ON SOME ROMANIAN TEXT SUMMARIZATION CONTRIBUTIONS

FLORENTINA HRISTEA, MARIUS POPESCU

Automatic summarization is a field which has been in existence since the 1960's. Overall, automatic summarization is a highly interdisciplinary application, involving natural language processing, information retrieval, library science, statistics, cognitive psychology and artificial intelligence. Although the field has its roots in the late 50's, and has been developed for decades, today it continues to grow and has become even more important taking into account the Internet and the WWW. Under these circumstances, it becomes crucial to note that attempts to perform text summarization with reference to the Romanian language date far back in time, with Romanian researchers showing an interest in the field ever since the early 70's.

Important early attempts to perform automatic summarization in Romanian belong to Erika Nistor and Eliza Roman who, in a group of papers published in the 70's, start out by using Luhn's algorithm for automatic creation of abstracts, described in Luhn (1958). This technique for "automatic creation of literature abstracts" relies on statistically determining the *key words* of the text. The authors (Nistor, Roman 1970a) perform tests for Romanian involving the manual statistical processing of 75 pages of text, representing 21 papers or short book chapters belonging to various fields (medical science, electronics, mathematics, history, music) and discuss the results while analyzing the difficulties which occurred during the experiment, some of which are typical of the Romanian language.

Let us note the usage of Luhn's method in early attempts to perform text summarization in Romanian, a method which relies on dividing each sentence into segments bracketed by significant terms (i.e. commonly-occurring, stoplist-filtered terms) not more than four non-significant terms apart. Luhn scores each segment by taking the square of the number of bracketed significant terms divided by the total number of bracketed terms. As it is commented in Mani (2001), "this type of segmentation method is less semantic in nature than other methods which involve using text cohesion for topic segmentation"[1]. It has been chosen by the Romanian

[1] Let us remind the reader that there are two broad approaches to summarization that can be identified: the *shallow approaches* which, as it is commented in Mani (2001): "do not venture beyond a syntactic level of representation, although different elements may be represented at different levels. For example, words may be analyzed to a semantic level, but sentences will be analyzed at most to a syntactic level. These approaches typically produce *extracts*, usually by extracting sentences". As

RRL, **LI**, *3–4*, p. 507–509, Bucureşti, 2006

authors working in the field in order to perform tests in the case of the Romanian language. On this occasion the authors have felt the need of having available a stoplist, namely a list of stop-words for Romanian. Thus, the absence of certain linguistic resources in electronic format, corresponding to the Romanian language, which do not fully exist even today, is noticed and noted by Romanian researchers in the field since the early 70's.

The mentioned research (Nistor, Roman 1970a) is extended in Nistor, Roman (1970b) and Nistor, Roman (1971), where the authors perform a text division into *kernel-sentences*, the so-called "information quanta", of standard form. The transformational approach used here follows Chomsky's suggestion to divide the text into kernel sentences. While Chomsky finds 9 categories of kernel sentences for English, which he calls "information quanta", Nistor and Roman attempt a similar study for Romanian and apply the results to Romanian text summarization. It is up to Romanian linguists to evaluate the quality of these results (both theoretical and practical), with respect to the Romanian language, but the importance of the attempt itself, as well as the necessity of this analysis are beyond any doubt.

After the text has been divided into kernel-sentences, bearing 8 standard forms established by the authors, the abstract of the text is automatically constructed. A question which naturally arises is the following: which are the most efficient key-words the most frequent kernel-sentences or the most frequent words? The authors come to the conclusion that "the words with the greatest frequency are at the same time the subjects of the kernel-sentences with the greatest frequency. This means that Luhn's method can be combined with the transformational one, building the abstract from the most frequent kernel-sentences".

In further work the two mentioned Romanian authors broaden their concept of abstract, which is an important step at the time. Thus, Nistor and Roman (1979) presents a method which attempts to identify the sentences referring to a subject area given a priori by the user's request. As the authors point out, "two ideas underlie this work: (1) to use fuzzy sets and their function to measure semantical similarity between document-text and key-words and (2) to use the thesaurus – available or especially constructed for this purpose – in order to quantify the distance between the text and the request". The main conclusion of this study is that "the relation between the set of documents and 'their' abstracts is no longer biunivocal. There exists a document and as many abstracts as many requests are referred to it". This is the reason why the authors call their method "dynamic abstracting".

A summary of the methods used by the mentioned authors for text summarization can be found in Nistor and Roman (1980).

opposed to them, the *deeper approaches* "usually assume at least a sentential semantics level of representation". As noted in Mani (2001), "they produce *abstracts*, and the synthesis phase here usually involves natural language generation from a semantic or discourse level representation".

An attempt to parallelize their text summarization techniques takes place in (Roman 1987), at a moment in time when parallel computing was still in its childhood.

For now let us note the importance of the two Romanian researchers having performed text summarization tests for the Romanian language since the early 70's. These contributions should stimulate current work in the field, both on the part of computer scientists and on that of the linguists, the latter being the only ones truly qualified to evaluate the results.

REFERENCES

Luhn, H. P., 1958, "The automatic creation of literature abstracts", *IBM Journal of Research and Development*, 2, 2, 159–165.

Mani, I., 2001, *Automatic Summarization*, Amsterdam/Philadelphia, John Benjamins Publishing Company.

Nistor, E., E. Roman, 1970a, "Câteva probleme legate de rezumarea automată a textelor româneşti", *Studii şi cercetări lingvistice*, XXI, 59–77.

Nistor, E., E. Roman, 1970b, "Transformations in automatical abstracting", *Cahiers de linguistique théorique et appliquée*, VII, 143–158.

Nistor, E., E. Roman, 1971, "Constructing automatical abstracts from kernel-sentences", *Cahiers de linguistique théorique et appliquée*, VIII, 249–256.

Nistor, E., E. Roman, 1979, "Dynamic abstracting", *Revue Roumaine de Linguistique*, XXIV, 186–191.

Nistor, E., E. Roman, 1980, "Încercări de rezumare automată", *Probleme de informare şi documentare*, 14, 1, 5–18.

Roman, E., 1987, "Calculul paralel şi rezumarea automată", *Informarea documentară în teorie şi în practică (mapă documentară)*, I, 1, 28–37.