Kaili Müürisep, Pilleriin Mutso. ESTSUM - Estonian newspaper texts summarizer. Proceedings of The Second Baltic Conference on Human Language Technologies. April 4-5, 2005. Tallinn, 2005. Pp. 311-316.

ESTSUM – ESTONIAN NEWSPAPER TEXTS SUMMARIZER

Kaili Müürisep, Pilleriin Mutso
University of Tartu, Estonia

Abstract

This article describes an experimental software system for automatic summary generation of Estonian newspaper texts called EstSum. EstSum constructs short summaries of text by selecting the key sentences that characterize the document. Sentences are ranked for potential inclusion in the summary using a weighted combination of statistical, linguistic and typographic features like the position, format and type of sentence, and the word frequency. During the testing, a corpus of 10 hand-created summaries of neswpaper articles was used. The summarizer's output was compared to the handmade summaries and the percentage of overlapping sentences was 60% in average.

Keywords: summarization, Estonian language

1. Introduction

As the amount of on-line information increases, more and more effort is dedicated to creating automatic summarization systems. Since the automatic text summarization is largely a language-specific task, suitable algorithms must be found for each natural language. This paper describes our summarization work for Estonian.

According to Radev et al, a summary can be loosely defined as a text that is produced from one or more texts, that conveys important information in the original text (s), and that is no longer than half of the original text(s) and usually sicnificantly less than that. In other words, the main goal of a summary is to present the main ideas in a document in less space (Radev et al. 2002).

Most research on summary generation techniques still relies on extraction of important sentences from the original document to form a summary. There are several methods for measuring the importance of a sentemce. Some algorithms calculate a weight for each sentence, taking into account the position of the sentence and word frequencies (Dalianis et al. 2003), while other algorithms use semantic information (e.g. WordNet), in order to find the hierarchy of concepts.

There are also different methods for summary generation from a single document and from multiple documents.

Summarization tool for Estonian (EstSum) focuses on extraction methods from a single document. Also the area of texts is limited: EstSum considers that the input text is formed as news text.

EstSum can be considered as the first tool for automatic summarization adapted for Estonian. It should be noted that there exist some summarization systems that are language-independent, with MS Office AutoSummarizer probably being the best known example (the algorithm behind it has not been publicly released). SweSum is another well-known language-independent summarizer which is also publicly available on the Internet (Dalianis 2000).

2. Overview of EstSum

EstSum has been written in Perl language, and it consists of three modules: HTML converter, sentence splitter and extractor.

HTML converter removes unimportant tags, normalizes the crossing labels and converts input to SGML format. It marks the headers and subheaders using font information, gives special labels to captions of photos and removes tables. It also preserves the important information about font, distinguishing between bold, italic and default font.

Sentence splitter is uses the rule-based approach for processing its input, employing 30 rules that consider the different cases of sentence beginnings and endings.

EstSum has two options for calculating text compression rates. With the first option, EstSum considers sentences as units, and when the text of 100 sentences is compacted by 30%, the generated summary has 30 sentences. With the second option, words are considered as units that sometimes helps to exclude long sentences from the summary.

EstSum extracts salient sentences from the text using location, format and keyword based information about sentence. The overall method of scoring sentences for extraction is based on a linear function of the weights of each of the three features, similar to Edmundson's style formula (Edmundson 1969; Mani 2001):

(1) $W(s)=\alpha P(s)+\beta F(s)+\gamma K(s)$

Here W(s) is the weight function of sentence s, P(s) is the position-based score function, F(s) the format-based score function and K(s) the keyword-based score function; α, β and γ are constants.

The feature weights and tuning parameters α, β and γ have been adjusted by hand using a manually created training corpus of extracts and the knowledge of authors.

The coprus of extracts is relatively small (only 20 texts) for drawing any final statistical conclusions. Despite this the main tendencies for selecting salient sentences are detectable. The length of extracts is 30% of length of the original texts. The smallest original text contained 4 sentences and the largest one 41 sentences, with the average text length being 18 sentences. The texts belonged to various genres - short news, columns, feature stories and one interview.

2.1. Position-based scoring

Position-based scoring considers the sentence location. In order to find appropriate weights for position-based scoring, we investigated how the summaries in the training corpus reflect the first 3 sentences of the original text, the first sentence after each subtitle and the first 2 sentences of each paragraph.

We established that the most influential sentences are the sentences following the title – the first sentence of the text was included in the summary in 100% of the cases, the second and the third sentence in 65% of the cases. The sentences immediately following the subtitles were included in the 60% of the cases.

We also found that the first sentence of the paragraph was included in the summary in 40% of the cases, and the second and the third in 20% of the cases. In addition, 20% of

the summaries contained the last sentence of the text. The position-based scores are given in Table 1.

Table 1. Position based scores

Feature	Percentage in extracts	Given score
1st sentence in article	100	10
2nd sentence in article	65	7
3rd sentence in article	65	7
1st sentence after subheader	60	6
1st sentence in paragraph	40	4
2nd sentence in paragraph	20	2
3rd sentence n paragraph	20	2
Other	6	0

The scores are normalized using formula (2).

$$n = \frac{p \cdot 100}{t} \quad (2)$$

Here n is normalized score, p is assigned score of the sentence and t is total of all position scores in the article.

2.2. Format-based scoring

Format-based scoring considers the sentence font (default, bold or italic) and punctuation marks (exclamation and question marks, double quotes). Figure captions and the text author are also detected and given minimum scores. Table 2 depicts the features and scores.

Table 2. Format based scores

Feature	Percentage	Given score
Default font	32	3
Bold or italic	70	10
Question or exclamation mark in sentence	10	0
Quotation marks in sentence	18	2
Captions, authors, subheaders	0	0

The scores are normalized using same algorithm as formula (2).

2.3. Keyword-based scoring

The first version of EstSum did not use any linguistic modules, so it was possible to use only word forms instead of roots.

Keyword-based scoring uses two techniques for detecting keywords: finding words that are relatively frequent in this article and not very frequent in general word frequency table; extracting words from the text title and all subtitles.

However, when inspecting the training corpus, we discovered that only 48% of the sentences containing words from the titles were included in summaries. Also, if extra score is assigned to sentences containing most frequent words, then only 25% of the sentences with highest scores are actually present in summaries. Therefore, when discovering frequent word forms, the summarizer must also employ a general word

frequency table for a given language, in order to estimate whether the word form appears more frequently than it normally does in texts written in that language.

Our keyword-based scoring algorithm also uses a general word frequency table that is generated from the newspaper texts of 400,000 words. The table lists word frequencies per 10,000 words and contains 1100 words that occur at least once in texts of 10,000 words.

The words belonging to the title (article headline) and subtitles are given extra scores. (5 and 2 points respectively). All the other words are put into the local frequency table with a weight 1.

2.4. Tuning general parameters

In order to tune general parameters, we measured how many of the sentences in summaries are found by applying each weight function separately. The position-based weight function assigned high scores to the first 3 sentences of the text, and the format-based function assigned high scores to the first 1-2 sentences of the text, while the rest of the sentences received relatively similar scores from all methods. Since the position- and format-based function yielded better results, we decided to use 0.4 as coefficient for them in the formula (1), while the coefficient for the keyword-based function was set to 0.2.

With these settings, 51% of the sentences present in the training corpus are also chosen by the EstSum summarizer for inclusion in the summaries (the figure of 51% does not reflect the text titles which belong to summaries by default).

3. Evaluation

Evaluating automatically generated summaries is not a straightforward process. The evaluation is usually made by comparing automatically generated summaries to summaries compiled by humans. Such evaluation gives good results in other domains of language technology like tagging and parsing, but sentence selection for summary is not so well defined task, and the summaries may be subjective depending on the author's interests and the mood of the moment. Hassel (2003) has found that at best there was a 70% agreement between summaries generated by two individuals, Radev and others have reported a figure of 60% (Radev et al. 2002).

3.1. Corpus for evaluation

The small corpus for evaluation consists of 11 texts with the average length of 321 words and 23 sentences. These texts are more uniform by their genre (front page stories, domestic news, business news and sports news from one newspaper).

3.2. Results

EstSum was able to choose 60% of sentences from the evaluation corpus as an average. In the best case the figure was 85.7% and in the worst case it was 0%. In the latter case the text was a very short newspaper article and EstSum chose the article title for the summary, while the manually compiled summary was longer than 30% of the words.

3.3. Comparison with other tools

SweSum is a freely available[1] summarizer that has some common features with EstSum. SweSum has been desinged for the processing newspaper text, and thus it uses so called position score: the sentences in the beginning of the text are given higher scores than the ones in the end. HTML tags which indicate sentences with bold text are given a higher score than the ones without bold text tagging, dito title tagging. Sentences containing numerical data are given a higher score than the ones without numerical values. Sentences which contain keywords are scored high. SweSum has a linguistic module for Swedish

[1] http://swesum.nada.kth.se/

texts that finds the stem of each word. For Estonian, we used SweSum with a generic language option. All the above parameters were normalized and put in a naïve combination function with no special weighting to obtain the total score of each sentence. (Dalianis 2000). SweSum without the linguistic module selected 41% of the sentences from the manually created test corpus.

Since 1997, the Microsoft Word editor has also a summarizer for documents. Unfortunately, it performs rather poorly on Estonian texts. For example, in a number of cases the summarizer is unable to detect sentence boundaries. During our experiments we found that approximately 25% of the extracted sentences were same as in the benchmark corpus.

4. Conclusions and future extensions

The automatic text summarizer tool EstSum presented in this paper can still be regarded as a prototype and is thus rather actively developed. Although the preliminary results by EstSum are excellent when compared to other two systems described in this paper, the evaluation was carried out on relatively small data sets, and therefore EstSum needs a considerable amount of further development and testing. Our current plans include the addition of a linguistic module to the EstSum framework for morphological analysis and morphosyntactic disambiguation. The employment of the linguistic module would make the keyword detection more efficient. We also plan to work on pronoun resolution, in order to make the summarized text more coherent.

Apart from developing EstSum, another important task is the creation of a larger training and test corpus that could be used for advanced statistical analysis and machine learning. The methods of measuring the summarizer performance are also a subject of further research.

References

Dalianis H. 2000. SweSum – A text summarizer for Swedish. *Technical report TRITA-NAP0015, IPLab-174*, NADA, KTH. October 2000. Retrieved February 27, 2004, from http://www.nada.kth.se/~hercules/Textsumsummary.html

Dalianis, H., M. Hassel, J. Wedekind, D. Haltrup, K. de Smedt and T.C. Lech. 2003. Automatic text summarization for the Scandinavian languages. In Holmboe, H. (ed.) *Nordisk Sprogteknologi 2002: Årbog for Nordisk Språkteknologisk Forskningsprogram 2000-2004*. Museum Tusculanums Forlag. 153-163.

Edmundson, H.P. 1969. New methods in automatic abstracting. In: *Journal of the Association for Computing Machinery 16 (2).* 264-285. Reprinted in: Mani, I.; Maybury, M.T. (eds.) *Advances in Automatic Text Summarization.* Cambridge, Massachusetts: MIT Press. 21-42.

Hassel, M. 2003. Exploitation of Named Entities in Automatic Text Summarization for Swedish. In the *Proceedings of NODALIDA '03* - 14th Nordic Conference on Computational Linguistics, May 30-31 2003, Reykjavik, Iceland.

Mani Inderjeet 2001 Automatic summarization. Amsterdam: John Benjamins Publishing Co.

Radev D. R., E. Hovy, K. McKeown. 2002. Introduction to the special issue on summarization. *Computational Linguistics.* Vol 28(4). 399-408.

KAILI MÜÜRISEP is a researcher at the Institute of Computer Science, University of Tartu. She received her Ph. D. (computer science) at the University of Tartu, dealing with automatic syntactic analysis of Estonian. Her research interests concern automatic syntactic analysis of written and spoken language, treebanks and automatic summary generation. E-mail: kaili.muurisep@ut.ee.

PILLERIIN MUTSO is a master student at the Institute of Computer Science, University of Tartu. Her research interests concern automatic summary generation and the evaluation of generated summaries. E-mail: pmutso@ut.ee.