Finding document topics for improving topic segmentation

Olivier Ferret
CEA LIST, LIC2M
18 route du Panorama, BP6
Fontenay aux Roses, F-92265 France
ferreto@zoe.cea.fr

Abstract

Topic segmentation and identification are often tackled as separate problems whereas they are both part of topic analysis. In this article, we study how topic identification can help to improve a topic segmenter based on word reiteration. We first present an unsupervised method for discovering the topics of a text. Then, we detail how these topics are used by segmentation for finding topical similarities between text segments. Finally, we show through the results of an evaluation done both for French and English the interest of the method we propose.

1 Introduction

In this article, we address the problem of linear topic segmentation, which consists in segmenting documents into topically homogeneous segments that does not overlap each other. This part of the Discourse Analysis field has received a constant interest since the initial work in this domain such as (Hearst, 1994). 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 (Hearst, 1994; Choi, 2000; Utiyama and Isahara, 2001; Galley et al., 2003) or discourse cues in (Passonneau and Litman, 1997; Galley et al., 2003). As such systems do not require external knowledge, they are not sensitive to domains but they are limited by the type of documents they can be applied to: lexical reiteration is reliable only if concepts are not too frequently expressed by several means (synonyms, etc.) and 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 (Kozima, 1993); a thesaurus in (Morris and Hirst, 1991); a large set of lexical co-occurrences collected from a corpus in (Choi et al., 2001). To a certain extent, these lexical networks enable topic segmenters to exploit a sort of concept reiteration. However, their lack of any explicit 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 (Yamron et al., 1998), (Blei and Moreno, 2001) or in one component of (Beeferman et al., 1999). 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: (Jobbins and Evett, 1998) combined word recurrence, co-occurrences and a thesaurus; (Beeferman et al., 1999) relied on both lexical modeling and discourse cues; (Galley et al., 2003) made use of word reiteration through lexical chains and discourse cues.

The work we report in this article takes place in the first category we have presented. It does not rely on any *a priori* knowledge and exploits word usage rather than discourse cues. More precisely, we present a new method for enhancing the results

Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 480–487,
Prague, Czech Republic, June 2007.

of segmentation systems based on word reiteration without relying on any external knowledge.

2 Principles

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 strict reiteration of words. But it can be enhanced by taking into account semantic relations between words. This was done for instance in (Jobbins and Evett, 1998) by taking semantic relations from Roget's Thesaurus. This resource was also used in (Morris and Hirst, 1991) where the similarity between discourse units was more indirectly evaluated through the lexical chains they share. The same approach was adopted in (Stokes et al., 2002) but with WordNet as the reference semantic resource.

In this article, we propose to improve the detection of topical similarity between text segments but without relying on any external knowledge. For each text to segment, we first identify its topics by performing an unsupervised clustering of its words according to their co-occurrents in the text. Thus, each of its topics is represented by a subset of its vocabulary. When the similarity between two segments is evaluated during segmentation, the words they share are first considered but the presence of words of the same topic is also taken into account. This makes it possible to find similar two segments that refer to the same topic although they do not share a lot of words. It is also a way to exploit long-range relations between words at a local level. More globally, it helps to reduce the false detection of topic shifts.

3 Unsupervised Topic Identification

The approach we propose first requires to discover the topics of texts. For performing such a task without using *a priori* knowledge, we assume that the most representative words of each of the topics of a text occur in similar contexts. Hence, for each word of the text with a minimal frequency, we collect its co-occurrents, we evaluate the pairwise similarity of these selected text words by relying on their co-occurrents and finally, we build topics by applying an unsupervised clustering method to them.

3.1 Building the similarity matrix of text words

The first step for discovering the topics of a text is a linguistic pre-processing of it. This pre-processing splits the text into sentences and represents each of them as the sequence of its lemmatized plain words, that is, nouns (proper and common nouns), verbs and adjectives. After filtering the low frequency words of the text (frequency < 3), the co-occurrents of the remaining words are classically collected by recording the co-occurrences in a fixed-size window (15 plain words) moved over the pre-processed text. As a result, each text word is represented by a vector that contains its co-occurrents and their co-occurrence frequency. The pairwise similarity between all the selected text words is then evaluated for building their similarity matrix. We classically apply the *Cosine* measure between the vectors that represent them for this evaluation.

3.2 From a similarity matrix to text topics

The final step for discovering the topics of a text is the unsupervised clustering of its words from their similarity matrix. We rely for this task on an adaptation of the Shared Nearest Neighbor (SNN) algorithm described in (Ertöz et al., 2001). This algorithm particularly fits our needs as it automatically determines the number of clusters – in our case the number of topics of a text – and does not take into account the elements that are not representative of the clusters it builds. This last point is important for our application as all the plain words of a text are not representative of its topics. The SNN algorithm

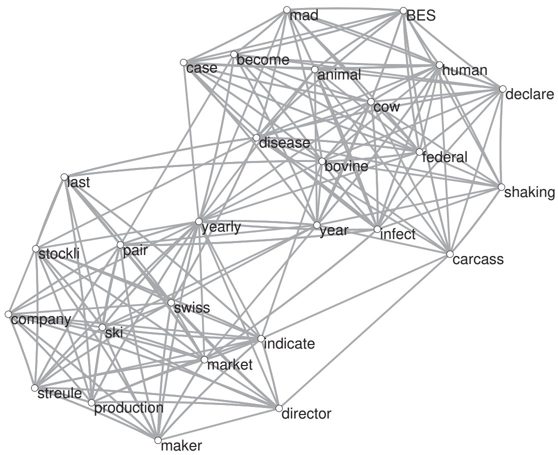

Figure 1: Similarity graph after its sparsification

(see Algorithm 1) performs clustering by detecting high-density areas in a similarity graph. In our case, the similarity graph is directly built from the similarity matrix: each vertex represents a text word and an edge links two words whose similarity is not null. The SNN algorithm splits up into two main stages: the first one finds the elements that are the most representative of their neighborhood. These elements are the seeds of the final clusters that are built in the second stage by aggregating the remaining elements to those selected by the first stage. This first stage

Algorithm 1 SNN algorithm
1. sparsification of the similarity graph
2. building of the SNN graph
3. computation of the distribution of strong links
4. search for topic seeds and filtering of noise
5. building of text topics
6. removal of insignificant topics
7. extension of text topics

starts by sparsifying the similarity graph, which is done by keeping only the links towards the k (k=10) most similar neighbors of each text word (step 1). Figure 1 shows the resulting graph for a two-topic document of our evaluation framework (see Section 5.1). Then, the similarity graph is transposed into a shared nearest neighbor (SNN) graph (step 2). In this graph, the similarity between two words is given by the number of direct neighbors they share in the similarity graph. This transposition makes the similarity values more reliable, especially for high-dimensional data like textual data. Strong links in the SNN graph are finally detected by applying a fixed threshold to the distribution of shared neighbor numbers (step 3). A word with a high number of strong links is taken as the seed of a topic as it is representative of the set of words that are linked to it. On the contrary, a word with few strong links is supposed to be outlier (step 4).

The second stage of the SNN algorithm first builds text topics by associating to topic seeds the remaining words that are the most similar to them provided that their number of shared neighbors is high enough (step 5). Moreover, the seeds that are judged as too close to each other are also grouped during this step in accordance with the same criteria. The last two steps bring small improvements to the results of this clustering. First, when the number of words of a topic is too small (size $<$ 3), this topic is judged as insignificant and it is discarded (step 6). Its words are added to the set of words without topic after step 5. We added this step to the SNN algorithm to balance the fact that without any external knowledge, all the semantic relations between text words cannot be found by relying only on co-occurrence. Finally, the remaining text topics are extended by associating to them the words that are neither noise nor already part of a topic (step 7). As topics are defined at this point more precisely than at step 4, the integration of words that are not strongly linked to a topic seed can be safely performed by relying on the average strength of their links in the SNN graph with the words of the topic. After the SNN algorithm is applied, a set of topics is associated to the text to segment, each of them being defined as a subset of its vocabulary.

4 Using Text Topics for Segmentation

4.1 Topic segmentation using word reiteration

As *TextTiling*, the topic segmentation method of Hearst (Hearst, 1994), the topic segmenter we propose, called F06, first evaluates the lexical cohesion of texts and then finds their topic shifts by identifying breaks in this cohesion. The first step of this process is the linguistic pre-processing of texts, which is identical for topic segmentation to the pre-processing described in Section 3.1 for the discovering of text topics. The evaluation of the lexical cohesion of a text relies as for *TextTiling* on a fixed-size focus window that is moved over the text to segment and stops at each sentence break. The cohesion in the part of text delimited by this window is evaluated by measuring the word reiteration between its two sides. This is done in our case by applying the *Dice coefficient* between the two sides of the focus window, following (Jobbins and Evett, 1998). This cohesion value is associated to the sentence break at the transition between the two sides of the window. More precisely, if W_l refers to the vocabulary of the left side of the focus window and W_r refers to the vocabulary of its right side, the cohesion in the window at position x is given by:

$$LC_{rec}(x) = \frac{2 \cdot card(W_l \cap W_r)}{card(W_l) + card(W_r)} \quad (1)$$

This measure was adopted instead of the *Cosine* measure used in *TextTiling* because its definition in terms of sets makes it easier to extend for taking into account other types of relations, as in (Jobbins and Evett, 1998). A cohesion value is computed for each sentence break of the text to segment and the final result is a cohesion graph of the text.

The last part of our algorithm is mainly taken from the *LCseg* system (Galley et al., 2003) and is divided into three steps:

- computation of a score evaluating the probability of each minimum of the cohesion graph to be a topic shift;
- removal of segments with a too small size;
- selection of topic shifts.

The computation of the score of a minimum m begins by finding the pair of maxima l and r around it. This score is then given by:

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

This score, whose values are between 0 and 1, is a measure of how high is the difference between the minimum and the maxima around it. Hence, it 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 our experiments).

4.2 Using text topics to enhance segmentation

The heart of the algorithm we have presented above is the evaluation of lexical cohesion in the focus window, as given by Equation 1. This evaluation is also a weak point as $card(W_l \cap W_r)$ only relies on word reiteration. As a consequence, two different words that respectively belongs to W_l and W_r but also belong to the same text topic cannot contribute to the identification of a possible topical similarity between the two sides of the focus window.

The algorithm F06T is based on the same principles as F06 but it extends the evaluation of lexical cohesion by taking into account the topical proximity of words. The reference topics for judging this proximity are of course the text topics discovered by the method of Section 3. In this extended version, the evaluation of the cohesion in the focus window is made of three steps:

- computation of the word reiteration cohesion;
- determination of the topic(s) of the window;
- computation of the cohesion based on text topics and fusion of the two kinds of cohesion.

The first step is identical to the computation of the cohesion in F06. The second one aims at restricting the set of topics that are used in the last step to the topics that are actually representative of the content of the focus window, *i.e.* representative of the current context of discourse. This point is especially important in the areas where the current topic is changing because amplifying the influence of the surrounding topics can lead to the topic shift being missed. Hence, 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[1] and by testing if the resulting value is higher than a fixed threshold (equal to 0.1 in the experiments of Section 5). It must be noted that several topics may be associated to the focus window. As the discovering of text topics is done in an unsupervised way and without any external knowledge, a theme of a text may be scattered over several identified topics and then, its presence can be characterized by several of them.

The last step of the cohesion evaluation first consists in determining for each side of the focus window the number of its words that belong to one of the topics associated to the window. The cohesion of the window is then given by Equation 3, that estimates the significance of the presence of the text topics in the window:

$$LC_{top}(x) = \frac{card(TW_l) + card(TW_r)}{card(W_l) + card(W_r)} \qquad (3)$$

where $TW_{i \in \{l,r\}} = (W_i \cap T_w) - (W_l \cap W_r)$ and T_w is the union of all the representations of the topics associated to the window. TW_i corresponds to the words of the i side of the window that belong to the topics of the window ($W_i \cap T_w$) but are not part of the vocabulary from which the lexical cohesion based on word reiteration is computed ($W_l \cap W_r$).

Finally, the global cohesion in the focus window is computed as the sum of the two kinds of cohesion, the one computed from word reiteration (see Equation 1) and the one computed from text topics (see Equation 3).

5 Evaluation

5.1 Evaluation framework

The main objective of our evaluation was to verify that taking into account text topics discovered without relying on external knowledge can actually improve a topic segmentation algorithm that is initially based on word reiteration. Since the work of Choi (Choi, 2000), the evaluation framework he proposed has become a kind of standard for the evaluation of topic segmentation algorithms. This framework is based on the building of artificial texts made of segments extracted from different documents. It has at least two advantages: the reference corpus is easy to build as it does not require human annotations; parameters such as the size of the documents or the segments can be precisely controlled. But it has also an obvious drawback: its texts are artificial. This is a problem in our case as our algorithm for discovering text topics exploits the fact that the words of a topic tend to co-occur at the document scale. This hypothesis is no longer valid for documents built according to the procedure of Choi. It is why we adapted his framework for having more realistic documents without losing its advantages. This adaptation concerns the way the document segments are selected. Instead of taking each segment from a different document, we only use two source documents. Each of them 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, *i.e.* one segment from a source document and the following from the other one, until 10 segments are extracted. Moreover, in order to be sure that the boundary between two adjacent segments of an evaluation document actually corresponds to a topic shift, the source documents are selected in such a way that they refer to different topics. This point was controlled in our case by taking documents from the corpus of the CLEF 2003 evaluation for crosslingual information retrieval: each evaluation document was built from two source documents that had been judged as relevant for two different CLEF 2003 topics. Two evaluation corpora made of 100 documents each, one in French and one in English, were built following this procedure. Table 1 shows their main characteristics.

	French	**English**
# source doc.	128	87
# source topics	11	3
segments/doc.	10 (84%) 8 (16%)	10 (97%) 8 (3%)
sentences/doc.	65	68
plain words/doc.	797	604

Table 1: Data about our evaluation corpora

[1]Each word of the topic vector has a weight equal to 1. In the window vector, this weight is equal to the frequency of the word in the corresponding side of the window.

5.2 Topic identification

As F06T exploits document topics, we also evaluated our method for topic identification. This evaluation is based on the corpus of the previous section. For each of its documents, a reference topic is built from each group of segments that come from the same source document by gathering the words that only appear in these segments. A reference topic is associated to the discovered topic that shares with it the largest number of words. Three complementary measures were computed to evaluate the quality of discovered topics. The main one is purity, which is classically used for unsupervised clustering:

$$Purity = \sum_{i=1}^{k} \frac{v_i}{V} P(Td_i) \tag{4}$$

where $P(Td_i)$, the purity of the discovered topic Td_i, is equal to the fraction of the vocabulary of Td_i that is part of the vocabulary of the reference topic Td_i is assigned to, V is the vocabulary of all the discovered topics and v_i is the vocabulary of Td_i. The second measure evaluates to what extent the reference topics are represented among the discovered topics and is equal to the ratio between the number of discovered topics that are assigned to a reference topic (*assigned discovered topics*) and the number of reference topics. The last measure estimates how strongly the vocabulary of reference topics is present among the discovered topics and is equal to the ratio between the size of the vocabulary of the assigned discovered topics and the size of the vocabulary of reference topics. Table 2 gives the mean

	purity	**reference topics (%)**	**ref. topic vocab. (%)**
French	0.771 (0.117)	89.5 (23.9)	29.9 (7.8)
English	0.766 (0.082)	99.0 (10.0)	31.6 (5.3)

Table 2: Evaluation of topic identification

of each measure, followed by its standard deviation. Results are globally similar for French and English. They show that our method for topic identification builds topics that are rather pure, *i.e.* each of them is strongly tied to a reference topic, but their content is rather sparse in comparison with the content of their associated reference topics.

5.3 Topic segmentation

For validating the hypothesis that underlies our work, we applied F06 and F06T to find the topic bounds in the documents of our two evaluation corpora. Moreover, we also tested four well known segmenters on our corpora to compare the results of F06 and F06T with state-of-the-art algorithms. We classically used the error metric P_k proposed in (Beeferman et al., 1999) to measure segmentation accuracy. P_k evaluates the probability that a randomly chosen pair of sentences, separated by k sentences, is wrongly classified, *i.e.* they are found in the same segment while they are actually in different ones (miss) or they are found in different segments while they are actually in the same one (false alarm). We also give the value of *WindowDiff* (WD), a variant of P_k proposed in (Pevzner and Hearst, 2002) that corrects some of its insufficiencies. Tables 3 and 4 show

systems	$\mathbf{P_k}$	$\mathbf{p_{val}}$**(F06)**	$\mathbf{p_{val}}$**(F06T)**	**WD**
U00	25.91	**0.003**	**1.3e-07**	27.42
C99	27.57	**4.2e-05**	**3.6e-10**	35.42
TextTiling*	21.08	0.699	**0.037**	27.43
LCseg	20.55	0.439	0.111	28.31
F06	21.58	/	**0.013**	27.83
F06T	*18.46*	**0.013**	/	*24.05*

Table 3: Evaluation of topic segmentation for the French corpus (P_k and WD as percentages)

the results of our evaluations for topic segmentation (smallest values are best results). U00 is the system described in (Utiyama and Isahara, 2001), C99 the one proposed in (Choi, 2000) and *LCseg* is presented in (Galley et al., 2003). *TextTiling** is a variant of *TextTiling* in which the final identification of topic shifts is taken from (Galley et al., 2003). All these systems were used as F06 and F06T without fixing the number of topic shifts to find. Moreover, their parameters were tuned for our evaluation corpus to obtain their best results. For each result, we also give the significance level p_{val} of its difference for P_k with F06 and F06T, evaluated by a one-side t-test with a null hypothesis of equal means. Levels lower than 0.05 are considered as statistically significant (bold-faced values). The first important point to notice about these tables is the fact that

systems	P_k	p_{val}(F06)	p_{val}(F06T)	WD
U00	19.42	**0.048**	**4.3e-05**	21.22
C99	21.63	**1.2e-04**	**1.8e-09**	30.64
TextTiling*	15.81	0.308	0.111	19.80
LCseg	14.78	**0.043**	0.496	19.73
F06	16.90	/	**0.010**	20.93
F06T	*14.06*	**0.010**	/	*18.31*

Table 4: Evaluation of topic segmentation for the English corpus (P_k and WD as percentages)

F06T has significantly better results than F06, both for French and English. Hence, it confirms our hypothesis about the interest of taking into account the topics of a text for its segmentation, even if these topics were discovered in an unsupervised way and without using external knowledge. Moreover, F06T have the best results among all the tested algorithms, with a significant difference in most of the cases.

Another notable point about these results is their stability across our two corpora, even if these corpora are quite similar. Whereas F06 and F06T were initially developed on a corpus in French, their results on the English corpus are comparable to their results on the French test corpus, both for the difference between them and the difference with the four other algorithms. The comparison with these algorithms also illustrates the relationships between them: *TextTiling**, *LCseg*, F06 and F06T share a large number of principles and their overall results are significantly higher than the results of U00 and C99. This trend is different from the one observed from the Choi corpus for which algorithms such C99 or U00 have good results (P_k for C99, U00, F06 and F06T is respectively equal to 12%, 10%, 14% and 14%). This means probably that algorithms with good results on a corpus built as the Choi corpus will not necessarily have good results on "true" texts, which agrees with (Georgescul et al., 2006). Finally, we can observe that all these algorithms have better results on the English corpus than on the French one. As the two corpora are quite similar, this difference seems to come from their difference of language, perhaps because repetitions are more discouraged in French than in English from a stylistic viewpoint. This tends to be confirmed by the ratio between the size of the lemmatized vocabulary of each corpus and their number of tokens, equal to 8% for the French corpus and to 5.6% for the English corpus.

6 Related Work

One of the main problems addressed by our work is the detection of the topical similarity of two text units. We have tackled this problem following an endogenous approach, which is new in the topic segmentation field to our knowledge. The main advantage of this option is that it does not require external knowledge. Moreover, it can integrate relations between words, such as proper nouns for instance, that are unlikely to be found in an external resource.

Other solutions have been already proposed to solve the problem we consider. Most of them consist of two steps: first, they automatically build a semantic representation of words from the co-occurrences collected from a large corpus; then, they use this representation for enhancing the representation of each text unit to compare. This overall principle is implemented with different forms by several topic segmenters. In CWM (Choi et al., 2001), 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 (Ponte and Croft, 1997), 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 (Caillet et al., 2004), 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. In fact, the way we use relations between words is closer to (Jobbins and Evett, 1998), even if the relations in this work come from a network of co-occurrences or a thesaurus rather than from text topics. In both cases the similarity of two text units is determined by the proportion of their words that are part of a relation across the two units.

More globally, our work exploits the topics of a text for its segmentation. This kind of approach was also explored in (Blei and Moreno, 2001) where probabilistic topic models were built in an unsupervised way. More recently, (Purver et al., 2006) has also proposed a method for unsupervised topic modeling to address both topic segmentation and identification. (Purver et al., 2006) is closer to our work than (Blei and Moreno, 2001) because it does not require to build topic models from a corpus but as in our case, its results do not outperform *LCseg* (Galley et al., 2003) while its model is far more complex.

7 Conclusion and Future Work

In this article, we have first proposed an unsupervised method for discovering the topics of a text without relying on external knowledge. Then, we have shown how these topics can be used for improving a topic segmentation method based on word reiteration. Moreover, we have proposed an adaptation of the evaluation framework of Choi that aims at building more realistic evaluation documents. Finally, we have demonstrated the interest of the method we present through its evaluation both on a French and an English corpus.

However, the solution we have proposed for improving the identification of topical similarities between text excerpts cannot completely make up for not using any external knowledge. Hence, we plan to use a network of lexical co-occurrences, which is a source of knowledge that is easy to build automatically from a large corpus. More precisely, we intend to extend our method for discovering text topics by combining the co-occurrence graph of a document with such a network. This network could also be used more directly for topic segmentation as in (Jobbins and Evett, 1998).

References

Doug Beeferman, Adam Berger, and John Lafferty. 1999. Statistical models for text segmentation. *Machine Learning*, 34(1):177–210.

David M. Blei and Pedro J. Moreno. 2001. Topic segmentation with an aspect hidden markov model. In *24th ACM SIGIR*, pages 343–348.

Marc Caillet, Jean-François Pessiot, Massih Amini, and Patrick Gallinari. 2004. Unsupervised learning with term clustering for thematic segmentation of texts. In *RIAO'04*, pages 1–11.

Freddy Y. Y. Choi, Peter Wiemer-Hastings, and Johanna Moore. 2001. Latent semantic analysis for text segmentation. In *EMNLP'01*, pages 109–117.

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

Levent Ertöz, Michael Steinbach, and Vipin Kuma. 2001. Finding topics in collections of documents: A shared nearest neighbor approach. In *Text Mine'01, Workshop of the 1st SIAM International Conference on Data Mining*.

Michel Galley, Kathleen McKeown, Eric Fosler-Lussier, and Hongyan Jing. 2003. Discourse segmentation of multi-party conversation. In *ACL'03*, pages 562–569.

Maria Georgescul, Alexander Clark, and Susan Armstrong. 2006. An analysis of quantitative aspects in the evaluation of thematic segmentation algorithms. In *7th SIGdial Workshop on Discourse and Dialogue*, pages 144–151.

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

Amanda C. Jobbins and Lindsay J. Evett. 1998. Text segmentation using reiteration and collocation. In *ACL-COLING'98*, pages 614–618.

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

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

Rebecca J. Passonneau and Diane J. Litman. 1997. Discourse segmentation by human and automated means. *Computational Linguistics*, 23(1):103–139.

Lev Pevzner and Marti A. Hearst. 2002. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36.

Jay M. Ponte and Bruce W. Croft. 1997. Text segmentation by topic. In *First European Conference on research and advanced technology for digital libraries*.

Matthew Purver, Konrad P. Körding, Thomas L. Griffiths, and Joshua B. Tenenbaum. 2006. Unsupervised topic modelling for multi-party spoken discourse. In *COLING-ACL 2006*, pages 17–24.

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

Masao Utiyama and Hitoshi Isahara. 2001. A statistical model for domain-independent text segmentation. In *ACL'01*, pages 491–498.

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