Improving text segmentation by combining endogenous and exogenous methods

Olivier Ferret
CEA, LIST
18 route du Panorama, BP6, Fontenay-aux-Roses, F-92265 France
olivier.ferret@cea.fr

Abstract

Topic segmentation was addressed by a large amount of work from which it is not easy to draw conclusions, especially about the need for knowledge. In this article, we propose to combine in the same framework two methods for improving the results of a topic segmenter based on lexical reiteration. The first one is endogenous and exploits the distributional similarity of words in a document for discovering its topics. These topics are then used to facilitate the detection of topical similarity between discourse units. The second approach achieves the same goal by relying on external resources. Two resources are tested: a network of lexical co-occurrences built from a large corpus and a set of word senses induced from this network. An evaluation of the two approaches and their combination is performed in a reference framework and shows the interest of this combination both for French and English.

1 Introduction

In this article, we address the problem of linear topic segmentation, which consists in segmenting documents into topically homogeneous non-overlapping segments. This Discourse Analysis problem has received a constant interest since works such as [11]. One criterion for classifying topic segmentation systems is the kind of knowledge they depend on. Most of them only rely on surface features of documents: word reiteration in [11, 4, 20, 10], and more recently [14, 7], or discourse cues in [16, 10]. As they don't exploit external knowledge, such systems are not domain-dependent but they can be successfully applied only to some types of documents: word reiteration is reliable only if concepts are not expressed by too different means (synonyms, etc.); discourse cues are often rare and corpus-specific.

To overcome these difficulties, some systems make use of domain-independent knowledge about lexical cohesion: a lexical network built from a dictionary in [13]; a thesaurus in [15]; a large set of lexical co-occurrences collected from a corpus in [5] or [6]. To some extent, these lexical networks enable segmenters to rely on a sort of concept reiteration. However, their lack of any topical structure makes this kind of knowledge difficult to use when lexical ambiguity is high.

The most simple solution to this problem is to exploit knowledge about the topics that may occur in documents. Such topic models are generally built from a large set of example documents as in [21], [2] or in one component of [1]. These statistical topic models enable segmenters to improve their precision but they also restrict their scope.

Hybrid systems that combine the approaches we have presented were also developed and illustrated the interest of such a combination: [12] combined word recurrence, co-occurrences and a thesaurus; [1] relied on both lexical modeling and discourse cues; [10] made use of word reiteration through lexical chains and discourse cues.

The work we report in this article takes place in the last category we have presented. More precisely, it first confirms the interest of combining lexical recurrence with an external resource about lexical cohesion of texts. Second, it shows that the improvement brought by the use of a resource about lexical cohesion and the improvement brought by an endogenous method such as the one presented in [9] are complementary and can be fruitfully combined.

2 Overview

In most of the algorithms in the text segmentation field, documents are represented as sequences of basic discourse units. When they are written texts, these units are generally sentences, which is also the case in our work. Each unit is turned into a vector of words, following the principles of the *Vector Space* model. Then, the similarity between the basic units of a text is evaluated by computing a similarity measure between the vectors that represent them. Such a similarity is considered as representative of the topical closeness of the corresponding units. This principle is also applied to groups of basic units, such as text segments, because of the properties of the *Vector Space* model. Segments are finally delimited by locating the areas where the similarity between units or groups of units is weak.

This quick overview highlights the important role of the evaluation of the similarity between discourse units in the segmentation process. When no external knowledge is used, this similarity is only based on the reiteration of words. But it can be enhanced by taking into account semantic relations between words. Such relations can be found in manually-built semantic resources such as WordNet or Roget's Thesaurus. Although their coverage is large, these resources can't obviously cover all domains in depth. As a consequence, it can also be interesting to rely on resources whose relations are less well defined but that can be easily extended to a new domain in an unsupervised way

International Conference RANLP 2009 - Borovets, Bulgaria, pages 88–93

from a representative corpus. The cohesion relations that are captured through the lexical co-occurrences extracted from a set of texts fulfill these constraints and were already exploited with success together with word reiteration in [12].

In [9], another way to improve the evaluation of the similarity between two sentences of a text is proposed: the idea is to define each topic of the text as a subset of its vocabulary and to use the implicit relation between each couple of words that are part of the same topic for detecting the topical similarity of sentences. A large repository of topics doesn't exist and similarly to semantic resources, it couldn't cover all domains. As a consequence, [9] performs the discovering of the topics of a text in an unsupervised way and its method can be qualified as endogenous as it doesn't rely on any external resource. This method is used in conjunction with a method based on word reiteration that is implemented in the same framework.

The first objective of the work we report here is to generalize this framework for evaluating the use of different types of lexical relations to improve the detection of similarity between text units in the context of topic segmentation. Its second objective is to test an exogenous approach in this framework in conjunction with word reiteration. More precisely, two sources of knowledge have been used to support the exogenous approach: a network of lexical co-occurrences and a set of word senses induced from this network. The first one aims at confirming the results of [12] while the second one aims at testing the interest of selecting the most significant co-occurrences from a semantic viewpoint. The last main objective of this work is to determine if the association in the same framework of two different kinds of methods for improving the detection of similarity between text units, endogenous and exogenous methods, can lead to better results.

3 The F06 framework for text segmentation

3.1 Overall principles

The F06 framework is globally based on the same principles as *TextTiling* [11]. Its first stage, the linguistic pre-processing of texts, splits each text into sentences and represents each of them as the sequence of its normalized plain words, that is, nouns (proper and common nouns), verbs and adjectives. It is performed by the *TreeTagger* tool both for French and English, our two target languages. Finally, each sentence is turned into a vector of normalized plain words.

The evaluation of the lexical cohesion of a text, the second stage, relies as for *TextTiling* on a fixed-size focus window that is moved over the text to segment and stops at each sentence break. A cohesion value is computed at each position of this window and is associated to the sentence break at the transition between the two sides of the window. The final result is a cohesion graph of the text.

The last stage of F06 is mainly taken from the *LCseg* system [10]. It starts by evaluating the likelihood of each minimum m of the cohesion graph to be a topic shift. First, the pair of maxima l and r around m is found. Then, its score as a topic shift is computed as:

$$score(m) = \frac{LC(l) + LC(r) - 2 \cdot LC(m)}{2} \tag{1}$$

This score favors as possible topic shifts minima that correspond to sharp falls of lexical cohesion.

The next step is done by removing as a possible topic shift each minimum that is not farther than 2 sentences from its preceding neighbor. Finally, the selection of topic shifts is performed by applying a threshold computed from the distribution of minimum scores. Thus, a minimum m is kept as a topic shift if $score(m) > \mu - \alpha \cdot \sigma$, where μ is the average of minimum scores, σ their standard deviation and α is a modulator ($\alpha = 0.6$ in all our experiments).

3.2 Evaluation of lexical cohesion

As pointed out in Section 2, the evaluation of the cohesion in the sliding window of the text segmenter is the most important stage of the segmentation process and is the focus of the improvements we explore in this article. Globally, this evaluation is performed following [12]: each side of the window is represented by a vector and the cohesion in the window is evaluated by applying the *Dice coefficient* between its two sides. When the evaluation of the cohesion is only based on word reiteration, this principle is applied literally: if W_l refers to the vocabulary of the left side of the focus window and W_r to the vocabulary of its right side, the cohesion in the window at a text position is given by:

$$LC_{rec} = \frac{2 \cdot card(W_l \cap W_r)}{card(W_l) + card(W_r)} \tag{2}$$

More generally, this definition is suitable for relations of equivalence between words, which are limited in the present case to the equality between lemmas. For the other types of lexical cohesion relations, the *Dice coefficient* is extended in the following way: instead of focusing on words that are shared by the two sides of the window, the measure takes into account the words of one side of the window that are linked to words of the other side of the window according to the considered type of relations. More precisely, the cohesion in the window for a relation type rel_i takes the form:

$$\begin{aligned} LC_{rel_i} = {} & \frac{card(W_{rel_i}(l) - W_{rec} - \bigcup_{j \neq i} W_{rel_j}(l))}{card(W_l) + card(W_r)} + \\ & \frac{card(W_{rel_i}(r) - W_{rec} - \bigcup_{j \neq i} W_{rel_j}(r))}{card(W_l) + card(W_r)} \end{aligned} \tag{3}$$

where

$W_{rel_i}(x)$ is the set of words from the (x=l)eft or the (x=r)ight side of the window that are selected according to lexical relations of type rel_i,

$W_{rec} = card(W_l \cap W_r)$, words in a recurrence relation,

$\bigcup_{j \neq i} W_{rel_j}(x)$ gathers the set of words from the x side of the window that are selected according to lexical relations that are different from rel_i.

By removing from $W_{rel_i}(x)$ words that also appear in W_{rec} or $\bigcup_{j \neq i} W_{rel_j}(x)$, we make LC_{rel_i} measure the specific contribution of rel_i relations to the detection of the cohesion between the two sides of the window.

Finally, the global cohesion inside the window is the result of the sum of the cohesion values computed for each kind of lexical relations:

$$LC = LC_{rec} + \sum_i LC_{rel_i} \quad (4)$$

4 Improving text segmentation by an endogenous method

In this section, we will not present F06T in detail, the method described in [9] for improving text segmentation in an endogenous way. We will only remind its main principles and show how it can be defined in the F06 framework.

The specificity of the F06T segmenter in relation to the F06 framework is the use of the topics of the text to segment in the evaluation of the cohesion of the focus window. These topics are identified in an unsupervised way by clustering the words of the text according to their co-occurrents in this text. Thus, each of its topic is represented by a subset of its vocabulary.

In this context, the evaluation of the cohesion of the focus windows starts by the determination of the topics of the window that are actually representative of its content. Several topics may be associated to the focus window as a theme of a text may be scattered over several identified topics due to the absence of external reference topics. A topic is considered as representative of the content of the focus window only if it matches each side of this window. In practice, this matching is evaluated by applying the *Cosine* measure between the vector that represents one side of the window and the vector that represents the topic.

The computation of the cohesion of the focus windows from these selected text topics first consists in determining for each side of this window the number of its words that belong to one of these topics. The topical cohesion of the window, LC_{top}, is then given by Equation 5, derived from Equation 3:

$$\frac{card(W_{top}(l) - W_{rec})}{card(W_l) + card(W_r)} + \frac{card(W_{top}(r) - W_{rec})}{card(W_l) + card(W_r)} \quad (5)$$

where $W_{top}(i)_{i \in \{l,r\}} = W_i \cap T_w$ and T_w is the union of all the representations of the topics associated to the window. $W_{top}(i)$ corresponds to the words of the i side of the window that belong to the topics of the window.

Finally, the global cohesion in the focus window for F06T is computed as the sum of the cohesion computed from word reiteration, LC_{rec}, and the one computed from text topics LC_{top}, in accordance with Equation 4.

5 Improving text segmentation by exogenous methods

We present now the use in the F06 framework of external resources about relations between words. We first consider lexical co-occurrences, a resource that can be extracted from large corpora in an easy way.

5.1 Using lexical co-occurrences

5.1.1 Co-occurrence networks

For the experiments of Section 6, two networks of lexical co-occurrences were built: one for French, from the *Le Monde* newspaper, and one for English, from the *L.A. Times* newspaper. The size of each corpus was around 40 million words.

The building process was the same for the two networks. First, the initial corpus was pre-processed similarly to the texts to segment (see Section 3.1). Co-occurrences were classically extracted by moving a fixed-size window on texts with parameters for catching topical relations: the window was rather large, 20-word wide, took into account the boundaries of texts and co-occurrences were indifferent to word order. We adopted the Pointwise Mutual Information as the cohesion measure of each co-occurrence. This measure was normalized according to the maximal mutual information relative to the considered corpus. After filtering the less frequent and cohesive co-occurrences, we got a network with approximately 23,000 words and 5.2 million co-occurrences for French, 30,000 words and 4.8 million co-occurrences for English.

5.1.2 Using co-occurrence networks for segmentation

Similarly to F06T, F06C, the topic segmenter based on lexical co-occurrences, evaluates the cohesion inside the focus window in two steps. First, it uses its resource for selecting the words of one side of the focus window ($W_{coo}(x)$) that are the most strongly linked to words of the other side of this window. As lexical co-occurrences are not as reliable as semantic relations coming from a resource such as WordNet, this selection is only based on the most cohesive co-occurrences and must rely on several co-occurrence relations. More precisely, a word of one side of the focus window is selected if:

- it has direct co-occurrence relations with at least N words of the other side of the window ($N = 2$ in the experiments of Section 6);
- the frequency and the cohesion of these support co-occurrence relations are higher than a fixed threshold (14 for frequency and 0.14 for cohesion).

The second step is the computation of the cohesion in F06C's focus window following Equations 3 and 4. $LC_{F06C} = LC_{rec} + LC_{coo}$, where LC_{coo}, the cohesion from co-occurrence relations, is equal to:

$$\frac{card(W_{coo}(l) - W_{rec})}{card(W_l) + card(W_r)} + \frac{card(W_{coo}(r) - W_{rec})}{card(W_l) + card(W_r)} \quad (6)$$

5.2 Using word senses

Lexical co-occurrences represent a rather crude resource and it is interesting to test if more elaborated resources can lead to better results. In this section, we consider word senses that were discriminated from a corpus, a resource that can be seen halfway between co-occurrences and semantic knowledge.

mouse-device	computer#n, disk#n, pc#n, software#n, user#n, machine#n, screen#n, compatible#a ...
mouse-animal	hormone#n, tumour#n, immune#a, researcher#n, animal#n, disease#n, gene#n ...

Table 1: *Two senses of the word "mouse"*

5.2.1 Word senses discriminated from texts

The word senses we use in this work were built according to the method described in [8]. More precisely, the building process starts from a network of lexical co-occurrences as the ones described in Section 5.1.1. First, the subgraph of the co-occurrents of the target word is delimited and turned into a similarity graph where the similarity between two co-occurrents is equal to their cohesion in the network. Then, a clustering algorithm is applied for detecting high-density areas in this graph. Finally, a word sense is defined from each resulting cluster. An example of such word senses is given by Table 1 with the two senses found for the noun *mouse*.

A set of word senses were built from the two co-occurrence networks of Section 5.1.1. Due to the sparsity of these networks, senses are not discriminated for all their words. For French, the word sense base is made of 7,373 lemmas with an average number of senses by word equal to 2.8 whereas for English, the base is made of 9,838 lemmas with 2.0 senses by word.

5.2.2 Using word senses for segmentation

The discrimination of the senses of a word evoked in the previous section can also be seen as a way of filtering and structuring its co-occurrents. As a consequence, the use of lexical co-occurrences for topic segmentation described in Section 5.1.2 can be extended rather straightforwardly to the use of such word senses. The resulting segmenter is called F06WS.

This extension mainly consists in adding a preliminary step: before selecting the words of each side of the focus window that are the most strongly linked to words of the other side of this window, a word sense disambiguation process is applied to them to determine which of their senses are actually present. The selection is then performed as in Section 5.1.2 except that it is only based on the co-occurrents that are part of the definition of the senses kept by the word sense disambiguation process. More precisely, this process follows the principles defined by Lesk: it selects a sense for a word according to the overlap between its definition and the side of the window this word is part of. This overlap is evaluated here by computing the *Cosine* measure between the definition of the sense and the side of the window, both turned into vectors of lemmas. The sense with the highest similarity is kept.

6 Evaluation

6.1 Evaluation methodology

Choi proposed in [4] an evaluation methodology for topic segmentation systems that has became a kind of standard. This methodology is based on the building of artificial texts made of segments extracted from different documents. Because it is not well adapted to the evaluation of endogenous approaches, [9] proposed to adapt this methodology concerning the way the document segments are selected.

Instead of taking each segment from a different document, [9] uses two source documents referring to two different topics. This ensures that the boundary between two adjacent segments of an evaluation document actually corresponds to a topic shift. Each of the two source documents is split into a set of segments whose size is between 3 and 11 sentences, as for Choi, and an evaluation document is built by concatenating these segments in an alternate way from the beginning of the source documents until 10 segments are extracted. The topics of the source documents are controlled by taking them from the corpus of the CLEF 2003 evaluation for crosslingual information retrieval: each evaluation document is built from two source documents that were judged as relevant for two different CLEF 2003 topics. We used for our evaluations the two corpora of [9], one for French, one for English, as the results of F06R and F06T on these corpora were already known.

6.2 Using external resources

First, we evaluated the interest of using external resources in F06 by applying F06C and F06WS to the two evaluation corpora. We classically used the error metric P_k proposed in [1] and its variant *WindowDiff* (*WD*) [17] to measure segmentation accuracy. P_k and *WD* are given as percentages in the next tables (smallest values are best results).

Systems	$\mathbf{P_k}$	WD
U00	25.91	27.42
C99	27.57	35.42
TextTiling*	21.08	27.43
LCseg	20.55	28.31
F06R	21.58	27.83
F06C	16.48	20.94
F06WS	18.17	23.14

Table 2: *Evaluation of F06 with external resources for the French corpus*

Tables 2 and 3 show both the results of our evaluations for F06C and F06WS and the results reported in [9] for F06R (F06 with only word recurrence) and several reference topic segmenters: U00 [20], C99 [4] and *LCseg* [10]; *TextTiling** is a variant of *TextTiling* in which the final identification of topic shifts from the cohesion graph is taken from [10]. All these systems were used without fixing the number of topic shifts to find. As pointed out in [9], the results of these reference systems show that the corpora we used are more difficult than Choi's corpus.

In the F06 framework, the results are globally similar for the two corpora: the use of external resources in

Systems	$\mathbf{P_k}$	WD
U00	19.42	21.22
C99	21.63	30.64
TextTiling*	15.81	19.80
LCseg	14.78	19.73
F06R	16.90	20.93
F06C	14.85	21.00
F06WS	15.89	19.30

Table 3: *Evaluation of F06 with external resources for the English corpus*

addition to word recurrence improves topic segmentation but the use of word senses instead of co-occurrence relations is not as interesting as we could expected.

Nevertheless, the detailed situation is a little bit different for French and English. For French, the results of F06C and F06WS are higher than those of F06R in a significant way[1] and the difference between F06C and F06WS is not significant. For English, the difference between F06C and F06R or between F06WS and F06R are not statistically significant. There is no obvious explanation of this fact but we can notice that the average level of P_k and WD values of methods based on word recurrence is clearly higher for English than for French, which means that word recurrence is a more reliable way to detect topic similarity in English than in French. As a consequence, the use of external resources is less necessary for English than for French. Moreover, the high level of results based on word recurrence make them difficult to increase.

The lack of interest of word senses is also difficult to interpret. Their use was supposed to restrict the number of words that are wrongly detected as topically linked in the focus window of the segmenter. The results show that in practice, such restriction is too strict and certainly leads to discard relevant links. This loss is at least partially due to two characteristics of the clustering algorithm used for discriminating senses: it removes some of the co-occurrents of the word and it performs hard clustering, which means that some co-occurrents that should be shared by several senses are assigned to only one of them.

6.3 Combining endogenous and exogenous approaches

The general definition of the cohesion in the focus window of a F06 segmenter given by Equation 3 offers a direct means of integrating the endogenous approach of F06T and the exogenous approaches of F06C and F06WS. More precisely, as the evaluation of the previous section has shown that F06WS tends to have worst results than F06C, even if the difference is not significant, we will only consider F06C for this integration. Hence, the cohesion in the focus window of this integrating segmenter, called F06CT, is given by this instance of Equation 7:

$$LC = LC_{rec} + LC_{top-coo} + LC_{coo-top} \qquad (7)$$

where $LC_{top-coo}$, the contribution of the endogenous approach, is equal to LC_{top} without taking into account words of $W_{coo}(x)$ and $LC_{coo-top}$, the contribution of the exogenous approach, is equal to $LC_{F06C} - LC_{rec}$ without taking into account words of $W_{top}(x)$.

Systems	$\mathbf{P_k}$	WD
F06R	21.58	27.83
F06T	18.46	24.05
F06C	16.48	20.94
F06CT	14.59	18.41

Table 4: *Evaluation of the combination of approaches in F06 for the French corpus*

Tables 4 and 5 show the results of F06CT on the two evaluation corpora together with the results of the other F06 segmenters. The first point to note is that F06CT have the highest results among all the F06 segmenters. Both for French and English, F06CT significantly outperforms F06. Moreover, [9] reports that F06T also outperforms F06 in both cases. But as for F06C, there are also some differences for the two languages: the difference between F06CT and F06C is not significant for French while it is significant for English and the difference between F06CT and F06T is significant for French but not for English. This globally confirms our findings from the first evaluation. For English, the use of lexical co-occurrences is less effective than for French. As a consequence, a significant part of F06CT's results for English can certainly be explained by its endogenous approach while for French, the dominant part in F06CT's results seems rather come from its exogenous approach.

Systems	$\mathbf{P_k}$	WD
F06R	16.90	20.93
F06T	14.06	18.31
F06C	14.85	21.00
F06CT	12.30	14.88

Table 5: *Evaluation of the combination of approaches in F06 for the English corpus*

Nevertheless, the results of this evaluation also show that the two kinds of approaches, exogenous and endogenous, are complementary: for French, F06CT significantly outperforms F06 and F06T while F06T significantly outperforms F06. This means that the cohesion relations brought by lexical co-occurrences are different from the endogenous relations extracted from the unsupervised topic identification and can be fruitfully associated.

7 Related work

Our work has two main characteristics. First, it focuses on the detection of the topical similarity of text units. Second, it tests and integrates several approaches for performing this detection. These two aspects are not tackled by many works as most of the

[1] The significance level of differences is evaluated by a one-side t-test with a null hypothesis of equal means. Levels lower than 0.05 are considered as statistically significant.

work in this field concentrates on the application of statistical models to topic segmentation and relies on a basic similarity measure between text units. This can be partly explained by the differences we have observed in our evaluations between French and English: almost all works are dedicated to English, a language in which word recurrence seems to be particularly effective for topic segmentation. As a consequence, the use of external resources is not considered as a priority. But we have seen that the situation can be different for other languages, such as French.

Some works were done following this trend nevertheless. One way that was commonly adopted for improving the evaluation of this similarity without being dependent of a particular domain was to exploit a semantic space built from a large corpus. In CWM [5], a variant of C99, each word of a sentence is replaced by its representation in a *Latent Semantic Analysis* (LSA) space. In the work of Ponte and Croft [18], the representations of sentences are expanded by adding to them words selected from an external corpus by the means of the *Local Context Analysis* (LCA) method. Finally in [3], a set of concepts are learnt from a corpus in an unsupervised way by using the X-means clustering algorithm and the paragraphs of documents are represented in the space defined by these concepts. Co-occurrence relations were more directly used by [6] for supporting a similarity measure between sentences. Works that exploit manually-built resources such as [13], [15] or [19] also exist but they generally don't use these resources for evaluating directly the similarity of text units: in [15] and [19] for instance, they help in identifying lexical chains.

More globally, these works explore one way to detect the similarity of text units but they don't try to integrate several approaches. Hybrid systems are rare for topic segmentation and as [10] or [1], they aims at integrating content-based approaches with discourse cues. The only work that can be compared to ours from this viewpoint is [12], which combines word recurrence, co-occurrence relations and relations from a thesaurus. Although their evaluation framework is different from ours, their results also confirm the interest of combining word recurrence with external resources.

8 Conclusion and future work

In this article, we have first proposed to generalize the framework of [9] for topic segmentation to integrate external resources about lexical cohesion. Then, we have presented how to exploit in this new framework, F06, two such resources: a network of lexical co-occurrences and a repository of word senses induced from this network. The evaluation of the segmenters integrating word recurrence with these resources have shown that both of them improve segmentation results but that word senses don't outperform co-occurrences. Finally, we have combined the endogenous approach of [9] and the best exogenous approach we have tested and shown the interest of such a combination.

As future work, we plan to extend this work in three ways. First, we want to add to the tested external resources a manually-built resource about synonymy relations between words. This will be comparable to the use of a thesaurus in [12]. The second extension will test further the use of word senses by considering senses defined by similar words instead of co-occurrents. The last extension concerns the unsupervised topic identification process that underlies the endogenous approach. In [9], this identification only relies on the distributional similarity of words in documents. Using external resources such as lexical co-occurrences or synonyms could also improve this process, with finally a positive impact on topic segmentation.

References

[1] D. Beeferman, A. Berger, and J. Lafferty. Statistical models for text segmentation. *Machine Learning*, 34(1):177–210, 1999.

[2] D. M. Blei and P. J. Moreno. Topic segmentation with an aspect hidden markov model. In *24th ACM SIGIR*, pages 343–348, 2001.

[3] M. Caillet, J.-F. Pessiot, M. Amini, and P. Gallinari. Unsupervised learning with term clustering for thematic segmentation of texts. In *RIAO'04*, pages 1–11, 2004.

[4] F. Y. Y. Choi. Advances in domain independent linear text segmentation. In *NAACL'00*, pages 26–33, 2000.

[5] F. Y. Y. Choi, P. Wiemer-Hastings, and J. Moore. Latent semantic analysis for text segmentation. In *EMNLP'01*, pages 109–117, 2001.

[6] G. Dias, E. Alves, and J. G. P. Lopes. Topic segmentation algorithms for text summarization and passage retrieval: An exhaustive evaluation. In *22th AAAI*, pages 1334–1340, 2007.

[7] J. Eisenstein and R. Barzilay. Bayesian unsupervised topic segmentation. In *EMNLP 2008*, pages 334–343, 2008.

[8] O. Ferret. Discovering word senses from a network of lexical cooccurrences. In *20th COLING*, pages 1326–1332, 2004.

[9] O. Ferret. Finding document topics for improving topic segmentation. In *45th ACL*, pages 480–487, 2007.

[10] M. Galley, K. McKeown, E. Fosler-Lussier, and H. Jing. Discourse segmentation of multi-party conversation. In *ACL'03*, pages 562–569, 2003.

[11] M. A. Hearst. Multi-paragraph segmentation of expository text. In *ACL'94*, pages 9–16, 1994.

[12] A. C. Jobbins and L. J. Evett. Text segmentation using reiteration and collocation. In *ACL-COLING'98*, pages 614–618, 1998.

[13] H. Kozima. Text segmentation based on similarity between words. In *ACL'93 (Student Session)*, pages 286–288, 1993.

[14] I. Malioutov and R. Barzilay. Minimum cut model for spoken lecture segmentation. In *COLING-ACL 2006*, pages 25–32, 2006.

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

[16] R. J. Passonneau and D. J. Litman. Discourse segmentation by human and automated means. *Computational Linguistics*, 23(1):103–139, 1997.

[17] L. Pevzner and M. A. Hearst. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36, 2002.

[18] J. M. Ponte and B. W. Croft. Text segmentation by topic. In *First European Conference on research and advanced technology for digital libraries*, 1997.

[19] N. Stokes, J. Carthy, and A. Smeaton. Segmenting broadcast news streams using lexical chains. In *STAIRS'02*, pages 145–154, 2002.

[20] M. Utiyama and H. Isahara. A statistical model for domain-independent text segmentation. In *ACL'01*, pages 491–498, 2001.

[21] J. Yamron, I. Carp, L. Gillick, S. Lowe, and P. van Mulbregt. A hidden markov model approach to text segmentation and event tracking. In *23th ICASSP*, pages 333–336, 1998.