Thematic segmentation of texts: two methods for two kinds of texts

Olivier FERRET
LIMSI-CNRS
Bât. 508 - BP 133
F-91403, Orsay Cedex, France
ferret@limsi.fr

Brigitte GRAU
LIMSI-CNRS
Bât. 508 - BP 133
F-91403, Orsay Cedex, France
grau@limsi.fr

Nicolas MASSON
LIMSI-CNRS
Bât. 508 - BP 133
F-91403, Orsay Cedex, France
masson@limsi.fr

Abstract

To segment texts in thematic units, we present here how a basic principle relying on word distribution can be applied on different kind of texts. We start from an existing method well adapted for scientific texts, and we propose its adaptation to other kinds of texts by using semantic links between words. These relations are found in a lexical network, automatically built from a large corpus. We will compare their results and give criteria to choose the more suitable method according to text characteristics.

1. Introduction

Text segmentation according to a topical criterion is a useful process in many applications, such as text summarization or information extraction task. Approaches that address this problem can be classified in knowledge-based approaches or word-based approaches. Knowledge-based systems as Grosz and Sidner's (1986) require an extensive manual knowledge engineering effort to create the knowledge base (semantic network and/or frames) and this is only possible in very limited and well-known domains.

To overcome this limitation, and to process a large amount of texts, word-based approaches have been developed. Hearst (1997) and Masson (1995) make use of the word distribution in a text to find a thematic segmentation. These works are well adapted to technical or scientific texts characterized by a specific vocabulary. To process narrative or expository texts such as newspaper articles, Kozima's (1993) and Morris and Hirst's (1991) approaches are based on lexical cohesion computed from a lexical network. These methods depend on the presence of the text vocabulary inside their network. So, to avoid any restriction about domains in such kinds of texts, we present here a mixed method that augments Masson's system (1995), based on word distribution, by using knowledge represented by a lexical co-occurrence network automatically built from a corpus. By making some experiments with these two latter systems, we show that adding lexical knowledge is not sufficient on its own to have an all-purpose method, able to process either technical texts or narratives. We will then propose some solutions to choose the more suitable method.

2. Overview

In this paper, we propose to apply one and the same basic idea to find topic boundaries in texts, whatever kind they are, scientific/technical articles or newspaper articles. This main idea is to consider smallest textual units, here the paragraphs, and try to link them to adjacent similar units to create larger thematic units. Each unit is characterized by a set of descriptors, i.e. single and compound content words, defining a vector. Descriptor values are the number of occurrences of the words in the unit, modified by the word distribution in the text. Then, each successive units are compared through their descriptors to know if they refer to a same topic or not.

This kind of approach is well adapted to scientific articles, often characterized by domain technical term reiteration since there is often no synonym for such specific terms. But, we will show that it is less efficient on narratives. Although the same basic principle about word distribution applies, topics are not so easily detectable. In fact, narrative or expository texts often refer to a same entity with a large set of different words. Indeed, authors avoid repetitions and redundancies by using hyperonyms, synonyms and referentially equivalent expressions.

To deal with this specificity, we have developed another method that augments the first method by making use of information coming from a lexical co-occurrence network.

This network allows a mutual reinforcement of descriptors that are different but strongly related when occurring in the same unit. Moreover, it is also possible to create new descriptors for units in order to link units sharing semantically close words.

In the two methods, topic boundaries are detected by a standard distance measure between each pair of adjacent vectors. Thus, the segmentation process produces a text representation with thematic blocks including paragraphs about the same topic.

The two methods have been tested on different kinds of texts. We will discuss these results and give criteria to choose the more suitable method according to text characteristics.

3. Pre-processing of the texts

As we are interested in the thematic dimension of the texts, they have to be represented by their significant features from that point of view. So, we only hold for each text the lemmatized form of its nouns, verbs and adjectives. This has been done by combining existing tools. MtSeg from the Multext project presented in Véronis and Khouri (1995) is used for segmenting the raw texts. As compound nouns are less polysemous than single ones, we have added to MtSeg the ability to identify 2300 compound nouns. We have retained the most frequent compound nouns in 11 years of the French *Le Monde* newspaper. They have been collected with the INTEX tool of Silberztein (1994). The part of speech tagger TreeTagger of Schmid (1994) is applied to disambiguate the lexical category of the words and to provide their lemmatized form. The selection of the meaningful words, which do not include proper nouns and abbreviations, ends the pre-processing. This one is applied to the texts both for building the collocation network and for their thematic segmentation.

4. Building the collocation network

Our segmentation mechanism relies on semantic relations between words. In order to evaluate it, we have built a network of lexical collocations from a large corpus. Our corpus, whose size is around 39 million words, is made up of 24 months of the *Le Monde* newspaper taken from 1990 to 1994. The collocations have been calculated according to the method described in Church and Hanks (1990) by moving a window on the texts. The corpus was pre-processed as described above, which induces a 63% cut. The window in which the collocations have been collected is 20 words wide and takes into account the boundaries of the texts. Moreover, the collocations here are indifferent to order.

These three choices are motivated by our task point of view. We are interested in finding if two words belong to the same thematic domain. As a topic can be developed in a large textual unit, it requires a quite large window to detect these thematic relations. But the process must avoid jumping across the texts boundaries as two adjacent texts from the corpus are rarely related to a same domain. Lastly, the collocation w1-w2 is equivalent to the collocation w2-w1 as we only try to characterize a thematic relation between w1 and w2.

After filtering the non-significant collocations (collocations with less than 6 occurrences, which represent 2/3 of the whole), we obtain a network with approximately 31000 words and 14 million relations. The cohesion between two words is measured as in Church and Hanks (1990) by an estimation of the mutual information based on their collocation frequency. This value is normalized by the maximal mutual information with regard to the corpus, which is given by:

$$I_{\max} = \log_2 N^2 (S_w - 1)$$

with N: corpus size and S_W: window size

5. Thematic segmentation without lexical network

The first method, based on a numerical analysis of the vocabulary distribution in the text, is derived from the method described in Masson (1995).

A basic discourse unit, here a paragraph, is represented as a term vector $G_i = (g_{i1}, g_{i2}, \dots, g_{it})$ where g_i is the number of occurrences of a given descriptor in G_i.

The descriptors are the words extracted by the pre-processing of the current text. Term vectors are weighted. The weighting policy is *tf.idf* which is an indicator of the importance of a term according to its distribution in a text. It is defined by:

$$w_{ij} = tf_{ij} . \log \frac{N}{df_i}$$

where tf_{ij} is the number of occurrences of a descriptor T_j in a paragraph i; df_i is the number of paragraphs in which T_j occurs and

N the total number of paragraphs in the text. Terms that are scattered over the whole document are considered to be less important than those which are concentrated in particular paragraphs.

Terms that are not reiterated are considered as non significant to characterize the text topics. Thus, descriptors whose occurrence counts are below a threshold are removed. According to the length of the processed texts, the threshold is here three occurrences.

The topic boundaries are then detected by a standard distance measure between all pairs of adjacent paragraphs: first paragraph is compared to second paragraph, second one to third one and so on. The distance measure is the *Dice coefficient*, defined for two vectors $X= (x_1, x_2, \dots , x_t)$ and $Y= (y_1, y_2, \dots , y_t)$ by:

$$C(X,Y)=\frac{2\sum_{i=1}^{t} w(x_i)w(y_i)}{\sum_{i=1}^{t} w(x_i)^2+\sum_{i=1}^{t} w(y_i)^2}$$

where $w(x_i)$ is the number of occurrences of a descriptor x_i weighted by *tf.idf* factor

Low coherence values show a thematic shift in the text, whereas high coherence values show local thematic consistency.

6. Thematic segmentation with lexical network

Texts such as newspaper articles often refer to a same notion with a large set of different words linked by semantic or pragmatic relations. Thus, there is often no reiteration of terms representative of the text topics and the first method described before becomes less efficient. In this case, we modify the vector representation by adding information coming from the lexical network.

Modifications act on the vectorial representation of paragraphs by adding descriptors and modifying descriptor values. They aim at bringing together paragraphs which refer to the same topic and whose words are not reiterated. The main idea is that, if two words *A* and *B* are linked in the network, then " *when A is present in a text, B is also a little bit evoked, and vice versa* ".

That is to say that when two descriptors of a text *A* and *B* are linked with a weight *w* in the lexical network, their weights are reinforced into the paragraphs to which they simultaneously belong. Moreover, the missing descriptor is added in the paragraph if absent. In case of reinforcement, if the descriptor *A* is really present *k* times and *B* really present *n* times in a paragraph, then we add *wn* to the number of *A* occurrences and *wk* to the number of *B* occurrences. In case of descriptor addition, the descriptor weight is set to the number of occurrences of the linked descriptor multiplied by *w*. All the couples of text descriptors are processed using the original number of their occurrences to compute modified vector values.

These vector modifications favor emergence of significant descriptors. If a set of words belonging to neighboring paragraphs are linked each other, then they are mutually reinforced and tend to bring these paragraphs nearer. If there is no mutual reinforcement, the vector modifications are not significant.

These modifications are computed before applying a *tf.idf* like factor to the vector terms. The descriptor addition may add many descriptors in all the text paragraphs because of the numerous links, even weak, between words in the network. Thus, the effect of *tf.idf* is smoothed by the standard-deviation of the current descriptor distribution. The resulting factor is:

$$\log(\frac{N}{dfj}\cdot(1+\frac{\sqrt{\sum_{k}(tfjk-\overline{tfj})^2}}{dfj}))$$

with *k*, the paragraphs where T_j occurs.

7. Experiments and discussion

We have tested the two methods presented above on several kinds of texts.

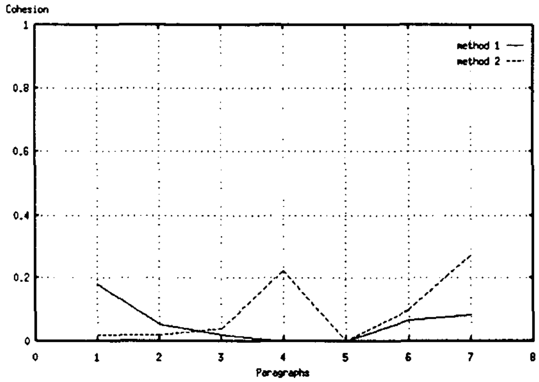

Figure 1 - Improvement by the second method with low word reiteration

Figure 1 shows the results for a newspaper article from *Le Monde* made of 8 paragraphs. The cohesion value associated to a paragraph i indicates the cohesion between paragraphs i and $i+1$. The graph for the first method is rather flat, with low values, which would a priori mean that a thematic shift would occur after each paragraph. But significant words in this article are not repeated a lot although the paper is rather thematically homogeneous. The second method, by the means of the links between the text words in the collocation network, is able to find the actual topic similarity between paragraphs 4 and 5 or 7 and 8.

The improvement resulting from the use of lexical cohesion also consists in separating paragraphs that would be set together by the only word reiteration criterion. It is illustrated in Figure 2 for a passage of a book by Jules Verne[1]. A strong link is found by the first method between paragraphs 3 and 4 although it is not thematically justified. This situation occurs when too few words are left by the low frequency word and *tf.idf* filters.

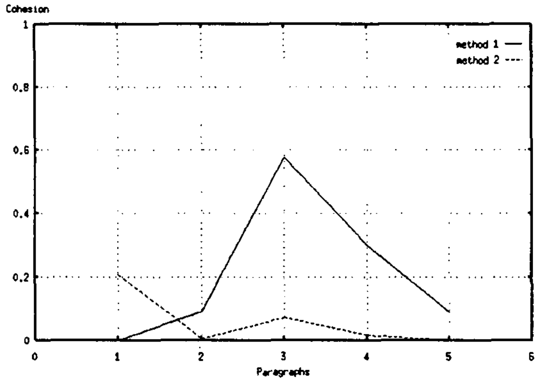

Figure 2 - Improvement by the second method when too many words are filtered

More generally, the second method, even if it has not so impressive an effect as in Figures 1 and 2, allows to refine the results of the first method by proceeding with more significant words. Several tests have been made on newspaper articles that show this tendency.

Experiments with scientific texts have also been made. These texts use specific reiterated vocabulary (technical terms). By applying the first method, significant results are obtained because of this specificity (see Figure 3, the coherence graph in solid line).

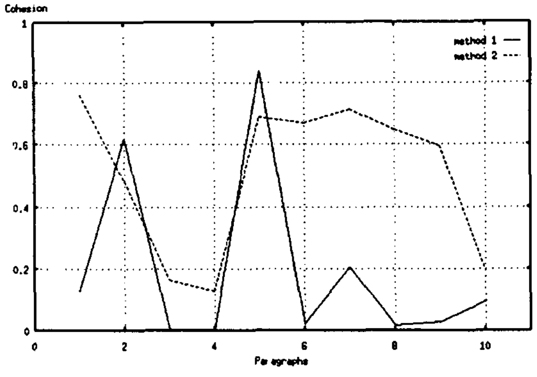

Figure 3 - Test on a scientific paper[2] in a specialized domain

On the contrary, by applying the second method to the same text, poor results are sometimes observed (see Figure 3, the coherence graph in dash line). This is due to the absence of highly specific descriptors, used for *Dice coefficient* computation, in the lexical network. It means that descriptors reinforced or added are not really specific of the text domain and are nothing but noise in this case.

The two methods have been tested on 16 texts including 5 scientific articles and 11 expository or narrative texts. They have been chosen according to their vocabulary specificity, their size (between 1 to 3 pages) and their paragraphs size. Globally, the second method gives better results than the first one: it modulates some cohesion values. But the second method cannot always be applied because problems arise on some scientific papers due to the lack of important specialized descriptors in the network. As the network is built from the recurrence of collocations between words, such words, even belonging to the training corpus, would be too scarce to be retained. So, specialized vocabulary will always be missing in the network. This observation has lead us to define the following process to choose the more suitable method:

```
Apply method 1;
If $x$% of the descriptors whose value is not
null after the application of tf.idf are not
found in the network,
        then continue with method 1
        otherwise apply method 2.
```

According to our actual studies, x has been settled to 25.

[1] De la Terre à la Lune.

[2] Le vin jaune, Pour la science (French edition of Scientific American), October 1994, p. 18

8. Related works

Without taking into account the collocation network, the methods described above rely on the same principles as Hearst (1997) and Nomoto and Nitta (1994). Although Hearst considers that paragraph breaks are sometimes invoked only for lightening the physical appearance of texts, we have chosen paragraphs as basic units because they are more natural thematic units than somewhat arbitrary sets of words. We assume that paragraph breaks that indicate topic changes are always present in texts. Those which are set for visual reasons are added between them and the segmentation algorithm is able to join them again. Of course, the size of actual paragraphs are sometimes irregular. So their comparison result is less reliable. But the collocation network in the second method tends to solve this problem by homogenizing the paragraph representation.

As in Kozima (1993), the second method exploits lexical cohesion to segment texts, but in a different way. Kozima's approach relies on computing the lexical cohesiveness of a window of words by spreading activation into a lexical network built from a dictionary. We think that this complex method is specially suitable for segmenting small parts of text but not large texts. First, it is too expensive and second, it is too precise to clearly show the major thematic shifts. In fact, Kozima's method and ours do not take place at the same granularity level and so, are complementary.

9. Conclusion

From a first method that considers paragraphs as basic units and computes a similarity measure between adjacent paragraphs for building larger thematic units, we have developed a second method on the same principles, making use of a lexical collocation network to augment the vectorial representation of the paragraphs. We have shown that this second method, if well adapted for processing such texts as newspapers articles, has less good results on scientific texts, because the characteristic terms do not emerge as well as in the first method, due to the addition of related words. So, in order to build a text segmentation system independent of the kind of processed text, we have proposed to make a shallow analysis of the text characteristics to apply the suitable method.

10. References

Kenneth W. Church and Patrick Hanks. (1990)*Word Association Norms, Mutual Information, And Lexicography*. Computational Linguistics, 16/1, pp. 22—29.

Barbara J. Grosz and Candace L. Sidner. (1986) *Attention, Intentions and the Structure of Discourse*. Computational Linguistics, 12, pp. 175—204.

Marti A. Hearst. (1997) *TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages*. Computational Linguistics, 23/1, pp. 33—64.

Hideki Kozima. (1993) *Text Segmentation Based on Similarity between Words*. In Proceedings of the 31th Annual Meeting of the Association for Computational Linguistics (Student Session), Colombus, Ohio, USA.

Nicolas Masson. (1995) *An Automatic Method for Document Structuring*. In Proceedings of the 18th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, USA.

Jane Morris and Graeme Hirst. (1991) *Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text*. Computational Linguistics, 17/1, pp. 21—48.

Tadashi Nomoto and Yoshihiko Nitta. (1994) *A Grammatico-Statistical Approach To Discourse Partitioning*. In Proceedings of the 15th International Conference on Computational Linguistics (COLING), Kyoto, Japan.

Helmut Schmid. (1994) *Probabilistic Part-of-Speech Tagging Using Decision Trees*. In Proceedings of the International Conference on New Methods in Language Processing, Manchester, UK.

Max D. Silberztein. (1994) *INTEX: A Corpus Processing System*. In Proceedings of the 15th International Conference on Computational Linguistics (COLING), Kyoto, Japan.

Jean Véronis and Liliane Khouri. (1995) *Étiquetage grammatical multilingue: le projet MULTEXT*. TAL, 36/1-2, pp. 233—248.