reports** the preliminary findings of a practicum that used the Research Libraries Group Conspectus as a basis for collection analysis in the humanities for a university library.

ABSTRACTOR II:

Discusses the application of cellular technology for cellular data communications (CDC). The cellular phone system industry and market **are described**; barriers to CDC **are explained**; CDC applications for online searchers **are discussed**; and problems with using CDC for online searching **are reviewed**...

ABSTRACTOR III:

Describes FOCES (Foster Care Expert System), a prototype expert system for choosing foster care placements for children which integrates information retrieval techniques with artificial intelligence. The use of prototypes and queries in Prolog routines, extended Boolean matching, and vector correlation **are explained**...

Figure 1. Professional abstracts with inserted predicates indicating types of information contained in the abstracted document.

one of the first works to examine the structure of abstracts. By asking abstractors to list typical components of information in abstracts, she created a formal model of the abstracts conceptual structure. [Swales 1990] was a pioneer in the study of the organization of the structure of scientific abstracts. [Teufel and Moens 1999] used a high-level rhetorical model (similar to Liddy's) for detecting information in scientific texts and extracting informational components for summaries. [Rino and Scott 1996] developed a text representation based on rhetorical and semantic relations rich enough to make possible the preservation of essential information for summary generation. Predicates used to signal the structure and content of the texts (e.g. present, discuss, includes) similar to those studied here, have been incorporated in a computer-assisted system for the creation of abstracts, the TEXT System [Craven 1998]. [Montesi and Owen 2007] argued that these predicates are used to make abstracts more clear and objective. In [Saggion 2009], a classification algorithm based on text content and contextual features was developed to predict which predicates to insert when creating indicative sentences while in [Saggion 2011], the same problem is addressed using transformation-based learning. In cut-and-paste summarization [Jing and McKeown 2000] some abstracting operations are simulated, however the insertion of rhetorical predicates is not taken into consideration. Abstracts studied here are similar to indicative summaries studied in [Kan and McKeown 2002] aiming at generating indicative sentences for bibliographical data. [Saggion and Lapalme 2002] studied to some degree the insertion of predicates to create topical sentences for indicative abstracts.

3. Data Collection

Abstracts for this study were collected from the ERIC Abstracting database (<http://www.eric.ed.gov>) available at our institution. Using the search facilities pro-

vided by the ERIC abstracting service we have collected over 1,000 abstracts written by professional abstractors. For this study, however, we have considered only abstracts similar to those presented in Figure 1 and reduced the size of the sample abstracts to 369 items. These abstracts which were produced by three different abstractors at ERIC (we will call them *Abs-I*, *Abs-II* and *Abs-III*) have the following characteristics: they introduce information by means of a set of *rhetorical predicates* which are prepended or appended to information from the abstracted document in order to create new sentences. This is a usual mechanism for creating indicative abstracts. The formal structure of these abstracts matches the following pattern:

$$\text{Abstract} \equiv \bigoplus_{i=1}^n \text{Component}_i \oplus \text{Connective}_i$$

where Component_i follows one of the patterns below:

$$\text{Component}_i \equiv \text{Pred}_i^A \oplus \beta_i$$

$$\text{Component}_i \equiv \beta_i \oplus \text{Pred}_i^P$$

Pred_i^A is a predicate in active voice (e.g., “Presents”) used to introduce the “content” β_i of sentence (or clause) i and Pred_i^P is a predicate in the passive voice (e.g., “are presented”) also introducing the content β_i , n is the number of sentences in the abstract, \bigoplus indicates multiple concatenation, and $X \oplus Y$ indicates the concatenation of X and Y . Connective_i is a clause connector such as “.” or “;” or “, and”, etc. Note that professional abstracts can take many forms, but in this paper and to reduce the complexity of this study we are only considering abstracts which follow the above “linear” structure.

One can think of these predicates as signalling specific information types one may find in the summarized document. We understand these predicates with the definitions given in Table 1: for example the sentence “Describes FOCES...” in one of the abstracts in Figure 1 suggests that in the abstracted document a description of system “FOCES” will be found; and the sentence “... boolean matching, and vector correlation are explained” indicates that explanations of “boolean matching” and “vector correlation” will be found in the document. From the text summarization point of view we argue that the identification of a description in the document to be summarized can inform a generation components about the type of predicate to select to convey in an indicative way the found information. The list of predicates used in this study is drawn directly from the abstracts; there are many more predicates abstractors could use in order to produce indicative sentences but we believe the list given in Table 1 contains predicates commonly used.

4. Data Processing

Each electronic version of the abstracts was processed using the freely available GATE text analysis software [Maynard et al. 2002]. First each abstract was analyzed by a text structure analysis program to identify the abstract meta-data. After this, each abstract and document title was tokenized, sentence splitted, part-of-speech tagged, and morphologically analysed. Each sentence in the abstract was analysed by a string matching algorithm to identify the rhetorical predicates in each sentence. Rhetorical predicates can appear the beginning of a sentence (e.g. “*Discusses* the idea of...”), in the middle of a sentence (e.g. “Difficulties that*are discussed*, and ... ”) or at the end of the sentence (e.g. ...performance measures “*are reported*.”). The identified predicate or phrase is normalised and

Predicate	Function
consider	thinking about an issue reported in the abstracted document
contain	signals information included in the abstracted document
describe	an entity is described in the document
discuss	an issue is discussed in the document
examine	an issue is examined in the document
explain	an explanation is given in the document
explore	a topic is addressed in the document
focus	particular attention is paid to a topic
include	some information items are included in the document
present	some entity or topic is presented in the document
provide	some information is given
report	the paper gives details on an topic or entity or issue
review	a review is given in the paper
suggest	some suggestions are put forward in the paper

Table 1. Rhetorical Predicates in Abstracts

used to annotate the sentence: for example for “Discusses” the predicate is “discuss” and for the phrase “are reviewed” the predicate is “review”. The tokens corresponding to the identified predicate or phrase were eliminated from the set of document tokens.

5. This Study

In previous experiments classification algorithms [Saggion 2009] and rule induction systems [Saggion 2011] have been applied to try to predict the predicates prepended to sentence fragments using as training a set of abstracts. In this study we are interested in “measuring” how close abstracts from different abstractors are in terms of the rhetorical predicates used to create inductive sentences.

5.1. Comparing Abstracts

Because there is only one version of each abstract, we can not compare directly the abstracts’ content, but we can compare the predicates used which are drawn from a list of predicates used in abstracting. We use a number of metrics to compare the structure of the abstracts:

- Predicate distribution: we compute the distribution of the predicates each abstractor used in the dataset. This information gives an indication of the types of information each abstractor will include in the abstract.
- Predicate distribution correlation: we rank predicates for each abstractor according to their frequency of occurrence and compare the distributions using rank cor-

Abs II	%	Abs III	%	Abs I	%
focus	0.9	explore	0.3	contain	0.4
explore	1.2	report	1.3	explain	0.9
consider	2.4	present	1.6	suggest	0.9
present	2.7	provide	1.6	consider	1.3
review	2.7	review	2.9	focus	2.2
suggest	3.6	focus	3.5	present	2.6
examine	4.8	suggest	3.5	review	3.0
report	3.6	contain	3.8	report	4.3
explain	4.8	examine	5.1	include	6.1
provide	12.1	explain	5.9	examine	7.8
discuss	13.6	consider	7.8	explore	7.8
contain	13.9	include	13.4	provide	10.8
include	14.2	describe	18.2	discuss	22.9
describe	21.5	discuss	31.1	describe	29.0

Table 2. Predicates and Distribution in the Corpus for 3 Abstractors

relation [Siegel and Castellan 1988]. A correlation of 1 indicates that the predicates are used with similar distribution and correlation of -1 indicates that the predicates are used with opposite distribution.

- Predicate prediction: we use abstractor’s X predicate model to simulate abstractor’s Y predicate model. This is carried out training a machine learning algorithm on abstractor’s X data and applying the learnt model to abstractor’s Y data. The resulting predictions are compared to the true predicates using *accuracy*: the proportion of times the predicate was correctly predicted (e.g., if there are 10 occurrences of predicate “present” in a set of abstracts and only 5 of those occurrences are correctly predicted, then accuracy will be 50%). This metric in a sense give us an idea of the level of similarity among internalised structures.

6. Results and Discussion

In Table 2 we show the distribution of the predicates for each of the three abstractors. The predicates are listed for each abstractor sorted by frequency of appearance in the corpus. As can be appreciated, abstractors use same predicates but with a rather different distribution. Table 3 provides more insight into this, since it presents correlation figures. It can be seen that two abstractors (II and III) are likely to use a set of predicates such as “explore”, “focus”, “present”, etc. with similar frequency.

Abstractor	Abs I	Abs II	Abs III
Abs I	1.00		
Abs II	0.30	1.00	
Abs III	0.11	0.57	1.00

Table 3. Predicate Distribution Correlation

Table 4 shows classification results (e.g., accuracy at predicate level) using a machine learning classification algorithm (similar to the one used in [Saggion 2009]) to pre-

dict each of the predicates in an abstract. We have used for training the classification system the following features:

- the first three token roots of each clause (and excluding the predicate),
- the first three nouns of each clause,
- the three parts of speech of each clause,
- the position of the clause in the abstract,
- a cohesion feature indicating a link with the previous clause (whether two sentences share a noun),
- the number of punctuation marks in the clause, and
- the presence of nouns from the title in the clause.

The reported accuracy numbers are aggregated over all abstracts for each abstractor. We can see for example that abstractor I is able to predict with 20% accuracy the structure of abstracts from abstractor II. Here again, the results indicate that abstractor II abstracts resemble more to abstractor III abstracts than to abstractor I abstracts. Note that the numbers in the diagonal of Table 4 show to what degree an abstractor is able to predict his/her own abstracts. This is a cross-validation experiment where a set of abstracts from abstractor X is used for training the classification algorithm and applied to a different set of abstracts from abstractor X.

Abstractor	Abs I	Abs II	Abs III
Abs I	0.34	0.20	0.23
Abs II	0.29	0.50	0.31
Abs III	0.26	0.39	0.39

Table 4. Machine Learning Experiments Results. Diagonal Contains Results for 10-fold Cross-validation Prediction.

Predicate	Abs I ACC		Abs II ACC		Abs III ACC	
	Abs II	Abs III	Abs III	Abs I	Abs II	Abs I
consider	0	0	12	0	0	0
contain	0	0	0	96	100	0
describe	54	25	51	25	37	34
discuss	28	57	40	44	28	43
examine	11	0	0		0	11
explain	0	0	0	19	5	0
explore	0	6	0	0	0	100
focus	0	40	0	67	0	8
include	64	79	13	40	78	22
present	17	0	11	0	0	0
provide	12	0	15	32	67	0
report	10	0	0	50	40	20
review	0	0	0	11	0	0
suggest	0	50	0	56	8	0

Table 5. Abstractor's Predictions. Numbers are percent accuracies for prediction of individual predicates.

Finally, Table 5 shows prediction results for individual predicates. For example, if we take predicate “Suggest” we can see that:

- Abstractor I is unable to predict this predicate on data from Abstractor II (e.g. zero accuracy) but is able to predict it on data from Abstractor III with 50% accuracy.
- Abstractor II is unable to predict this predicate on data from Abstractor III (e.g. zero accuracy) but is able to predict it on data from Abstractor I with 56% accuracy.
- Abstractor III is unable to predict this predicate on data from Abstractor I (e.g. zero accuracy) but is able to predict it on data from Abstractor II with 8% accuracy.

Predicates such as “describe”, “discuss”, “include” and “provide” can all be predicted with some success.

As a general observation, here again, it is difficult to predict abstractor’s II and III data from abstractor’s I data. But many predicates in abstractor’s I data can be predicted using the others abstractors’ data. Prediction is more accurate when data for training and testing is drawn from the same abstractor.

7. Conclusions

In this paper we have presented a study into the organization of abstracts written by professional abstractors. We believe that it is the first study that compares the rhetorical organization of abstracts in terms of the predicates used to signal information components. The results appear to indicate that the relevance of information categories present in abstracts may vary from one abstractor to another. But because we have based our comparison on the superficial form of the predicates used to introduce the information, these results should be taken with caution. We suggest that the use of clustering techniques and dictionaries could be used to group predicates with same use. Experiments on that direction could help better assess the proximity between the abstracts.

Acknowledgments

We would like to thank two reviewers for the detailed analysis of the paper, we have tried to follow their suggestions to produce the final version of the paper. We are grateful to Programa Ramón y Cajal from Ministerio de Ciencia e Innovación, Spain.

References

- Banko, M., Mittal, V. O., and Witbrock, M. J. (2000). Headline generation based on statistical translation. In *ACL '00: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 318–325, Morristown, NJ, USA. Association for Computational Linguistics.
- Barzilay, R. and Lee, L. (2004). Catching the Drift: Probabilistic Content Models, with Applications to Generation and Summarization. In *Proceedings of HLT-NAACL 2004*.
- Craven, T. (1998). Human creation of abstracts with selected computer assistance too. *Information Research*, 3(4).

- Jing, H. and McKeown, K. (2000). Cut and Paste Based Text Summarization. In *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 178–185, Seattle, Washington, USA.
- Kan, M.-Y. and McKeown, K. R. (2002). Corpus-trained text generation for summarization. In *Proceedings of the Second International Natural Language Generation Conference (INLG 2002)*, pages 1–8, Harriman, New York, USA.
- Liddy, E. D. (1991). The Discourse-Level Structure of Empirical Abstracts: An Exploratory Study. *Information Processing & Management*, 27(1):55–81.
- Maynard, D., Tablan, V., Cunningham, H., Ursu, C., Saggion, H., Bontcheva, K., and Wilks, Y. (2002). Architectural Elements of Language Engineering Robustness. *Journal of Natural Language Engineering – Special Issue on Robust Methods in Analysis of Natural Language Data*, 8(2/3):257–274.
- Montesi, M. and Owen, J. M. (2007). Revision of author abstracts: how it is carried out by LISA editors. *Aslib Proceedings*, 59(1):26–45.
- Oakes, M. P. and Paice, C. D. (2001). Term extraction for automatic abstracting. In Bourigault, D., Jacquemin, C., and L’Homme, M.-C., editors, *Recent Advances in Computational Terminology*, volume 2 of *Natural Language Processing*, chapter 17, pages 353–370. John Benjamins Publishing Company.
- Rino, L. H. and Scott, D. (1996). A Discourse Model for Gist Preservation. In Borges, D. and Kaestner, C., editors, *Proceedings of the 13th Brazilian Symposium on Artificial Intelligence, SBIA’96*, Advances in Artificial Intelligence, pages 131–140. Springer.
- Saggion, H. (2009). A classification algorithm for predicting the structure of summaries. In *UCNLG+Sum ’09: Proceedings of the 2009 Workshop on Language Generation and Summarisation*, page 31–38, Morristown, NJ, USA. Association for Computational Linguistics, Association for Computational Linguistics.
- Saggion, H. (2011). Learning Predicate Insertion Rules for Document Abstracting. In *CICLing*, pages 301–313, Tokyo, Japan. Springer.
- Saggion, H. and Lapalme, G. (2002). Generating Indicative-Informative Summaries with SumUM. *Computational Linguistics*.
- Siegel, S. and Castellan, N. J. (1988). *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, New York, second edition.
- Soricut, R. and Marcu, D. (2007). Abstractive headline generation using WIDL-expressions. *Inf. Process. Manage.*, 43(6):1536–1548.
- Swales, J. (1990). *Genre Analysis. English in Academic and Research Settings*. Cambridge. Applied Linguistics.
- Teufel, S. and Moens, M. (1999). Argumentative classification of extracted sentences as a first step towards flexible abstracting. In Mani, I. and Maybury, M., editors, *Advances in Automatic Text Summarization*, pages 155–171. The MIT Press.