AUTOMATIC ABSTRACTING IN A LIMITED DOMAIN

Ruslan Mitkov

Machine Translation Unit
School of Computer Science
University of Science Malaysia
11800 Penang, Malaysia
Fax (60-4) 873335
Email ruslan@cs.usm.my

ABSTRACT

The present paper discusses a new knowledge-based approach proposed for automatic abstracting (AA) in a limited domain. Unlike most of the automatic abstracting models which do not make use of domain knowledge (an therefore are hardly capable of differentiating the "important" from the other concepts in the area) and linguistic knowledge (the summaries obtained usually do not undergo any additional modifications and the uncontrolled linkage of the text picked up for summary may sometimes be unacceptable), our approach is based on empirical, domain and linguistic knowledge, the latter enabling an additional treatment of the extracted text in order to obtain more coherent and natural text. The principal ideas presented in the paper are incorporated in the development of an automatic abstracting program.

INTRODUCTION

Research in automatic abstracting, which is usually concerned with the automatic construction of a short abstract from a long document, has concentrated so far in two approaches [Tong et al.91], and these are complementary rather than competing approaches. The first approach - *automatic text extraction* - is largely derived from the work of researches in the field of information science, where simple techniques of word frequency, cue words and other clues based on document structure are used to determine the set of important sentences to be extracted from the original document. The second approach - *automatic text summarization* - uses artificial intelligence (AI) and natural language processing (NLP) techniques to derive an understanding of a text document before generating an original summary.

Text extraction has had a longer history than text summarization, beginning with the first study by Luhn in 1958. Text extraction systems are traditionally confined to ad-hoc and simple techniques, without any symbolic or linguistic processing, and this limits the quality of extracts that can be produced.

Text summarization, on the other hand, tackles the text understanding problem head-on, but to date, there have been very few successful systems, and most of these are restricted to documents in small well-defined domains only. This is similar to the barrier faced by the fields of AI and NLP, where success in the small artificial world is rarely extensible to the large real world.

Given the complexity of the problem, we have oriented our research in a limited domain, which to our belief could be at present the only successful direction in AA.

The sublanguage of school geometry in Bulgarian has been chosen as a test language for our experiments. If the sample texts do not always represent perfect English, it is because they are literal translation from Bulgarian.

SHORT SURVEY OF PREVIOUS WORK

The first experiment on text extraction was by Luhn (1959), followed by the important work of Edmundson (1969), Earl (1970), Rush et al. (1971), Skorokhod'ko (1972), and Paice (1981). More recent works are reported by Black & Johnson (1988), Miller et al. (1990) and Paice (1990).

The general approach adopted by most automatic text extraction systems is first to figure out the important words in a document, then compute a score for each sentence based on these keywords, and finally generate an extract containing the higher ranked sentences. Various positive or negative clues may be used to identify the keywords and the sentences to be extracted, including the use of a stoplist (a list of common function words and noncontent-bearing words) and a frequency count (this is known as the frequency-keyword method, where a threshold value is used to separate out the keywords).

Research reported in text summarization is still *a long way from* truly general text understanding systems. Since any form of understanding must involve a huge amount of knowledge, it is necessary, from the practical point of view, to restrict the content of this knowledge base to a specific domain only. Systems that claim certain success in text understanding are those dealing with specific applications and in well-defined and restricted domains.

Important work in the area of text summarization include among others the works by Dejong (1982), Lehnert (1982), Froscher et al. (1983), Fum et al. (1985), Hayes (1985), Lytinen & Gershman (1986), Chiaramella and Defude (1987), Jacobs & Rau (1990), Reimer & Hahn (1988), Anderson et al. (1992), Yamaguchi et al.(92), Tsou et al.(92).

To make the difficulty of text understanding even clearer, we would mention a recent development ([Tsou et al.92]) which produces summaries *after a dialog with the user* who is being questioned by the system on the essential content of the text to be summarized. Since text understanding is a very complicated problem many current systems do not aim at really understanding the text but rather use other techniques ([Anderson et al.92]).

3. CONCLUSIONS AND A PROPOSAL FOR A NEW KNOWLEDGE-BASED APPROACH FOR AUTOMATIC ABSTRACTING IN A LIMITED DOMAIN

Text summarization as an approach is quite ambitious and though it is our distant goal, we are aware that its true implementation in general domains, particularly for industrial applications, is unrealistic. For instance, in order to derive the intention of the speaker/author from a paragraph, the discourse analysis should identify such aspects as focus and goal. Automatic recognition of the latter features, however, is an extremely complicated task. For practical reasons we propose at the initial stage of the project an intermediate paragraph partial understanding alternative, which is confined to recognizing the main domain concepts and the relations between them. In perspective, however, we envisage as complete as possible linguistic analysis of the input text.

Moreover, to our knowledge, no AA system tries to paraphrase the obtained summary. This, however, may lead often to unacceptable texts. Imagine a few sentences picked up to be important and all of them start with the same subject. Obviously, in most of the cases such sentences have to be coordinated into one sentence with the single subject. This simple observation illustrates the necessity of additional processing of the generated summary. Our approach is linguistically motivated and does not only generate a summary, but also *revises it* until most acceptable text is obtained.

We claim that knowledge is an indispensable prerequisite for the successful operation of an intelligent automatic abstracting system. An intelligent AA system should be able to deal with various heuristics derived from empirical observation which suggest when a text should be selected as essential or rejected as not important enough for the summary. Empirical rules which guide the process of AA could be very useful and practical.

However, empirical knowledge may not be sufficient. How could the program decide which concept (topic) is important for certain domain and which not, if it does not have the necessary knowledge? Previous methods have made use of calculating the frequency of the words encountered, but obviously this method does not always give reliable results. Recognizing when certain concepts are important for a domain is possible only if domain knowledge is available.

Linguistic knowledge is needed for preliminary analysis (understanding the text, anaphora resolution etc.) and final paraphrasing, in order to obtain good quality final summary (generation of a revised version of the text). However, we should be also aware of many complicated linguistic problems that may arise which could imply some additional efforts.

Our approach makes use of empirical, domain and linguistic knowledge for extracting text and its additional processing. The model developed *integrates* the three types of knowledge into an uniform architecture of separate but flexible and interrelated modules.

The approach is also sublanguage-oriented, since the knowledge incorporated is domain- (and therefore sublanguage-) dependent. Therefore we have concentrated on Automatic Abstracting in limited domains.

4.0 STRUCTURE OF THE AUTOMATIC ABSTRACTING MODEL

4.1 Integration of different types of knowledge into an uniform architecture

Our model is an integrated knowledge-based architecture which processes the input text at three levels: empirical, domain and linguistic (see Figure 1). First the text is processed in an interactive way by the empirical, domain and linguistic modules which determine which texts are important and offer a selection as a summary. Finally, the linguistic module processes the selected texts to generate a coherent final summary.

4.1 Empirical module: rules for selection and rejection of text sequences.

Since an automatic abstracting program should know how to select the essential text and to reject the texts which are not so important to the topic, we have developed an empirical module containing "rules of selection" and "rules of rejection" (we call them also "summary rules"). These rules, which simulate the experience of a human abstractor, are based on empirical observations and are to be constantly updated: our project aims at finding an optimal set of such rules. The rules have the form "if A then B" e.g. "if C is a substring of S and $C \in DKB \Rightarrow$ include S", where S is a sentence, DKB is a domain knowledge base of concepts and the string C represents a concept. The rules of selection propose for instance the inclusion of

- texts with key concepts (as defined in the knowledge base)
- texts which contain critical attributes describing these concepts
- texts with emphasis on certain fact (e.g. signalled by "it is important", "it is essential").

The rejection propose inter alia the elimination of

-examples
-certain comparisons (e.g. signaled by "unlike")
-consequences
-additional information in brackets
-notes.

The rules to be proposed are sublanguage-dependent, which in our case, is suitable for the considered domain.

4.2. Domain Module: what is important for a domain

The availability of an underlying knowledge base, describing the key concepts in a domain, is crucial. It is essential for a program to "understand" which are and subsequently to extract the most important concepts in a text and what is more - to capture their properties. The envisaged domain module contains a semantic description of the important for a domain concepts and their relations. In developing our domain module we have used the extended version of Markle&Tieman's model for semantic concept representation [Mitkov90]. This version describes each concept as a set of critical and variable attributes. The general semantic knowledge representation of a concept is given as:

- Concept: (superordinate / critical attribute 1,..., critical attribute n / variety 1 of variable attribute 1,..., variety m of variable attribute 1;...; variety 1 of variable attribute k,...,variety s of variable attribute k / an animate object indicator)

A particular concept from the domain of geometry - "triangle" is described in terms of the above model as follows:

Triangle (geometrical figure / plane, convex, straight linear, three sides / acute-angled, right-angled, obtuse-angled; equilateral, isosceles, scalene/0).

4.3 Linguistic module: restricted analysis and paraphrasing of the selected text into a final summary

Empirical and domain knowledge only are not sufficient. First, the input text should be at least partially understood. For instance, the linguistic module is supposed resolve references before figuring out the importance of the sentence. Moreover, in order to obtain a final summary, the obtained texts should paraphrased. The idea is to obtain more coherent and natural text through linguistic operations such as coordination, ellipsis construction, superordinate construction and various stylistic rules. These linguistic operations and the conditions under which they can be applied, are represented in the linguistic module. We should point out that for practical reasons this module contains more sublanguage knowledge than general language knowledge.

4.4 Sample text processing

Consider the following sample text from the sublanguage of school geometry:

The triangle is a plane geometrical figure with three sides. It is also straight linear: its sides are obviously straight linear. The triangle is convex. Unlike the other polygons, triangles cannot be concave.

The triangle can be isosceles, equilateral or scalene according to the nature of its sides. For instance, if two of its sides are equal, it is isosceles. So if it has three equal sides, it is called equilateral. and in case no sides are equal , the triangle is scalene.

Because of the different types of angles, the triangle can be right-angled, acute-angled or obtuse-angled according to the nature of its angles: if there is one angle right, the triangle is called right, if it has one angle obtuse, it is obtuse-angled and otherwise is acute-angled.

If this text is given as an input to the program, after application of the summary rules, it will yield:

The triangle is a plane geometrical figure with three sides. It is straight linear. The triangle is convex. The triangle can be isosceles, equilateral or scalene according to its sides. The triangle can be right-angled, acute-angled or obtuse-angled according to its angles.

Obviously this text is not fluent enough and cannot be presented as a final summary. The linguistic module coordinates sentences, generates elliptical constructions (but before that, it resolves the anaphor "it") in order to produce the following more fluent text:

> The triangle is a straight linear or convex plane geometrical figure with three sides. According to its sides, the triangle can be isosceles, equilateral or scalene and according to its angles, right-angled, acute-angled and obtuse-angled .

This example illustrates how important is that the obtained summary should be further revised in order to obtain more coherent natural language text.

The three modules are separate, but flexibly interdependent. The input text is analyzed with respect to its empirically expected significance, domain importance and linguistic references by all the three modules in an interactive way. The rules of the empirical and domain modules suggest which text sequences are to be included in the preliminary summary and the linguistic module generates its final revised version.

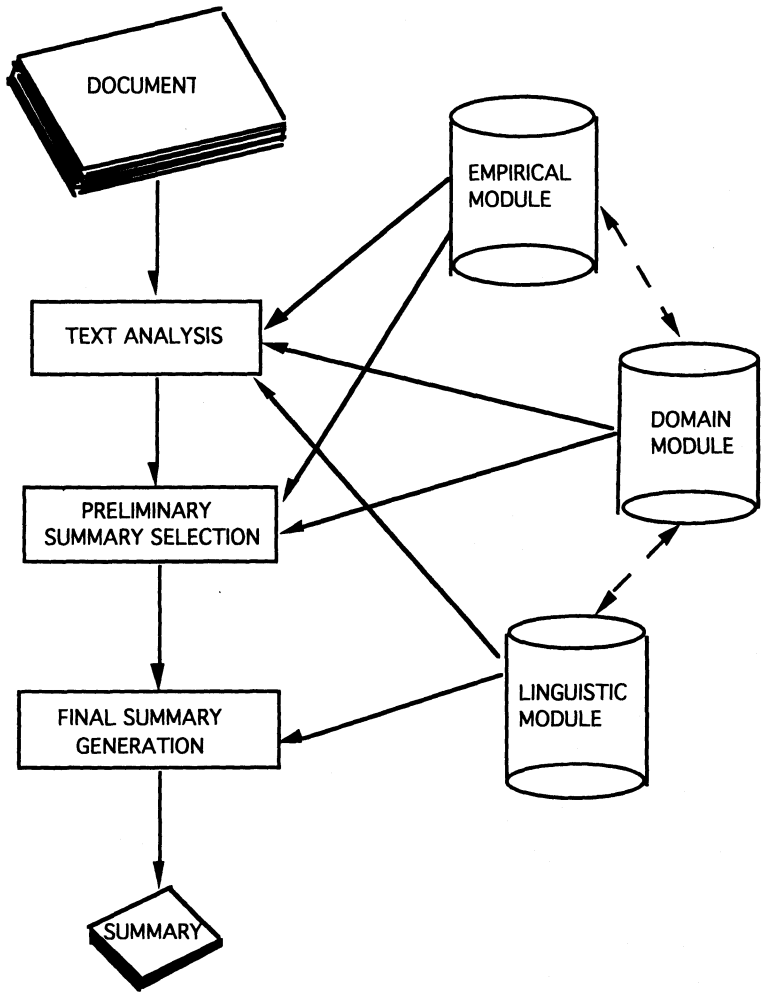

Fig. 1 Structure of the automatic abstracting model

5.0 CONCLUSION

The paper presents an preliminary model of a new knowledge-based approach for automatic abstracting in a limited domain. It is expected to have certain advantages in comparison with the so far known approaches and above all in the integrating different types of knowledge and summary revision. During the research, however, new interesting research problems such as domain-dependent anaphora resolution have arisen, and it will be worth

investigating them carefully too. Further investigations in rejection/selection criteria, paraphrase processing and knowledge representation methods are also desirable. The set of selection/rejection rules should be able to operate as a small expert system, which is supposed to stimulate the process of an experienced abstractor who selects the important parts of a text. Such rules are to be based on a general study on textual linguistics, an inquiry among professional abstractors and the results of a psychological experiments. So far as paraphrase generation is concerned, it is a very problematic research field; it may seem worthy to consider paraphrase operation models such as the one described by ([Harris76]). It would useful also to investigate contemporary efforts in generation of natural language with respect to text revision ([Kentaro et al. 91]). Finally, semantic knowledge representation models describing objects by the means of attributes and characteristic features may not be the most suitable in all domains, so that it would be necessary to study various approaches especially for more general applications.

Since at the present stage of NLP research real results can be expected only in limited domains, our project is initially restricted to a limited domain, but future investigations in more general domains should not be ruled out.

REFERENCES

[Anderson et al.92] Anderson P., Hayes P., Huetner A., Nirenburg I., Schamandt L, Weinstein - *Automatic Extraction of facts from press releases to generate news stories.* In Proceedings of the 3rd Conference on Applied Natural Language Processing, Trento, Italy, 31.March-3.April 1992

[Black 88] Black W., Johnson F. - A practical evaluation of two rule-based automatic abstracting techniques. Expert systems for information management, 1(3)

[Chiaramella et al 87] Chiaramella Y., Defude B. - A prototype of an intelligent system for information retrieval: IOTA. Information processing and management, 23(4), 1987

[De Jong82] De Jong G. - *An overview of the FRUMP system.* In Strategies for natural language processing . W. Lehnert and M. Ringle (Ed.), Hillsdale, NJ, Lawrence Erllbaum Associates, 1982

[Edmundson69] Edmundson H. P. - *New methods in automatic abstracting* - Journal of the ACM - 16 (2) , 1969

[Fum et al.85] Fum D., Guida G., Tasso C. - *Evaluating importance: a step towards text summarization.* IJCAI-85, Los Angeles

[Froscher J et al 83] Froscher J. Grishman R., Brachenko J. and Marsh E. - A linguistically motivated approach to automated analysis of military messages. AAAI-83, 1983

[Harris76] Harris, Z. - *Notes du cours de syntaxe* - Paris : Seuil, 1976

[Hayes 85] Hayes Y. - Automatic classification and summarization of banking telexes. IEEE Conference on AI Applications, Wahington DC, 1985

[Jacobs 92] Jacobs P. - *Joining statistics with NLP for text categorization.* In Proceedings of the 3rd Conference on Applied Natural Language Processing, Trento, Italy, 31.March-03.April 1992

[Jacobs et al.90] Jacobs P., Rau L.F. - *SCISOR; extracting information from on-line news.* Communications of ACM, 33(11), 1990

[Karlgren88] Karlgren G. - *Automatic Abstracting of content in text* - Nordic Journal of Linguistics - 11 , 1988

[Kentaro et al. 91] Kentaro I., Takenobu T., Hozumi T. - *An architecture for text revision.* Proceedings of the Natural Language Processing Pacific Rim Symposium, 25-26 November 1991, Singapore

[Lehnert82] Lehnert W. - *Plot units: a narrative summarization strategy.* In "Strategies for natural language processing", W. Lehnert and M. Ringle (Ed), Hillsdale, NJ, Lawrence Erllbaum Associates, 1982

[Le Roux90] Le Roux D. - *Automatisation de l'activité résumante* - essai de typologie, Colloque "Le resumé de textes", 12-14 September 1990

[Lytinen et al. 86] Lytinen S., Gershman A. - ATRANS: automatic processing of money transfer messages, AAAI-86, 1986

[Miller 90] Miller I., Ibuki J., Nishino F. - The construction and evaluation of ECON - an English text condensing system. Proceedings of PRICAI-90

[Mitkov90] Mitkov R. - *Generating explanations of geometrical objects.* Computers and Artificial Intelligence, 9, 1990

[Mitkov et al. 93] Mitkov R., Le Roux D. , Descles J.P. - *Knowledge-based automatic abstracting: experiments in the sublanguage of elementary geometry.* Proceedings of the I International Conference on Mathematical Linguistics, Barcelona, 5-7 April 1993

[Paice81] Paice C. D. - *Automatic generation of literature abstracts: AN approach based on the identification of self indicating phrases.* In Information Retrieval Research. Oddy R., Robertson S., Van Rijsbergen C., Williams P. (Ed.), Butterworth, London, 1981

[Paice90] Paice C. D. - *Construction literature abstracts by computer: Techniques and prospects.* Information processing and Management, 26(1), 1990

[Reimer 88] Reimer U., Hahn U. - Text condensation as knowledge base abstraction. IEEE Conference on AI applications, 1988

[Rush71] Rush J., Salvador R., Zamora A. - *Automatic abstracting and indexing. Production of indicative abstracts by application of contextual inference and syntactic coherence criteria.* Journal of American Society for Information Science, 22, 1971

[Sharp89] Sharp B. - *Elaboration and testing of new methodologies for automatic abstracting*, Ph.D. thesis, University of Birmingham, 1989

[Sharp91] Sharp B. - *Informex - An Information Extractor System* .In Proceedings from the International Conference "Current Issues in Computational Linguistics", Penang, Malaysia, 10-14 June, 1991

[Skorohod'ko72] Skorohod'ko E - *Adaptive method of automatic abstracting and indexing.* Information Processing 71, North Holland, 1972

[Tong et al.91] Tong L. Ch. , Low P. L. - *Automatic text abstraction - prospects and proposed R&D plan.* Information technology, September 1991, Vol. 4 No.2

[Tsou et al.92] Tsou B., Ho H., Lai T.B., Lun C., Lin H. - *A knowledge-based machine-aided system for Chinese text abstraction.* Proceedings of the 14. International Conference on Computational Linguistics, COLING'92, 23-28 July 1992, Nantes

[Yamaguchi et al.92] Yamaguchi H., Aichi K., Nagai H., Nomura H. - *Analysis of text and figure/table relationships and its application to chart style summarization.* Proceedings of the International Symposium on Natural Language Understanding and Artificial Intelligence, Fukuoka, Japan, 13-15 July, 1992