How to thematically segment texts by using lexical cohesion?

Olivier Ferret
LIMSI-CNRS
BP 133
F-91403 Orsay Cedex, FRANCE
ferret@limsi.fr

Abstract

This article outlines a quantitative method for segmenting texts into thematically coherent units. This method relies on a network of lexical collocations to compute the thematic coherence of the different parts of a text from the lexical cohesiveness of their words. We also present the results of an experiment about locating boundaries between a series of concatened texts.

1 Introduction

Several quantitative methods exist for thematically segmenting texts. Most of them are based on the following assumption: the thematic coherence of a text segment finds expression at the lexical level. Hearst (1997) and Nomoto and Nitta (1994) detect this coherence through patterns of lexical cooccurrence. Morris and Hirst (1991) and Kozima (1993) find topic boundaries in the texts by using lexical cohesion. The first methods are applied to texts, such as expository texts, whose vocabulary is often very specific. As a concept is always expressed by the same word, word repetitions are thematically significant in these texts. The use of lexical cohesion allows to bypass the problem set by texts, such as narratives, in which a concept is often expressed by different means. However, this second approach requires knowledge about the cohesion between words. Morris and Hirst (1991) extract this knowledge from a thesaurus. Kozima (1993) exploits a lexical network built from a machine readable dictionary (MRD).

This article presents a method for thematically segmenting texts by using knowledge about lexical cohesion that has been automatically built. This knowledge takes the form of a network of lexical collocations. We claim that this network is as suitable as a thesaurus or a MRD for segmenting texts. Moreover, building it for a specific domain or for another language is quick.

2 Method

The segmentation algorithm we propose includes two steps. First, a computation of the cohesion of the different parts of a text is done by using a collocation network. Second, we locate the major breaks in this cohesion to detect the thematic shifts and build segments.

2.1 The collocation network

Our collocation network has been built from 24 months of the French *Le Monde* newspaper. The size of this corpus is around 39 million words. The cohesion between words has been evaluated with the mutual information measure, as in (Church and Hanks, 1990). A large window, 20 words wide, was used to take into account the thematic links. The texts were pre-processed with the probabilistic POS tagger TreeTagger (Schmid, 1994) in order to keep only the lemmatized form of their content words, i.e. nouns, adjectives and verbs. The resulting network is composed of approximatively 31 thousand words and 14 million relations.

2.2 Computation of text cohesion

As in Kozima's work, a cohesion value is computed at each position of a window in a text (after pre-processing) from the words in this window. The collocation network is used for determining how close together these words are. We suppose that if the words of the window are strongly connected in the network, they belong to the same domain and so, the cohesion in this part of text is high. On the contrary, if they are not very much linked together, we assume that the words of the window belong to two different domains. It means that the window is located across the transition from one topic to another.

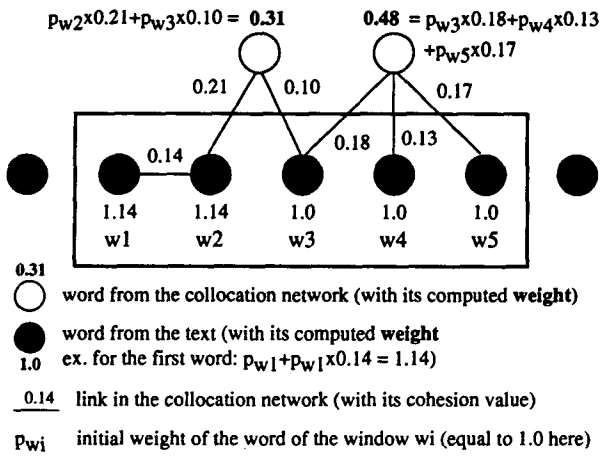

Figure 1: Computation of word weight

In practice, the cohesion inside the window is evaluated by the sum of the weights of the words in this window and the words selected from the collocation network common to at least two words of the window. Selecting words from the network linked to those of the texts makes explicit words related to the same topic as the topic referred by the words in the window and produces a more stable description of this topic when the window moves.

As shown in Figure 1, each word w (from the window or from the network) is weighted by the sum of the contributions of all the words of the window it is linked to. The contribution of such a word is equal to its number of occurrences in the window modulated by the cohesion measure associated to its link with w. Thus, the more the words belong to a same topic, the more they are linked together and the higher their weights are. Finally, the value of the cohesion for one position of the window is the result of the following weighted sum:

$$coh(p) = \sum_i sign(w_i) \cdot wght(w_i), \text{ with}$$

$wght(w_i)$, the resulting weight of the word w_i, $sign(w_i)$, the significance of w_i, i.e. the normalized information of w_i in the *Le Monde* corpus. Figure 2 shows the smoothed cohesion graph for ten texts of the experiment. Dotted lines are text boundaries (see 3.1).

2.3 Segmenting the cohesion graph

First, the graph is smoothed to more easily detect the main minima and maxima. This operation is done again by moving a window on the text. At each position, the cohesion associated to the window center is re-evaluated as the mean of all the cohesion values in the window.

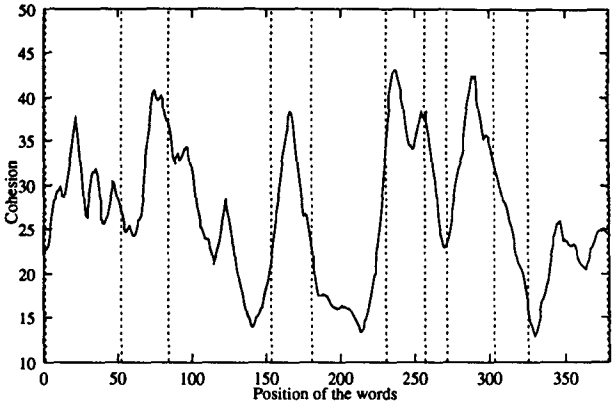

Figure 2: The cohesion graph of a series of texts

After this smoothing, the derivative of the graph is calculated to locate the maxima and the minima. We consider that a minimum marks a thematic shift. So, a segment is characterized by the following sequence: minimum - maximum - minimum. For making the delimitation of the segments more precise, they are stopped before the next (or the previous) minimum if there is a brutal break of the graph and after this, a very slow descent. This is done by detecting that the cohesion values fall under a given percentage of the maximum value.

3 Results

A first qualitative evaluation of the method has been done with about 20 texts but without a formal protocol as in (Hearst, 1997). The results of these tests are rather stable when parameters such as the size of the cohesion computing window or the size of the smoothing window are changed (from 9 to 21 words). Generally, the best results are obtained with a size of 19 words for the first window and 11 for the second one.

3.1 Discovering document breaks

In order to have a more objective evaluation, the method has been applied to the "classical" task of discovering boundaries between concatened texts. Results are shown in Table 1. As in (Hearst, 1997), boundaries found by the method are weighted and sorted in decreasing order. Document breaks are supposed to be the boundaries that have the highest weights. For the first N_b boundaries, N_t is the number of boundaries that match with document breaks. Precision is

N_b	N_t	Precision (P)	Recall (R)
10	5	0.5	0.13
20	10	0.5	0.26
30	17	0.58	0.45
38	19	0.5	0.5
40	20	0.5	0.53
50	24	0.48	0.63
60	26	0.43	0.68
67(N_b max)	26	0.39	0.68

Table 1: Results of the experiment

given by N_t/N_b and recall, by N_t/N, where N is the number of document breaks. Our evaluation has been performed with 39 texts coming from the *Le Monde* newspaper, but not taken from the corpus used for building the collocation network. Each text was 80 words long on average. Each boundary, which is a minimum of the cohesion graph, was weighted by the sum of the differences between its value and the values of the two maxima around it, as in (Hearst, 1997). The match between a boundary and a document break was accepted if the boundary was no further than 9 words (after pre-processing).

Globally, our results are not as good as Hearst's (with 44 texts; N_b: 10, P: 0.8, R: 0.19; N_b: 70, P: 0.59, R: 0.95). The first explanation for such a difference is the fact that the two methods do not apply to the same kind of texts. Hearst does not consider texts smaller than 10 sentences long. All the texts of this evaluation are under this limit. In fact, our method, as Kozima's, is more convenient for closely tracking thematic evolutions than for detecting the major thematic shifts. The second explanation for this difference is related to the way the document breaks are found, as shown by the precision values. When N_b increases, precision decreases as it generally does, but very slowly. The decrease actually becomes significant only when N_b becomes larger than N. It means that the weights associated to the boundaries are not very significant. We have validated this hypothesis by changing the weighting policy of the boundaries without having significant changes in the results.

One way for increasing the performance would be to take as text boundary not the position of a minimum in the cohesion graph but the nearest sentence boundary from this position.

4 Conclusion and future work

We have presented a method for segmenting texts into thematically coherent units that relies on a collocation network. This collocation network is used to compute a cohesion value for the different parts of a text. Segmentation is then done by analyzing the resulting cohesion graph. But such a numerical value is a rough characterization of the current topic.

For future work we will build a more precise representation of the current topic based on the words selected from the network. By computing a similarity measure between the representation of the current topic at one position of the window and this representation at a further one, it will be possible to determine how thematically far two parts of a text are. The minima of the measure will be used to detect the thematic shifts. This new method is closer to Hearst's than the one presented above but it relies on a collocation network for finding relations between two parts of a text instead of using the word recurrence.

References

K. W. Church and P. Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

M. A. Hearst. 1997. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

H. Kozima. 1993. Text segmentation based on similarity between words. In *31th Annual Meeting of the Association for Computational Linguistics (Student Session)*, pages 286–288.

J. Morris and G. Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1):21–48.

T. Nomoto and Y. Nitta. 1994. A grammatico-statistical approach to discourse partitioning. In *15th International Conference on Computational Linguistics (COLING)*, pages 1145–1150.

H. Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing.*