Text Segmentation based on Document Understanding for Information Retrieval

Violaine Prince, Alexandre Labadié

Text segmentation based on document understanding for information retrieval

Violaine Prince and Alexandre Labadié

LIRMM,
161 rue Ada 34392 Montpellier Cedex 5, France
{prince,alexandre.labadie}@lirmm.fr

Abstract. Information retrieval needs to match relevant texts with a given query. Selecting appropriate parts is useful when documents are long, and only portions are interesting to the user. In this paper, we describe a method that extensively uses natural language techniques for text segmentation based on topic change detection. The method requires a NLP-parser and a semantic representation in Roget-based vectors. We have run the experiment on French documents, for which we have the appropriate tools, but the method could be transposed to any other language with the same requirements. The article sketches an overview of the NL understanding environment functionalities, and the algorithms related to our text segmentation method. An experiment in text segmentation is also presented and its result in an information retrieval task is shown.

1 Introduction

Information retrieval needs to match relevant texts with a given query. The latter is seldom expressed as a sentence, but most frequently as a set of key-words, all in natural language (NL). If several research works have been dealing with the problem of matching the query content with available documents (on the Web for instance), the issue we are here focusing on is how to provide the user, not only with the relevant document, but with the most appropriate fragments of this document relevant to his/her queries.
Selecting appropriate parts is useful when documents are long, and only portions are interesting to an user. This approach has already been proned by ([22], [8], [4], [14]) and many other works. Two major techniques are to be applied :

- Looking for the fragment that contains the biggest set of words of the query, and selecting the n sentences containing it ([22],[15]). It provides the involved segments, but could be silent about portions, semantically related to the query as consequences or causes, that do not directly contain the query keywords.
- Segmenting the retrieved text into parts that are topically based, and matching the query content with these segments ([8]) being one of the first to suggest it.

In this paper, we describe a method belonging to the second category. Its advantage is that it extensively uses NL techniques for text segmentation based on topic change detection. This means that it undertakes a task of document understanding. Its advantage in information retrieval is that it might select as relevant text fragments semantically and topically related to the query, about which word-based methods are silent.
Since the method needs an NL-parser and a semantic representation in Roget-based vectors (a first major use of Roget-based representations in NL processing is described in ([21]), we have run the experiment on French documents, for which we have the required NL-environment. Transposition for English or other languages could be made with the appropriate parsers.
In section 2 we describe text segmentation as an issue, briefly browsing its state-of-the-art, related to information retrieval, and present the grounds on which our method is founded. In section 3 we provide an overview of our NL understanding environment functionalities, and the algorithms related to our text segmentation method. In section 4 we describe an experiment in text segmentation, and show its output to queries. Finaly we conclude about the accomplished research and its possible extensions for information retrieval.

2 Topical Text segmentation

2.1 What is Text Segmentation

Topic based text segmentation consists in finding, inside a text, sentences that will be borderlines of topical segments. There are three main approaches to detect these sentences :

- Similarity based methods, which measure proximity between sentences by using (most of the time) the cosine of the angle between vectors representing sentences. The c99 algorithm ([1]) for example uses a similarity matrix to generate a local classification of sentences and isolate topical segments.
- Graphical methods, which graphically represent terms frequencies and use these representations to identify topical segments (which are dense dot clouds on the graphic). The Dotplotting algorithm ([19]) is the most common example of the use of a graphical approach of text segmentation.
- Lexical chains based methods, which links multiple occurrences of a term and consider a chain is broken when there are too much sentences between two occurrences of a term. Segmenter ([12]) uses this methods for text segmentation with a subtle adjustment as it determine the number of necessary sentences to break a chain in function of the syntactical category of the term.

These methods are all word / term based, and so view the text as a "bag of words". If they can help retrieving relevant segment of text in big documents, they cannot solve the problem of relevant segments not using the same lexical field than the query.

2.2 Text-segmentation Based Approaches in Information Retrieval

Since text segmentation could be associated to the fact that segments could be named and indexed, it was an evidence that text segmentation was a requirement for information retrieval. However, a great majority of the recent literature is devoted to word segmentation as a major issue and not to topical fragments retrieval. Also, most recent papers are related to languages like Chinese where word segmentation is a real ambiguous problem ([9] is, in this respect, one of the most cited papers in the domain). Among this literature, [20] suggest to detect indexing segments in Chinese texts, with a heuristic based method that outperformed the boundary method previously used (a method close to lexical chains).However, their method is limited to words, and does not undertake a complete document understanding. Topic change detection methods applied to information retrieval are present in many works inspired from NL processing, mostly in the preceding decade. [18] describe a topic-based text segmentation. [7] suggests a text tiling algorithm detecting subtopics of a given topic. In the same year, [10] insist on redefining segments retrieval methods. Nevertheless, all these methods are lexically based, either on lexical cohesion determination ([16]) or on lexical chains delimitation. The few methods that enlarge the horizon of topic change detection towards discourse function (style, syntax, etc.) are found in ([11]) for stylistic variation. The latter is also used in multimedia information retrieval especially with speech and speaker recognition. More oriented towards syntax, [17] describe a grammar-based method for discourse partitioning.
One of the limitations to be found with the most popular methods in topic change detection is that although lexical cohesion, as a ground assumption, is a good candidate for defining a topic it has the following drawbacks:

- Most words of any natural language are polysemous. Their multiple meanings are only disambiguated by understanding the sentences they are in, because sentences are a natural way to select a word meaning for a human reader/speaker. So a representation of the word as modulated by the sentence is necessary to constrain word sense disambiguation, a thing that word-based methods tend to neglect.
- A text segment might be related to a topic by directly using the words that are prototypical of this topic. Describing the consequences of an action might not necessarily contain the action name. So word-based methods overlook these segments in their passages retrieval.

2.3 Requirements for an Adapted Text Segmentation

Text segmentation, to be useful in an information retrieval task, doesn't need to have very precise topical boundaries, so boundary based methods such as [19] are not necessarily the most adapted. Note that is was also a result found by [20] for their word segmentation for information retrieval. The rationale is that one or two sentences of margin won't significantly affect information retrieval

performances, from the user point of view. But to really improve results on this kind of task, a text segmentation method should :

- Represent text segments in a simple and lexically independent way : lexical dependency might burden information retrieval with side effects such as polysemy (introducing noise) or synonymy (introducing silence). The previous subsection discussions show that representations of syntax (for the sentence) and discourse relations are necessary to retrieve the best segments (note that [4] reaches a similar conclusion for enriching the idea of local coherence).
- Allow to match topically close segments together: matching methods are several; they could be by measuring length (on chains), similarity (on vectors), or distance (in bayesian networks in clustering methods).
- Allow to match queries with text segments.

It, indeed, also needs to find cuts between segments, but, as it is said before, fuzzy boundaries should work as well as precise ones.
In the next section we present a tool based on NL processing. It detects topic coherence by using a deep syntactic analysis employed as an input for semantic calculus of the sentence. The local topics of the sentence are thus determined as related to concepts defined in a thesaural ontology. Afterwards, each sentence is agglutinated to the preceding and a new calculus is performed. Topic change is detected when a new sentence (or a new bundle of sentences) strongly differs from the preceding one. Therefore, document understanding and segments topic comparison are performed by the environment and method described hereafter.

3 A natural language environnement for topic change detection

3.1 A NL Parser Providing Syntactic and Semantic Text Analysis

The NL environment we use is composed of a parser that provides constituents and dependencies in the sentence. Constituents are words that have a part-of-speech atomic tag such as Noun, Verb, Adjective, and so forth, but also sets of words labelled with compound tags such as Noun Phrase, Verb Phrase, Prepositional Noun Phrase, etc. Dependencies are relations between constituents that determine semantic and syntactic functions in the sentence. Subject, Object, Complement are the basic dependencies. They tend to express a notion of government (defined by Chomsky) thus showing that constituents are not equal in importance as semantic elements in a given sentence. The impact of dependencies on defining the general semantics of a sentence is great, and might influence the relevance of this sentence to a given topic (and onwards, to a given query). For instance, if the word "doctor" belongs to a query and appears in a given sentence of a given text as a very secondary complement, the semantic impact of this word on the sentence meaning is weak. Therefore, retrieving this sentence as a core for a relevant text segment would be introducing noise. Whereas if the

word is central and governor (like a subject, or sometimes an object), then the sentence containing it could be seen as an interesting candidate for relevance to the query.

The Parser in a Nutshell The parser we use is called SYGFRAN([3]) and works with 12,000 rules written with a Markov's algorithm formalism. Its characteristics, calculated on a French corpus of 300,000 sentences of an average of 25 words each are given in table 1. SYGFRAN guaranties a 34% precision in complete sentence analysis for any corpus (it has been run on several different corpora and the ratio does not change). However in all other cases, SYGFRAN is not silent : it provides a recall of 85% in partial sentence analysis. Both measures are intimately related to dependencies detection measures.

	Recall	Precision
Constituents detection (atomic and compound)	100%	97%
Dependencies detection	85%	34%
Partial sentence analysis	85%	85%
Complete Sentences deep and surface analysis	34%	34%

Table 1. SYGFRAN parser measures

Semantic calculus This parser calculates a semantic representation of the sentence based on a vector representation. Vectors are inspired from the Roget approach in NLP, which has already been proved as interesting for lexical semantics ([21]) and corpus based research ([5]). All words of the language are represented in a dictionary as vectors in a space of a fixed dimension of 873 for French (1000 for English). The basic 873 (respectively 1000) are organized as a conceptual ontology defined in ([24]) and every word is indexed by one or many elements of this ontology.

The technique for calculating each sentence vector is based on calculating each constituent vector, and then using dependencies to define the impact of each constituent on the sentence meaning ([2]). Once the impact defined, the sentence vector is calculated by linear combination of constituent vectors. The more the constituent has impact on the sentence meaning, the more weight it will have in the linear combination (for example, verbs will be more important than adjectives).

A segment vector is defined as a centroid of the sentence vectors it contains (centroids are already used in the domain in ([6]) : among centroid possibilities, we choose to represent segment vectors by a barycenter. But we apply different weights on sentence vectors depending on the position of the sentence

represented in the segment. First sentences of the segment have great weights, and weights decrease progressively as we advance in the segment. In a classical structure of an argumentative paragraph, first sentences carry the main subject (topic) of the paragraph and as we advance, we encounter examples and explanations. Finally, a "good" argumentative paragraph ends with some transition sentences which conclude the current paragraph and introduce the next one. We supposed that a topical segment should have the same structure.

Thematic Distance Most methods using a vectorial representation of text use the cosine as a similarity measure. We preferred to use the angular distance, which we call thematic distance in this case, to compare sentences or text segments. So the thematic distance between X and Y should be :

$$D_A(X,Y) = \arccos \frac{\boldsymbol{X}.\boldsymbol{Y}}{|\boldsymbol{X}|.|\boldsymbol{Y}|} \tag{1}$$

Where X and Y are vectorial representation of sentences. This measure seems better to us for two main reasons :

- It is a mathematical distance. So we can use it more freely than the cosine that is, most of the time, wrongly considered as probability.
- It is a decreasing function which is strongly non-linear between $\frac{\pi}{4}$ and 0. This property is very interesting in our case, because it allows us to be more precise when vectors are close.

The thematic distance will help us finding text segments and frontiers during the segmentation process, but also identifying relevant segments of text.

3.2 Defining Topical Text Segmentation in this Environment

Our segmentation method use the thematic distance and the vectorial representation of segments and sentence to identify what we call "transition zones". We made the hypothesis that topics' boundaries aren't, most of the time, standalone sentences, but small group of sentences concluding the previous topic and introducing the next one. So we can represent two successive topic segment as in Figure 1.

Fig. 1. Topical Structure in the sentence s stream.

To detect transition zones, we use a window which slides along the text and gives to the sentence in the middle of the window a value called transition value.

This value is calculated by considering the first half of the window (current sentence excluded) as a topical segment and the second half as well (current sentence included). We calculate the centroid of each supposed segment and then the thematic distance between them. This distance becomes the transition value of the current sentence.
The transition value of each sentence is compared to a threshold value that has been learnt on three different thematic corpora (law, computer science, and political discourses) of respectively 433456, 4722 and 303373 sentences provided in the evaluation campaign DEFT 2006, as proposed by [13]. What these authors seemed to hint at is that the threshold behaves a sort of a constant in text building. For a given domain, topical units tend to have a more or less fixed amount of sentences. If more, they tend to split into sub-topics. If the transition value is higher than the threshold, we consider this sentence to be a candidate for a transition. If two or more consecutive sentences are higher than the threshold value then we have a transition zone. Transitions zones, and their computing are illustrated in Figures 2 and 3.
Finding the right boundary sentence in the transition zone can be done by many means, the simplest (and the one we used here) is to select the sentence with the highest transition value. Other methods are currently experimented.

Fig. 2. Computing Transition Zones for Topic Change Detection, 1st step

Fig. 3. Computing Transition Zones for Topic Change Detection, 2nd step

4 Experimenting on a corpus with queres

During the segmentation task, the centroid of each identified segment is saved. Finding the right segments only means comparing the semantic vector of the query with the centroid of each segment using the thematic distance and the threshold which has been learnt on the corpora described in subsection 3.2. To evaluate the capabilities of our approach as a question-answer system, we have used a corpus of about 15,000 sentences, with an average of 27 words per sentence. Its domain is law texts, and it served as a test corpus for the DEFT06 evaluation conference on text segmentation. Our questions were the following: 1. What are the official languages in Europe? 2. What is the regulation concerning employment in the nuclear industry? 3. Which are the rules of formation of a limited company? 4. What is the regulation concerning the marketing of medical drugs?

4.1 The Method

Each question, given as one or more sentences in natural language, was projected in the vector space and its semantic vector calculated (see section 3). The corpus is already segmented, before entering the query-answer evaluation, and independently from the query contents. The idea is to compare the semantic vector of the query with all detected segments (with a transition of 2 sentences), and retrieve those segments whose angular distance with the query does not exceed 0.8. This value roughly corresponds to an angle of 45 ($\frac{\pi}{4}$) or less, which is half the maximum possible angular distance according to the formula given before. If two vectors make an angle of 45 and less, they are considered to be relatively close to each other. Transposed as a relationship between query and fragments, this means that the fragment is (semantically, topically) relevant to the query. The closer to 0 the angle is, the more relevant the fragment is.

4.2 The Results

Obtained results are summarized in table 2. The segmentation evaluation was made by the organizers of DEFT06 competition, so we just reproduce the values relative to the law test corpus (two other corpora of different domains were provided). The evaluation of segments relevance to queries was made by another group of persons. The idea was the following: a segment was considered as lacking if the human jury considered this segment as relevant to a query and not provided by the system (this played on the recall percentages). The segment was considered as totally relevant and scored 1 in the total if it was a very close or exact answer to the question. It was considered as partially relevant, and score 0.5 in the total if it was sufficiently related to the query to be seen as "interesting". Both values affected the precision percentages.

	Recall	Precision
Segmentation results	**0.806**	0.164
Question 1	0.666	0.518
Question 2	0.16	1
Question 3	0.96	0.36
Question 4	0.29	0.18

Table 2. Segmentation and Query-answer results

5 Conclusion

First obtained results are encouraging. The advantages of this approach could be listed as follows: (1)a query could be a big fragment or a small one, a question or

a text, this doesn?t temper with the fragment retrieval method. (2) Fragments containing other words than those in the query have been retrieved. They were judged partially and sometimes totally relevant by the human jury. With a word-based method, they would have been discarded. (3) Small fragments have been retrieved, which is much easier to read for a human user, and the ?informative power? of these fragments is higher than a big text into which the relevant part is littered with irrelevant segments.
Moreover, numerical results don't show some interesting links established during the process. The method, sometimes, bing back sentences and segment which aren't "officially" answer to the question, but that make sense.

References

1. F. Y. Y. Choi, Advances in domain independent linear text segmentation. Proc. NAACL-00 2000 p. 26-33
2. J. Chauché and V. Prince and S. Jaillet and M. Teisseire, Classification Automatique de Textes partir de leur Analyse Syntaxico-Smantique Proc. 12th International Conference on Natural Language Processing (TALN) 2003 p. 55-65
3. J. Chauché, Un outil multidimensionnel de l'analyse du discours. Proc. Coling'84. 1984 p. 11-15
4. S. W. K. Chan, Using heterogeneous linguistic knowledge in local coherence identification for information retrieval Journal of Information Science 2000 p. 313-328
5. J. Ellman and J. Tait, Roget's thesaurus: An additional Knowledge Source for Textual CBR? Proc. 19th SGES Int. Conf. on Knowledge-Based and Applied AI. 1999 p. 204-217
6. H. Eui-Hong and G. Karypis, Centroid-Based Document Classification: Analysis and Experimental Results. Proc. PKDD 2000 p. 424-431
7. M. A. Hearst, TextTiling: Segmenting text into multi-paragraph subtopic passages. Computational Linguistics. 1997 23 p. 33-64
8. M. A. Hearst and C. Plaunt, Subtopicstructuring for full-length document access. Proc. ACM SIGIR-93. 1993 p. 59-68
9. X. Huang and F. Peng and D. Schuurmans and N. Cercone and S.E. Robertson, Applying Machine Learning to Text Segmentation for Information Retrieval. Information Retrieval 2003
10. M. Kaszkiel and J. Zobel, Passage retrieval revisited. Proc. Twentieth International Conference on Research and Development in Information Access. 1997 p. 178-185
11. J. Karlgren, Stylistic variation in an information retrieval experiment. Proc. NeMLaP-2 Conference 1996
12. M. Kan and J. L. Klavans and K. R. McKeown, Linear segmentation and segment significance. Proceedings of WVLC-6 1998
13. A. Labadié and J. Chauché, Segmentation thmatique par calcul de distance smantique. Proc.DEFT06 2006
14. F. Llopisand and A. Ferrandezand and J. L. Vicedoand and A. Gelbukh, Textsegmentation for efficient information retrieval. Proc. CICLing. 2002 p. 373-380
15. A. Moffat and R. Sacks-Davis and R. Wilkinson and J. Zobel, Retrieval of partial documents. Proc. of the Second Text Retrieval Conference TREC-2. 1994 p. 181-190
16. J. Morris and G. Hirst, Lexical cohesion computed by thesaural relations as anindicator of the structure of text. Computational Linguistics. 1991 p. 21-48

17. T. Nomoto and N. Yoshihiko, A grammatico-statistical approach to discoursepartitioning. Proc. COLING'94 1994 p. 1145-1150
18. J. Ponte and B. Croft, Text segmentation by topic. Proc. First EuropeanConference on Research and Advanced Technology for Digitial Libraries. 1997 p. 1145-1150
19. J. C. Reynar, Topic Segmentation: Algorithms and Applications. Phd thesis, University of Pennsylvania 1998
20. C. C. Yang and K. W. Li, A heuristic method based on a statistical approach for chinese text segmentation. Journal of the American Society for Information Science and Technology 2005 p. 438-1447
21. D. Yarowsky, Word Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora. Proc. Coling'92 1992 p. 454-460
22. G. Salton, Automatic text processing: the transformation, analysis, and retrieval of information by computer. Addison-Wesley Longman Publishing Co. Boston, MA, USA 1989
23. P. Roget, Thesaurus of English Words and Phrases. Longman London, England 1992
24. Larousse, Thesaurus Larousse Larousse Paris, France 1992