# Generic Topic Segmentation of Document Texts.

**Conference Paper** · January 2001

DOI: 10.1145/383952.384065 · Source: DBLP



**2 authors:**

Marie-Francine Moens
KU Leuven
**324** PUBLICATIONS **5,060** CITATIONS
SEE PROFILE

Rik De Busser
National Chengchi University
**42** PUBLICATIONS **125** CITATIONS
SEE PROFILE

# Generic Topic Segmentation of Document Texts

Marie-Francine Moens
Katholieke Universiteit Leuven, Belgium
Interdisciplinary Centre for Law & IT
Tiensestraat 41
B-3000 Leuven, Belgium
xx/32/16/325383
marie-france.moens@law.kuleuven.ac.be

Rik De Busser
Katholieke Universiteit Leuven, Belgium
Interdisciplinary Centre for Law & IT
Tiensestraat 41
B-3000 Leuven, Belgium
tel: xx/32/16/325256
rik.debusser@law.kuleuven.ac.be

## ABSTRACT

Topic segmentation is an important initial step in many text-based tasks. A hierarchical representation of a text's topics is useful in retrieval and allows judging relevancy at different levels of detail. This short paper describes on-going research on generic algorithms for topic detection and segmentation that are applicable on texts of heterogeneous types and domains.

## Keywords

Topic detection; text indexing; summarization.

## 1. INTRODUCTION

Topic segmentation of texts concerns the detection of the overall organization of the text into themes or topics (of the thematic structure) and the identification of text segments that correspond to these general and more specific topics. Once the segments are found, they can be described by key terms, making the automatic generation of a detailed table of content of the text possible. The technique of topic segmentation is useful for retrieving documents (e.g. zoom in and out on the content). Furthermore, it is a practical initial step in information extraction, text summarization and question answering. It also is a valuable tool when linking the textual content of documents.

This short paper reports on research into topic segmentation algorithms that we recently started in the context of a project on generic technologies for information extraction from texts.

## 2. RESEARCH PROBLEM

The mental representation that humans make of the topics of a text – i.e. its macrostructure – gives them an insight in the global text topic (the kernel of the content) and its subtopics (more detailed meanings of the content). Although the thematic structure of a particular text often relies on unwritten guidelines for a specific text type, the author of a text has great freedom in choosing the thematic structure that best conveys his or her message. Document collections that contain different text types (e.g., on the World Wide Web) confront us with many instances of thematic organization. Among the most prominent thematic patterns in human discourse are: the hierarchical structuring of the main topic into subtopics, which in their turn can be divided recursively into smaller units; the progressive pattern that describes different concatenated topics; and the recurrent pattern or 'semantic return' by which a topic is suspended at one point and resumed later in the discourse. Real texts exhibit combinations of these patterns. For instance, one can tell a story in successive episodes, each of which is composed of subtopic sequences.

When we want to automatically detect the thematic structure, we will have to rely on surface markers of topicality and possibly on external knowledge sources. These topical cues can be text type dependent (e.g., lead paragraph in news story, conclusions in a scientific article), but there also exist many generic topical cues, amongst which the content terms of the texts.

Using these universal textual properties, we develop and test algorithms for detecting the thematic structure of a random text. The result is a topic segmentation of the text, indicating the more general and specific segments and their relationships (hierarchic, successive, semantic return).

## 3. METHODS

### 3.1 Content terms and their distribution

Content terms in the texts are detected by removing stopwords and recognition of proper names. The use of external knowledge sources renders the distribution of actual concepts more accurate. Synonyms can be detected with the help of a thesaurus. We also consider a shallow syntactic parse to detect better content terms and to resolve anaphors. Topically related terms can be detected by computing the correlations between terms based upon their occurrence in the same sentence.

### 3.2 Lexical chains

We construct lexical chains [1, 4] out of the selected content terms and their related terms. They represent the lexical cohesive


*SIGIR '01*, September 7-12, 2001, New Orleans, Louisiana, U.S.A.

structure of a text. A chain is built for each important content term, containing the term and its related terms (possibly including the resolved anaphors). In this process, the selection of relevant topical terms is of crucial importance. For each term of the chain, the position in the text in terms of the sentence number is recorded. The chains are stored as lists of terms and their position information. The terms are sorted in the reading order of the text. The information in the chains will be used for topic segmentation.

### 3.3 Topic segmentation

The aim is to compute the best possible segmentation of the text that takes into account a hierarchy of topics. The starts, interruptions, and terminations of lexical chains give valuable information on topic boundaries. However, we assume that the detected chains to a certain extent misrepresent the true segmentation. A lexical chain might overlap with another chain in text positions without being a subtopic of the latter. On the other hand, different chains can topically correlate. Some chains might also contain noise.

We implement and compare different algorithms for inferring the latent thematic patterns from the lexical chains. One is listed below.

We group chains with a cluster algorithm that in consequent steps merges two chains if the merge results in a better segmentation than the one obtained so far. For each pair of candidate chains the fit of the resulting segmentation is compared and the pair that produces the best fit is considered for the merging. The goodness or fit of a segmentation is computed as the combination of two factors: the total overlap of the chains based on their start and end position in the text, which must be minimal and does not take into account nested chains; and the total cohesiveness of the texts between the start and end positions of the chains computed, which must be maximal The latter is computed as a function of the length of the chain (length of the text between the start and end position of the chain) and number of members. A nested chain is a chain that, by looking at the text positions of its members, clearly represents a subpassage of the text represented by another chain.

This procedure allows filtering out part of the less significant chains and detecting nested topical segments. The hierarchy depth of segments can be computed from the number of topic segments it is nested in. The next step is to find the topic boundaries between consecutive segments that still have an overlap in text positions. Here, some heuristic cues might be handy: priority for paragraph boundary, taking the end position of the first segment in the sequence, etc. Semantic returns can be detected as large gaps in positions between two members of the lexical chain.

### 3.4 Test corpora

The techniques are being tested upon different test corpora. The kinds of texts we want to segment have a length up to 20 pages. The corpora include magazine articles, web pages and technical texts. Evaluation is difficult, since no evaluation criteria exist and segmentation by human evaluators proves to be highly unreliable. For this reason, we include texts in our test set that are already segmented by headings and subheadings, which are removed before processing.

## 4. RELATED RESEARCH

Topic segmentation was first studied by [5] who grouped paragraphs based on a threshold overlap of their content terms. Tuning the threshold allowed for more general or more specific topical groups. The most famous and useful algorithm for topic segmentation probably is TextTiling, which detects the boundaries of a text's subtopics [2]. The main topics of the opinions of court decisions were detected by [3], who applied a cluster algorithm based on the selection of representative objects. The last two approaches lack a hierarchical discrimination of the topics, which is useful if one wants to zoom in and out the content. Lexical chains are very useful to detect the main topics of a text by using the information of synonyms and related terms [1]. It is useful to add anaphoric referents to the chain. It is acknowledged that lexical chains could be exploited for hierarchical topic segmentation [6].

## 5. CONCLUSIONS

In this paper, we propose a method for a hierarchical topic segmentation of texts. We use lexical chains for representing the lexical cohesiveness of the texts and infer the thematic structure from them. We are currently testing the algorithms on different text corpora and refining the techniques in order to prove their generic character.

## 6. ACKNOWLEDGMENTS

We thank the *Vlaams Instituut voor de bevordering van het Wetenschappelijk-Technologisch onderzoek in de industrie* (*IWT*), Belgium for its research funding.

## 7. REFERENCES

[1] Barzilay, R. & Elhadad, M. (1999). Using lexical chains for text summarization. In *I. Mani & M.T. Maybury (Eds.), Advances in Automatic Text Summarization* (pp. 11-121). Cambridge, MA: MIT Press.

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

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

[4] Morris, J. & Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics* 17 (1), 21-43.

[5] Salton, G., Allan, J., Buckley, C., & Singhal, A. (1994). Automatic analysis, theme generation, and summarization of machine-readable texts. *Science*, 264, 1421-1426.

[6] Yaari, Y. (2000). NLP-assisted exploration of texts. In *Proceedings RIAO'2000 Content-Based Multimedia Information Access.*