Towards a Topic Driven Access to Full Text Documents

Caterina Caracciolo, Willem van Hage, and Maarten de Rijke

Informatics Institute, University of Amsterdam,
Kruislaan 403, 1098 SJ Amsterdam, The Netherlands
{caterina, wrvhage, mdr}@science.uva.nl

Abstract. We address the issue of providing a topic driven access to full text documents. The methodology we propose is a combination of topic segmentation and information retrieval techniques. By segmenting the text into topic driven segments, we obtain small and coherent documents that can be used as a basis for the automatic generation of links, and as a visualization aid for the reader who is presented with a focused and restricted text snippet. In presence of a concept hierarchy (ontology), the information retrieval step would connect the obtained segments to concepts in the ontology. In this paper we concentrate on the text segmentation phase: we describe our apporach, discuss some related issues and report on preliminary results.

1 Introduction

The full text documents accessible in a digital library can be rather long, potentially with a loose structure or no structure at all. In such a context, a search system that would provide the user with a document relevant for a given information need, and then leave up to the user to navigate within the document through a combination of "control F" and scrolling, would be very unsatisfactory.

We address the issue of providing focused access to the full text (scientific) documents in a digital environment, so as to enhance readability and minimize the browsing and scrolling effort. Specifically, we work in the setting of a collection of an electronic handbook consisting of "authoritative" (and usually lengthy) survey chapters. To provide access to the collection, a concept hierarchy (or "ontology") has been developed, consisting of concepts, and lexical relations (e.g., parent-child) and navigational relations (e.g., "see also") between those concepts. The concept hierarchy serves as a map of the handbook's domain, and, after browsing around the map, users jump from a concept to highly relevant text snippets, not complete chapters, in our collection.

We propose to use topic segmentation techniques as a way to subdivide documents into smaller documents (subdocuments), that are homogeneous in topic: used as link *targets*, these subdocuments will provide the reader with a notion of borders of the relevant document to read. As the *sources* of the links we use the concepts in our concept hierarchy.

In this paper we focus on the topic segmentation phase: in Section 2 we mention some previous work on topic segmentation and present our approach; in Section 3 we discuss the issue of evaluation for such a task and present current results. In Section 4 we draw preliminary conclusions and future work.

2 Topic segmentation

Previous work on text segmentation has focused on improving retrieval [1,2], or topic tracking of broadcast speech data [3]. The underlying theory of discourse can hypotesise that the text is linear [4], or hierarchical theory of text [5]. Skorochod'ko's seminal work [6] influences many approaches to topic segmentation, among others the Hearst's algorithm TextTiling [2] which we take as basis for our own work. According to Skorochod'ko's topologies [6], the overlap of words in sentences is an indicator of the semantic structure of the text.

Let us take a closer look at TextTiling. It performs a linear segmentation by using patterns of lexical connectivity (i.e., repetition of words through the text). The algorithm first compares adjacent *blocks* of text and assigns them a similarity value. The resulting sequence of similarity values is smoothed. Then the smoothed values are examined and each gap is given a score computed by averaging the difference between the smoothed similarity value at the gap and the peak to the left and to the right. The tunable parameter in the algorithm s the size of the blocks used for comparison: Hearst found that for newspaper corpora a block of 6 sentences (where a sentence is a sequence of n words, default is 20) is optimal. Similarity among two blocks is computed with a cosine similarity. Note that the actual value of similarity is not used for computing a breaks: the algorithm only looks at relative differences. We adopted Hearst algorithm because it looks at word repetition (which we considered appropriate to the writing style in a scientific domain). In our current implementation of TextTiling, blocks are real paragraphs.

Creation of a segmented corpus, manually annotated. Our work, and the experiments on which we report below, take place in the setting of a digital library project. Specifically, the *Logic and Language Links* (LoLaLi) project [7] explores methods to extend the traditional form of scientific handbooks with electronic tools These tools should help the reader explore the content of the handbook and to make it easier to locate relevant information. As a case study the project focuses on the *Handbook of Logic and Language* [8] (20 chapters, 1200 pages), and uses a WordNet-like concept hierarchy to provide access to (an electronic version of) the handbook [1]. For the work on which we report in this paper, we use the LaTeX sources of the book as our corpus, which amounted to about 4.5MB of text.

To develop a gold standard, to be used for assessing our segments, we selected two chapters from the collection of 20, and annotated the topic segmentation manually and form a reference system. The two chapters were chosen on the

[1] For more details see the project page: http://lolali.net

basis of the coverage in the LoLaLi concept hierarchy [9,7] and of the differences in style. Two annotators annotated the text independently, then discuss critical cases to agree on a unique annotation. The annotators were given indications about minimal and maximal size of a segment (a paragraph, an entire section). No other references to the layout structure of the text were made.

One of the two chapter had a rather formal style, with many tables, figures and formulas, either in-line or as separate objects. Here the difficulty laid in the treatment of those objects. The second chapter was written in a more narrative style, with fewer tables and pictures: here the annotators had difficulties with the rhetorical style of writing, because almost all paragraphs referred to previous paragraphs.

The annotators agreed on a large number of breaks, that we therefore consider more fundamental or evident than others. Within these breaks one of the two would mark more breaks, while the other would mark fewer breaks: following [10] we call the former type splitter and the latter type lumper. The annotation resulting from the confrontation of the two annotators can be taken as more splitter than lumper.

3 Evaluation issues

The evaluation of a topic segmentation system can be either task independent or task dependent. If task independent, the evaluation is done by comparing the result of the system against an annotated corpus, while a task independent evaluation would look at the how the segmentation improves other computational task. Here we concentrate on task independent evaluation, made on the basis of our manually annotated corpus. The most commonly used measures are precision and recall, applied to topic breaks or entire segments. Precision gives the proportion of hypotesised topic breaks (segments) that are correct, recall gives the proportion of correct topic breaks (segments) that are hypothesised. The two measure are often combined by using the F-measure.

When P and R are applied to paragraph breaks, they give a measure of how good the system is at recognising topic shifts; when applied to entire segments, they give a measure of how good the system is at recognising hogeneouity in topic. Precision and recall are quite crude measures in this context, because they do not give a measure of how distant the hypotesised segment break is from the real break. For this reason, Reyman introduces a range of tolerance, while [11] introduce the P_k precision measure, giving the probability that a randomly chosen pair of unit (for example paragraphs or sentences) be correctly classified. This measure is function of the distance between the elements of the pair.

We run some preliminary experiments to see how sensitive the used algorithm is to the different writing style, and as an evaluation measure we used precision and recall on the number of breaks. We found better values for recall than for precision, meaning that the segments tend to be shorter than those annotated by the annotators. Our current results are preliminary and mainly exploratory,

but they rise natural questions about the optimal size of a segment, from the point of view of a reader in an electronic environment.

4 Conclusions and future work

In this paper we reported on a work in progress about the application of text segmentation techniques in a digital library environment. The overall aim of our work is to apply these techniques for the automatic generation of links to full text documents, in particular we are interested in the application of these techniques for generation of links from ontologies to corpora of full text documents.

Future work will focus on a larger task independent evaluation of the text segmentation phase, possibly with the use of different segmentation algorithms. Then, we plan to use the documents resulting from the topic segmentation as target of a information retrieval system, where our queries will be concepts in the ontology we developed within the LoLaLi project [2].

We also plan to compare the semantic structure provided by our topic segmentation system with the layout structure of XML documents: we expect to learn useful lessons about the optimal size of a segment to return to a reader in an electronic environment.

Acknowledgements

The authors thank Joost Kircs and David Ahn for the interesting discussions.

References

1. Salton, G., Singhal, A.: Automatic text theme generation and the analysis of text structure. Technical Report 94–1483, Cornell Computer Science Technical Report (1994)
2. Hearst, M.A.: Context and Structure in Automated Full-text Information Access. PhD thesis (1994)
3. Ponte, J.M., Croft, W.B.: Text segmentation by topic. In: European Conference on Digital Libraries. (1997) 113–125
4. Min-Yen Kan, J.L.K., McKeown, K.R.: Linear segmentation and segment relevence. In: Proceedings of 6th International Workshop of Very Large Corpora (WVLC-6). (1998) 197–205
5. Yaari, Y.: Segmentation of expository text by hierarchical agglomerative clustering. In: Proceeding of the Conference on Recent Advances in Natural Language Processing. (1997) 59–65
6. Skorochod'ko, E.: Adaptive method of automatic abstracting and indexing. Information processing **71** (1092) 1179–1182
7. Caracciolo, C.: Towards modular access to electronic handbooks. JODI - Journal of Digital Information **3** (2003) http://jodi.ecs.soton.ac.uk/Articles/v03/i04/Caracciolo/.

[2] See http://lolali.net

8. van Benthem, J., ter Meulen, A., eds.: Handbook of Logic and Language. Elsevier (1997)
9. Caracciolo, C., de Rijke, M.: Structured access to scientific information. In: Proceeding of First Global WordNet Conference. (2002)
10. Klavans, J., McKeown, K., Kan, M., Lee, S.: Resources for the evaluation of summarization techniques. In: Proceedings of the 1st International Conference on Language Resources and Evaluation, Grenada, Spain. (1998)
11. Beeferman, D., Berger, A., Lafferty, J.D.: Statistical models for text segmentation. Machine Learning **34** (1999) 177–210