The Use of Topic Segmentation for Automatic Summarization

Roxana Angheluta, Rik De Busser and Marie-Francine Moens
Katholieke Universiteit Leuven
Interdisciplinary Centre for Law & IT
Tiensestraat 41, B-3000 Leuven, Belgium
{roxana.angheluta,rik.debusser,marie-france.moens}
@law.kuleuven.ac.be

Abstract

Topic segmentation can be used as a pre-processing step in numerous natural language processing applications. In this short paper, we will discuss how we adapted our segmentation algorithm for automatic summarization.

1 Introduction

Human readers are able to construct a mental representation of the organization of a text in an efficient and intuitive way. Despite the immense variation of a text's thematic structures, some general patterns return, such as the hierarchical organization of a text into topics and subtopics, topic concatenation, and semantic return. We have developed a topic segmentation algorithm, which detects thematic structures in texts using generic text structure cues. It associates key terms with each topic or subtopic and outputs a tree-like table of content (TOC). We refer to this process as 'layered topic segmentation'. For the DUC 2002 summarization test, we used these TOCs for automatic summarization, which is possible because the text structure trees reflect the most important terms at general and more specific levels of topicality and indicate topically coherent segments from which sentences are mined for inclusion into summaries.
We used the TOCs for constructing both the single-document abstracts and the multi-document abstracts and extracts. For the 50-, 100- and 200-word abstracts as well as for the 200- and 400-word extracts of multiple documents we have clustered individual sentences from single-document summaries and have extracted the representative object (medoid) of each cluster to be included in the summary.

2 Layered topic segmentation

The topic segmentation algorithm uses generic topical cues for detecting the thematic structure of a text (Moens and De Busser 2001). After the text is tagged and chunked[1], three processes interact to construct a topic hierarchy.
In a first optional step, lexical chains are built for the nouns in the text, using synonymy relations in WordNet. We use an algorithm that is comparable to the one developed by Barzilay and Elhadad (1999). The words of the text are replaced by their most representative synonym (i.e. the most frequent member of the chain that first occurs in the text). Words that bear little on the content and whose elimination does not harm to the grammaticality and coherence of the text (e.g. common adjectives) might be removed. Collocations of two or three words are extracted from the text, using an algorithm that combines frequency counting with likelihood ratios (Dunning 1993).
In a second step, the main topic of each sentence is determined, i.e. the content word or word group that reflects the aboutness or topical participant. We identified two heuristics that are applicable to the languages we work with: the initial position of noun phrases and persistency of the topic term (cf.

[1] Edinburgh Language Technology Group - LT-CHUNK tagger and chunker

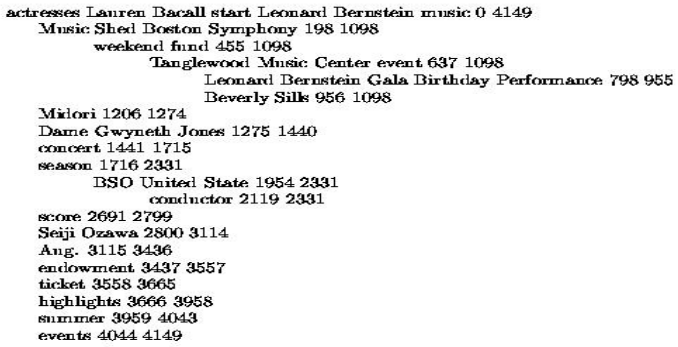

Figure 1: Example of a table of content made of doc. AP880720-0262.S, set d072f

Givón 2001). In languages that primarily have an SVO order – such as English, French and Dutch – noun phrases in a clause-initial position tend to be indicative of the topic of the sentence and of its most important information. Also, the main topic of a sentence usually occurs persistently in consecutive sentences. Other generic heuristics, such as definiteness or noun phrase embedding will be implemented in the future.

A third step in determining the topics and subtopics takes into account the distribution of topic terms in the text. It is generally agreed upon that the main topics of a text are signaled by terms that occur throughout the text, while subtopics are signaled by terms that are aggregated in limited passages (Hearst 1997).

Detection of the main sentence topics and of the term distribution identifies topically coherent segments and aids in detecting topic shifts, nested topics and semantic returns and in finding the most appropriate segmentation.

As more information becomes available from these heuristics, a table of content – a tree-like structure indicating the organization of topics and subtopics in a text – is gradually built and corrected. For each topic, the coordinates of the corresponding text segment and topic terms are added (see Figure 1).

Layered topic segmentation – i.e. topic segmentation that takes into account topic hierarchies – could be a useful preprocessing step in NLP applications such as information retrieval and information extraction.

3 Summarization

By restricting the number of levels of the TOC, it could already be used as a kind of short summary. For DUC 2002, we exploit the TOCs for text summarization in alternative ways.

For the summarization of single documents we use the hierarchical structure of the TOC: the predefined length of the summary dictates the level of topical detail of the summary as it can be derived from the TOCs. The first sentence of each topical segment at the chosen level of detail is included in the summary.

For the 10-word abstracts of multiple documents we select the 10 non-redundant topic terms with highest coverage in the articles computed with the coordinates of the TOCs, giving priority to terms that also occur in the articles' headlines (when they are present) and possibly ordering them as they appear in the original sentences.

For the 50-, 100- and 200-word abstracts of multiple documents, we start from the summaries of the single texts. We cluster the term vectors of sentences (which are restricted to nouns, adjectives and verbs, which are all open word classes) with two different methods: covering and *k*-medoid. In the covering clustering algorithm possible representative sentences (medoids) are considered for a potential grouping, each sentence having at least a

given similarity with the medoid of its cluster. The medoids are included in the summaries (cf. Moens et al. 1999). The objective is to minimize the number of medoids while fitting the predefined length of the summary.

The *k*-medoid method attempts to detect *k*-clusters for which the total similarity of each sentence and its medoid is maximized. The value *k* is determined as the clustering that, within the allowed summary length, maximizes the similarity between a sentence and its medoid and minimizes the similarity of the sentence with its second choice cluster.

For the 200- and 400-word extracts of multiple documents we cluster the term vectors of the sentences of single-document summaries as they occur in the text.

4 Results

For the DUC 2002 summarization test, we tried to match the length of our summaries as closely as possible to the required word length, which means that we neglected the parameter of brevity. This explains why the mean coverage values obtained for our summaries tend to be better than our mean length adjusted coverage values, unlike the results of other systems (see Figures 2, 4).

Evaluation of the single-document summaries gives us some insights into the applicability of the topic segmentation to automatic summarization (see Figure 2). A plot on the coverage for the single-document summaries is in Figure 3. The single summaries are quite satisfactory given that they are solely based upon the technique of layered topic segmentation.

	Mean	Our team	Best result	Worst result
Mean coverage	0.30438	0.361	0.388	0.057
Mean length adjusted coverage	0.25861	0.251	0.339	0.213
Mean quality questions	0.64192	0.660	0.407	1.281

Figure 2: The complete results for the single-document abstracts

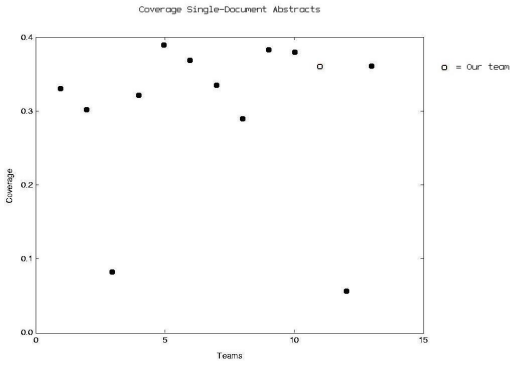

Figure 3: The results for the coverage for the single-document abstracts

In some preliminary experiments we tried out replacing words by representative synonyms, using the WordNet synonym relationships. However, we found out that neither for single document, nor for multiple document summaries, it did substantially improve the quality. Disregarding grammaticality issues (replacing each member of a lexical chain by the most representative member of a chain can result in an incorrect agreement between nouns as subjects and verbs) the number of good summaries from texts in which words had been replaced by synonyms is more or less equal to the ones for texts in which no replacements were made. In the DUC corpus the synonym replacement did not affect much the topic segmentation and the subsequent summaries.

With regard to the results sent to the DUC, we only used synonym replacement for the 10-word abstracts.

The results of the 10-word abstracts are not particularly impressive (see Figure 5). This can be explained by the fact that we extracted isolated words rather than phrases, and single words rarely match the peer units used by the human abstracters in evaluation.

For the multi-document summarization tasks – i.e. for the 50-, 100-, and 200-word abstracts and the 200- and 400-word extracts – we tested two clustering algorithms on the term vectors of sentences of the single-document summaries (see Figures 4 and 5). Restricting the term vectors to words that are nouns, verbs or adjectives seems to be fruitful. After more evaluation of the clustering algorithms,

it seemed that the covering method performs better than the k-medoid, but this is a hypothesis that needs further verification.

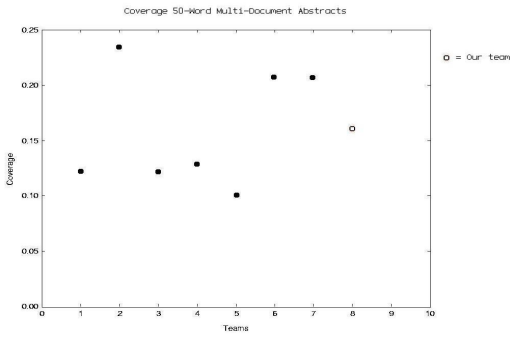

Figure 4: The results for the coverage for the 50-word multi-document abstracts

	Mean	Our team	Best result	Worst result
10-word abstracts				
Mean coverage	0.1865	0.091	0.390	0.091
Mean length adjusted coverage	0.19816	0.060	0.305	0.060
50-word abstracts				
Mean coverage	0.16012	0.161	0.234	0.100
Mean length adjusted coverage	0.149	0.145	0.180	0.102
Mean quality questions	0.77937	0.754	0.461	1.295
100-word abstracts				
Mean coverage	0.17362	0.141	0.235	0.122
Mean length adjusted coverage	0.13525	0.111	0.178	0.094
Mean quality questions	0.95837	1.008	0.735	1.259
200-word abstracts				
Mean coverage	0.202	0.165	0.253	0.151
Mean length adjusted coverage	0.14912	0.147	0.184	0.104
Mean quality questions	1.04162	1.085	0.897	1.243

Figure 5: The complete results for the multi-document abstracts

For the 200- and 400-word extracts, a considerable improvement is made by simply using 50-word single summaries for clustering instead of 100-word summaries (see Figures 6, 7).[2]

	Mean	Our team	Best result	Worst result
200-word extracts starting with 100-word single-document summaries				
Mean F-measure	0.13210	0.102	0.211	0.042
200-word extracts starting with 50-word single-document summaries				
Mean F-measure	0.137	0.151	0.211	0.042
400-word extracts starting with 100-word single-document summaries				
Mean F-measure	0.198	0.179	0.290	0.097
400-word extracts starting with 50-word single-document summaries				
Mean F-measure	0.2063	0.262	0.290	0.097

Figure 6: The complete results for the multi-document extracts

[2] We thank Hans van Halteren for evaluating the extracts based upon the 50-word single summaries.

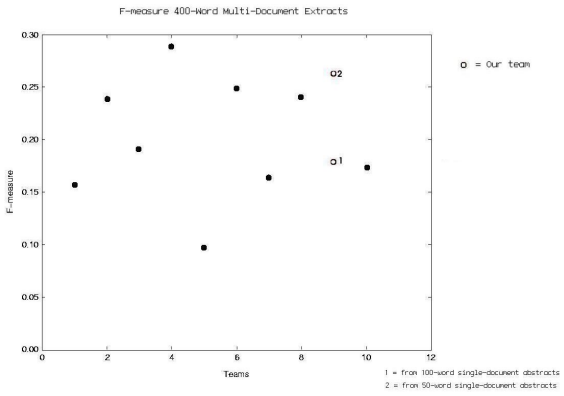

Figure 7: The F-measure for the 400-word multi-document extracts

In the near future we will evaluate the clustering of single document summaries for multi-document summarization by constructing ideal single summaries from 100-word human abstracts and by manually replacing sentences in these abstracts by the sentences from the original texts that best correspond to them.

Our abstracts still contain a lot of grammatical errors and incohesive passages, and have a rather sloppy organization (see Figures 2, 5, 8). Some of these errors can be attributed to the fact that we have largely neglected the preprocessing of the texts and postprocessing of the summaries. Altogether, given the fact that it is the first time that our research group participates in the DUC track and since we primarily focused on a few basic techniques that do not require a priori training, we are quite happy with the results.

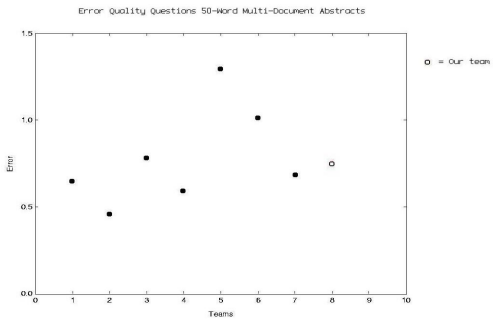

Figure 8: The mean quality questions errors for the 50-word multi-document abstracts

5 Future improvements

As far as the topic segmentation of single documents is concerned, we might improve the detection of sentence topics by considering a probabilistic approach. For the abstracts of single and multiple documents, the approach could be refined by condensing the sentences to their essential content without losing their grammatical well-formedness (e.g. especially in the case of direct speech). We will further investigate the effect of the removal of adjectives, adverbs and subclauses on the main propositional content of sentences. Also, the clustering might be refined by bringing back sentences to their more essential propositional content and by finding better cluster medoids. With regard to multi-document abstracts we want to look into matters of cohesion and more specifically into ways of improving the temporal order of the sentences.

6 Conclusion

Topic segmentation seems a valuable first step in automatic summarization, especially for summarizing expository text. It yields good summaries in the form of TOCs and acceptable summaries of single and multiple documents. The algorithms for topic segmentation and clustering the term vectors of sentences do not require prior training, which gives them the advantage of being generally applicable.

7 Acknowledgements

We thank Donna Harman, Paul Over and Hans Van Halteren for their help with the evaluation.

References

Barzilay R. & Elhadad M. (1999). *Using lexical chains for text summarization*. In "Advances in Automatic Text Summarization", I. Mani & M.T. Maybury, eds. MIT Press, Cambridge MA, pp. 111-121.

Dunning T. (1993). *Accurate methods for the statistics of surprise and coincidence*. In Computational Linguistics, 19, 61-74.

Givón T. (2001). *Syntax: Volume II,* John Benjamins, Amsterdam.

Hearst M.A. (1997). *TextTiling: segmenting text into multi-paragraph subtopic passages*. In Computational Linguistics 23(1), 33-64.

Moens, M.-F. & De Busser R. (2001). *Generic topic segmentation of document texts.* In "Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval", ACM, New York, pp. 418-419.

Moens, M.-F., Uyttendaele, C., & Dumortier, J. (1999). *Abstracting of legal cases: the potential of clustering based on the selection of representative objects*. In Journal of the American Society for Information Science 50 (2), 151-161.