Discovering Topic Boundaries for Text Summarization Based on Word Co-occurrence

Gaël Dias* and Elsa Alves*†

*HULTIG, Department of Computer Science, University of Beira Interior, Covilhã, Portugal
ddg@di.ubi.pt
†GLINT, Department of Computer Science, New University of Lisbon, Lisbon, Portugal
elsalves@zmail.pt

Abstract

Topic Segmentation is the task of breaking documents into topically coherent multi-paragraph subparts. In particular, Topic Segmentation is extensively used in Text Summarization to provide more coherent results by taking into account raw document structure. However, most methodologies are based on lexical repetition that show evident reliability problems or rely on harvesting linguistic resources that are usually available only for dominating languages and do not apply to less favored and emerging languages. In order to tackle these drawbacks, we present an innovative Topic Segmentation system based on a new informative similarity measure based on word co-occurrences and evaluate it on a set of web documents belonging to a single domain.

1. Introduction

This paper introduces a new technique for improving access to information dividing lengthy documents into topically coherent sections. This research area is commonly called Topic Segmentation and can be defined as the task of breaking documents into topically coherent multi-paragraph subparts.
Topic Segmentation has extensively been used in Text Summarization where it serves as the basic text structure in order to apply sentence extraction and sentence compression techniques (Boguraev and Neff, 2000; Angheluta *et al.*, 2002; Farzindar and Lapalme, 2004). In this paper, we present an innovative Topic Segmentation system based on a new informative similarity measure that takes into account word co-occurrence in order to avoid the accessibility to existing linguistic resources such as electronic dictionaries or lexico-semantic databases.
In particular, our architecture solves three main problems evidenced by previous research. First, systems based uniquely on lexical repetition show reliability problems (Hearst, 1994; Reynar, 1994; Sardinha, 2002) as common writing rules prevent from using lexical repetition. Second, systems based on lexical cohesion, using existing linguistic resources that are usually only available for dominating languages like English, French or German, do not apply to less favored and emerging languages (Morris and Hirst, 1991; Kozima, 1993). Third, systems that need previously existing harvesting training data (Beeferman *et al.*, 1997) do not adapt easily to new domains as training data is usually difficult to find or build depending on the domain being tackled. Instead, our architecture proposes a language-independent unsupervised solution, similar to (Phillips, 1985; Ponte and Croft, 1997), defending that Topic Segmentation should be done "on the fly" on any text thus avoiding the problems of domain, genre, or language-dependent systems.
In order to show the results of our system in real-world conditions, we propose an evaluation on a set of web documents belonging to a single domain unlike other methodologies that have been evaluated on (Choi, 2000)'s data set that relies on small texts of different domains within which lexical repetition is high. It is clear that this situation does not correspond to real-world conditions for Text summarization as documents to segment are usually from a same domain and do not use repetition.
This paper is divided into four sections. First, we show the weighting process of each word of the input text corpus. Second, we introduce our main contribution i.e. the informative similarity measure. Third, we define how subparts can be elected from the values of the informative similarity measure. And finally, we propose an evaluation on a real-world situation for Text Summarization.

2. Weighting Score

Our algorithm is based on the vector space model which determines the similarity of neighboring groups of sentences and places subtopic boundaries between dissimilar blocks. In our specific case, each sentence in the corpus is evaluated in terms of similarity with the previous block of *k* sentences and the next block of *k* sentences.

The simplest form of the vector space model treats a document (in our case, a sentence or a group of sentences) as a vector whose values correspond to the number of occurrences of the words appearing in the document as in (Hearst, 1994). Although (Hearst, 1994) showed successful results with this weighting scheme, we strongly believe that the importance of a word in a document does not only depend on its frequency. Indeed, frequency can only be reliable for technical texts where ambiguity is drastically limited and word repetition largely used. But unfortunately, these documents are an exception in the global environment of the internet for example. According to us, two main factors must be taken into account to define the relevance of a word for the specific task of Topic Segmentation: its semantic importance and its distribution across the text. For that purpose, we propose a new weighting scheme based on three heuristics: the well-known *tf.idf* measure, the adaptation of the *tf.idf* measure for sentences, the *tf.isf*, and a new density measure that calculates the density of each word in the text.

2.1 The *tf.idf* Score

The basic idea of the *tf.idf* score (Salton *et al.*, 1975) is to evaluate the importance of a word within a document based on its frequency and its distribution across a set of documents. The *tf.idf* is defined in equation 1 where *w* is a word and *d* a document.

$$tf.idf(w,d) = \frac{tf(w;d)}{|d|} \times \log_2 \frac{N}{df(w)} \qquad \textbf{(1)}$$

However, not all relevant words in a document are useful for Topic Segmentation. For instance, relevant words appearing in all sentences will be of no help to segment the text into topics. For that purpose, we extend the idea of the *tf.idf* to sentences.

2.2 The *tf.isf* Score

The basic idea of the *tf.isf* score is to evaluate each word in terms of its distribution over the document. Indeed, it is obvious that words occurring in many sentences within a document may not be useful for Topic Segmentation purposes. So, we will define the *tf.isf* to evaluate the importance of a word within a document based on its frequency within a given sentence and its distribution across all the sentences within the document. The *tf.isf* score is defined in equation 2 where *w* is a word, *s* a sentence, s*tf(w; s)* the number of occurrences of *w* in *s*, |*s*| the number of words in *s*, *Ns* the number of sentences within the document and *sf(w)* the number of sentences in which the word *w* occurs.

$$tf.isf(w,s) = \frac{stf(w;s)}{|s|} \times \log_2 \frac{Ns}{sf(w)} \qquad \textbf{(2)}$$

However, we can push even further our idea of word distribution. Indeed, a word *w* occurring 3 times in 3 different sentences may not have the same importance in all cases. Let's exemplify. If the 3 sentences are consecutive, the word *w* will have a strong influence on what is said in this specific region of the text. On the opposite, it will not be the case if the word *w* occurs in the first sentence, in the middle sentence and then in the last sentence. For that purpose, we propose a new density measure that calculates the density of each word in a document.

2.3 The Word Density Score

The basic idea of the word density measure is to evaluate the dispersion of a word within a document. So, very disperse words will not be as relevant as dense words. In order to evaluate the word density, we propose a new measure based on the distance of all consecutive occurrences of the word in the document. We call this measure *dens* and is defined in equation 3.

$$dens(w,d) = \sum_{k=1}^{|w|-1} \frac{1}{\ln(dist(occur(k), occur(k+1)) + e)} \qquad \textbf{(3)}$$

For any given word *w*, its density *dens(w,d)* in document d, is calculated from all the distances between all its occurrences, |*w*|. So, *occur(k)* and *occur(k+1)* respectively represent the positions in the text of two consecutive occurrences of the word *w* and *dist(occur(k), occur(k+1))* calculates the distance that separates them in terms of words within the document. Thus, by summing their inverse distances, we get a density function that gives higher scores to highly dense words. As a result, a word, the occurrences of which appear close to one another, will show small distances and as a result a high density. On the opposite, a word, the occurrences of which appear far from each other, will show high distances and as a result a small word density.

2.4 The Weighting Score

The weighting score of any word in a document can be directly derived from the previous three heuristics by combining these three scores as in equation 4 where each score is normalized so that they can be combined.

$$weight(w,d) = \left\| tf.idf(w,d) \right\| \times \left\| tf.isf(w,s) \right\| \times \left\| dens(w,d) \right\| \qquad \textbf{(4)}$$

The next step of the application of the vector space model aims at determining the similarity of neighboring groups of sentences. For that purpose, it is important to define an appropriate similarity measure. That is the objective of our next section.

3. Similarity Measure

There are a number of ways to compute the similarity between two documents. However, we show that classic similarity measures evidence problems in dealing with semantic information. Most similarity measures determine the distance between two vectors associated to two documents (i.e. Vector Space Model). However, when applying the classic similarity measures between two documents, only the identical indexes of the row vectors X_i and X_j are taken into account. However, this is not tolerable. Indeed, it is clear that both sentences (1) and (2) are similar although they do not share any word in common:

(1) *Ronaldo defeated the goalkeeper once more.*
(2) *Real Madrid striker scored again.*

The most interesting idea to avoid word repetition problems is certainly to identify lexical cohesion relationships between words. Indeed, systems should take into account semantic information that could, for instance, relate *Ronaldo* to *Real Madrid striker*. For that purpose, many authors have proposed to computationally identify these relationships (in particular, the synonym relation) using large linguistic resources such as Wordnet (Angheluta *et al.*, 2002), Roget's thesaurus (Morris and Hirst, 1991) or LDOCE (Kozima, 1993). However, these huge resources are only available for dominating languages and as a consequence do not apply to less favored languages. A much more interesting research direction is proposed by (Ponte and Croft, 1997) that propose a Topic Segmentation technique based on the Local Content Analysis (Xu and Croft, 1996), allowing substituting each sentence with words and phrases related to it. Our methodology is based on this same idea but differs from it as the word co-occurrence information is directly embedded in the calculation of the similarity between blocks of sentences thus avoiding an extra-step in the topic boundaries discovery. For that purpose, we propose a new informative similarity measure that includes in its definition the Equivalence Index Association Measure (*EI*) proposed by (Muller *et al.*, 1997) as in equation 5.

$$EI(w_1, w_2) = p(w_1 \mid w_2) \times p(w_2 \mid w_1) = \frac{f(w_1, w_2)^2}{f(w_1) \times f(w_2)} \qquad \textbf{(5)}$$

The frequency of co-occurrence $f(w_1, w_2)$ between w_1 and w_2 is calculated within a context window from a collection of documents. Our informative similarity measure is defined in equation 6 where $EI(W_{ik}, W_{jl})$ is the Equivalence Index value between W_{ik}, the word that indexes the vector of the document *i* at position *k*, and W_{jl}, the word that indexes the vector of the document *j* at position *l*.

$$S_{ij} = \text{infosimba}(X_i, X_j) = \frac{\sum_{k=1}^{p}\sum_{l=1}^{p} X_{ik} \times X_{jl} \times EI(W_{ik}, W_{jl})}{\sqrt{\sum_{k=1}^{p}\sum_{l=1}^{p} X_{ik} \times X_{il} \times EI(W_{ik}, W_{il})} \times \sqrt{\sum_{k=1}^{p}\sum_{l=1}^{p} X_{jk} \times X_{jl} \times EI(W_{jk}, W_{jl})}} \qquad \textbf{(6)}$$

The next step of the application aims at placing subtopic boundaries between dissimilar blocks. For that purpose, we propose a detection methodology based on the standard deviation algorithm proposed by (Hearst, 1994).

4. Topic Boundary Detection

Different methodologies have been proposed to place subtopic boundaries between dissimilar blocks depending on the models used to determine similarity between blocks of sentences (Kozima, 1993; Hearst, 1994; Beeferman *et al.*, 1997; Ponte and Croft, 1997; Stokes, *et al.*, 2002). Taking as reference the idea of (Ponte and Croft, 1997) who take into account the preceding and the following contexts of a segment, we calculate the informative similarity of each sentence in the corpus with its surrounding pieces of texts i.e. its previous block of *k* sentences and its next block of *k* sentences. The basic idea is to know whether the focus sentence is more similar to the preceding block of sentences or to the following block of sentences. For that purpose, we propose a score for each sentence as (Beeferman *et al.*, 1997) compare short and long-range models. It is defined in equation 7.

$$ps(S_i) = \log_2 \frac{\text{infosimba}(S_i, X_{i-1})}{\text{infosimba}(S_i, X_{i+1})} \qquad \textbf{(7)}$$

In order to better understand the variation of the *ps* score, each time its value goes from positive to negative between two consecutive sentences, there exits a topic shift. We will call this phenomenon a downhill. In fact, it means that the previous sentence is more similar to the preceding block of sentences and the following sentence is more similar to the following block of sentences thus representing a shift in topic in the text. A downhill is simply defined in equation 8 whenever the value of the *ps* score goes from positive to negative between two consecutive sentences S_i and S_{i+1}.

$$downhill(S_i, S_{i+1}) = ps(S_i) - ps(S_{i+1}) \qquad \textbf{(8)}$$

However, not all downhills identify the presence of a new topic in the text. Indeed, only deeper ones must be taken into account. In order to automatically identify these downhills, and as a consequence the topic shifts, we adapt the algorithm proposed by (Hearst, 1994) to our specific case. Downhills are topic boundaries if they satisfy the constraint expressed in equation 9 where c is a constant to be tuned and $\bar{x}$ is the average of all downhills and σ the standard deviation.

$$downhill\left(S_i, S_{i+1}\right) \geq \bar{x} + c\sigma \qquad \mathbf{(9)}$$

By applying this threshold, we obtain promising results for the discovery of topic boundaries for the specific case of web news segmentation. We illustrate these results in the next section.

5. Results

Topic Segmentation systems (Ferret, 2002; Xiang and Hongyuan, 2003) have usually been evaluated on (Choi, 2000)'s data set that represents the standard for evaluation. However, many authors have discussed the validity of this test corpus (Ferret, 2002; Xiang and Hongyuan, 2003) and proposed their own test corpus. Indeed, (Choi, 2000)'s data set, also called c99, evidences two major drawbacks: (1) it deals with segments of different domains and (2) lexical repetition is high within each segment. It is clear that the c99 corpus does not apply for an evaluation oriented towards Text Summarization. Indeed, in this case, the texts must cover a single domain and intra-segment lexical repetitions are not used as much as in the c99 corpus. However, it is likely that there exist inter-segment lexical repetitions which unease the process of boundary detection. By tackling this particular situation, we propose a new challenge compared to other works that have been proposed so far and use test corpora based on multi-domain and multi-genre segments as in (Ferret, 2002). In fact, the most similar experiment, to our knowledge, is the one proposed by (Xiang and Hongyuan, 2003) who use the *Mars* novel. However, their segments are 2650 words-long while we deal with segments around 100 words each. In fact, we aim at proposing a fine-grained system capable of finding topic boundaries with high precision in a single domain and in short texts. To our knowledge, such a challenge has never been attempted so far.

In order to evaluate our system, we propose an evaluation on a set of web documents about a unique domain using words as the basic textual information. In order to run our experiments, we built our own corpus by taking from two Portuguese soccer websites, a set of 100 articles of more or less 100 words each. Then, we built 10 test corpora by choosing randomly 10 articles from our database of 100 articles leading to 10 texts of around 1000 words-long[1].

A classical way of evaluating retrieval systems is to use Precision, Recall and F-measure. So, we show these results on our test corpus in Table 1.

	Measures	c=-1.5
T1	Precision	0,64
	Recall	0,78
	F-measure	0,70
T2	Precision	0,67
	Recall	0,67
	F-measure	0,67
T3	Precision	0,80
	Recall	0,89
	F-measure	0,84
T4	Precision	0,73
	Recall	0,89
	F-measure	0,80
T5	Precision	0,60
	Recall	0,67
	F-measure	0,63
T6	Precision	0,73
	Recall	0,89
	F-measure	0,80
T7	Precision	0,80
	Recall	0,89
	F-measure	0,84
T8	Precision	0,64
	Recall	0,78
	F-measure	0,70
T9	Precision	0,60
	Recall	0,67
	F-measure	0,63
T10	Precision	0,70
	Recall	0,78
	F-measure	0,74
Average	Precision	0,69
	Recall	0,79
	F-measure	0,73

Table 1. Quantitative Results

The results are surprisingly good considering the challenging task we were facing. Indeed, by using words as basic textual units, the average F-measure reaches 73% being Recall 79% and Precision 69%. After different tuning, the best results were obtained for c=-1.5. In any case, these global results hide most of the behavior of our system and a more detailed evaluation is needed. As (Reynar, 1994) evidences, Precision and Recall measures are overly strict. By taking into account only Precision and Recall, a hypothesized boundary close to a real segment boundary is equally detrimental to performance as one far from a boundary. This definitely should not be the case. As a consequence, we present, in Table 2, quantitative results by taking into account, as correct boundaries, all correct boundaries and all near misses with ± 1 sentence.

Precision	0,83
Recall	0,95
F-measure	0,89

Table 2. Estimated Results

We can see from these results that we would obtain 89% F-measure, which means that our system fails most correct topic for only one sentence.

[1] The chosen parameters of our experiments were the following: block size=2 sentences and EI window=10 words.

The results presented in this section are promising as we deal with a very difficult challenge which is working without any linguistic knowledge, on the basis of small mono-domain texts with many inter-segments lexical repetitions. As we said earlier, to our knowledge, such a challenge has never been attempted so far.

6. Conclusions and Future Work

In this paper, we proposed a language-independent unsupervised Topic Segmentation system based on word-co-occurrences that avoids the accessibility to existing linguistic resources such as electronic dictionaries or lexico-semantic databases. In particular, our architecture proposes a system that solves three main problems evidenced by previous research: systems based uniquely on lexical repetition that show reliability problems, systems based on lexical cohesion using existing linguistic resources that are usually available only for dominating languages and as a consequence do not apply to less favored and emerging languages and finally systems that need previously existing harvesting training data. Our evaluation has evidenced promising results showing an average F-measure of 73% being Recall 79% and Precision 69%. As immediate future work, we intend to test our system by integrating Multiword Units. Indeed, on-going results seem to lead to more accurate figures. The system and its evolutions will be available for download as a GPL license at the following address: http://asas.di.ubi.pt.

References

(Angheluta et al., 2002) Angheluta, R., De Busser, R., Moens, M-F. 2002. The Use of Topic Segmentation for Automatic Summarization. In Workshop on Text Summarization in Conjunction with the ACL 2002 and including the DARPA/NIST sponsored DUC 2002 Meeting on Text Summarization. July 11-12, Philadelphia, Pennsylvania, USA.

(Beeferman et al., 1997) Beeferman, D., Berger, A., and Lafferty, J. 1997. Text segmentation using exponential models. In Proceedings of the Second Conference on Empirical Methods in Natural Language Processing, 35--46.

(Boguraev and Neff, 2000) Boguraev, B. and Neff, M. 2000. Discourse segmentation in aid of document summarization. In Proceedings of Hawaii International Conference on System Sciences (HICSS- 33), Minitrack on Digital Documents Understanding, Maui, Hawaii. IEEE.

(Choi, 2000) Choi, F.Y.Y. 2000. Advances in Domain Independent Linear Text Segmentation. In Proceedings of NAACL'00, Seattle, April 2000. ACL.

(Cleuziou et al., 2003) Cleuziou G., Clavier V., Martin L. 2003. Une méthode de regroupement de mots fondée sur la recherche de cliques dans un graphe de cooccurrences. In Proceedings of the 5èmes rencontres Terminologie et Intelligence Artificielle), LIIA - ENSAIS ed., 179--182, Strasbourg, France.

(Farzindar and Lapalme, 2004) Farzindar, A. and Lapalme, G. 2004. Legal text summarization by exploration of the thematic structures and argumentative roles. In Text Summarization Branches Out Conference held in conjunction with ACL 2004, Barcelona, Spain, 27-38

(Ferret, 2002) Ferret, O. 2002.*Using Collocations for Topic Segmentation and Link Detection*. In Proceedings of COLING 2002, 19th International Conference on Computational Linguistics, August 24 - September 1, 2002, Taipei, Taiwan.

(Hearst, 1994) Hearst, M. 1994. Multi-Paragraph Segmentation of Expository Text, In Proceedings of the 32nd Meeting of the Association for Computational Linguistics, Las Cruces, New Mexico, June, 9--16.

(Kozima, 1993) Kozima, H. 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, 286--288.

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

(Muller, 1997) Muller, C., Polanco, X., Royauté, J. and Toussaint, Y. 1997. Acquisition et structuration des connaissances en corpus: éléments méthodologiques. Technical Report RR-3198, Inria, Institut National de Recherche en Informatique et en Automatique.

(Phillips, 1985) Phillips, M. 1985. Aspects of Text Structure: An Investigation of the Lexical Organisation of Text, North Holland Linguistic Series, North Holland, Amsterdam.

(Ponte and Croft, 1997) Ponte J.M. and Croft W.B. 1997. Text Segmentation by Topic. In Proceedings of the 1st European Conference on Research and Advanced Technology for Digitial Libraries.120--129.

(Reynar, 1994) Reynar, J.C. 1994. An Automatic Method of Finding Topic Boundaries. In Proceedings of the 32th Annual Meeting of the Association for Computational Linguistics, Las Cruces, USA.

(Salton et al., 1975) Salton, G., Yang, C.S., and Yu, C.T. 1975. A theory of term importance in automatic text analysis. Amer. Soc. Inf. Sc~ 26, 1, 33--44.

(Sardinha, 2002) Sardinha, T.B. 2002. Segmenting corpora of texts. DELTA, 2002, 18(2), 273--286. ISSN 0102-4450.

(Stokes et al., 2002) Stokes, N., Carthy, J. and Smeaton, A.F. 2002. Segmenting Broadcast News Streams Using Lexical Chains. In Proceedings of 1st Starting AI Researchers Symposium (STAIRS 2002), volume 1. 145--154.

(Xiang and Hongyuan, 2003) Xiang, J. and Hongyuan, Z. 2003. Domain-independent Text Segmentation Using Anisotropic Diffusion and Dynamic Programming. In proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Toronto, Canada. pp.322—329.

(Xu and Croft, 1996) Xu, J. and Croft, W.B. 1996. Query Expansion Using Local and Global Document Analysis. In Proceedings of the Nineteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 4--11.