Statistical Models for Topic Segmentation

Jeffrey C. Reynar[1]
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052 USA
jreynar@microsoft.com

Abstract

Most documents are about more than one subject, but many NLP and IR techniques implicitly assume documents have just one topic. We describe new clues that mark shifts to new topics, novel algorithms for identifying topic boundaries and the uses of such boundaries once identified. We report topic segmentation performance on several corpora as well as improvement on an IR task that benefits from good segmentation.

Introduction

Dividing documents into topically-coherent sections has many uses, but the primary motivation for this work comes from information retrieval (IR). Documents in many collections vary widely in length and while the shortest may address one topic, modest length and long documents are likely to address multiple topics or be comprised of sections that address various aspects of the primary topic. Despite this fact, most IR systems treat documents as indivisible units and index them in their entirety.

This is problematic for two reasons. First, most relevance metrics are based on word frequency, which can be viewed as a function of the topic being discussed (Church and Gale, 1995). (For example, the word *header* is rare in general English, but it enjoys higher frequency in documents about soccer.) In general, word frequency is a good indicator of whether a document is relevant to a query, but consider a long document containing only one section relevant to a query. If a keyword is used only in the pertinent section, its overall frequency in the document will be low and, as a result, the document as a whole may be judged irrelevant despite the relevance of one section.

The second reason it would be beneficial to index sections of documents is that, once a search engine has identified a relevant document, users would benefit from direct access to the relevant sections. This problem is compounded when searching multimedia documents. If a user wants to find a particular news item in a database of radio or television news programs, they may not have the patience to suffer through a 30 minute broadcast to find the one minute clip that interests them.

Dividing documents into sections based on topic addresses both of these problems. IR engines can index the resulting sections just like documents and subsequently users can peruse those sections their search engine deems relevant. In the next section we will discuss the nature of our approach, then briefly describe previous work, discuss various indicators of topic shifts, outline novel algorithms based on them and present our results.

1 Our Approach

We treat the process of creating documents as an instance of the noisy channel model. In this idealization, prior to writing, the author has in mind a collection of disjoint topics that she intends to address. During the writing process, due to the goals of writing smooth prose and knitting her document into a coherent whole, she blurs the boundaries between these topics. Thus, we assume there is a correct segmentation that has been hidden from our view. Our goal, therefore, is to model the clues about the original segmentation that were not obliterated while writing.

We view segmentation as a labeling task. Given the text of a document and a collection of putative topic boundary locations—which could correspond to sentence boundaries, paragraph boundaries, pauses between utterances, changes in speaker or some arbitrary list of choice points—

[1] This work was conducted as part of my Ph.D. thesis work at the University of Pennsylvania.

we label each of them as either the location of a topic boundary or not. We perform this labeling using statistical algorithms that combine diverse sources of evidence to determine the likelihood of a topic boundary.

2 Previous Work

Much research has been devoted to the task of structuring text—that is dividing texts into units based on information within the text. This work falls roughly into two categories. Topic segmentation focuses on identifying topically-coherent blocks of text several sentences through several paragraphs in length (e.g. see Hearst, 1994). The prime motivation for identifying such units is to improve performance on language-processing or IR tasks. Discourse segmentation, on the other hand, is often finer-grained, and focuses on identifying relations between utterances (e.g. Grosz and Sidner, 1986 or Hirschberg and Grosz, 1992).

Many topic segmentations algorithms have been proposed in the literature. There is not enough space to review them all here, so we will focus on describing a representative sample that covers most of the features used to predict the location of boundaries. See (Reynar, 1998) for a more thorough review.

Youmans devised a technique called the Vocabulary Management Profile based on the location of first uses of word types. He posited that large clusters of first uses frequently followed topic boundaries since new topics generally introduce new vocabulary items (Youmans, 1991).

Morris and Hirst developed an algorithm (Morris and Hirst, 1991) based on lexical cohesion relations (Halliday and Hasan, 1976). They used Roget's 1977 Thesaurus to identify synonyms and other cohesion relations.

Kozima defined a measure called the Lexical Cohesion Profile (LCP) based on spreading activation within a semantic network derived from a machine-readable dictionary. He identified topic boundaries where the LCP score was low (Kozima, 1993).

Hearst developed a technique called TextTiling that automatically divides expository texts into multi-paragraph segments using the vector space model from IR (Hearst, 1994). Topic boundaries were positioned where the similarity between the block of text before and after the boundary was low.

In previous work (Reynar, 1994), we described a method of finding topic boundaries using an optimisation algorithm based on word repetition that was inspired by a visualization technique known as dotplotting (Helfman, 1994).

Ponte and Croft predict topic boundaries using a model of likely topic length and a query expansion technique called Local Content Analysis that maps sets of words into a space of concepts (Ponte and Croft, 1997).

Richmond, Smith and Amitay designed an algorithm for topic segmentation that weighted words based on their frequency within a document and subsequently used these weights in a formula based on the distance between repetitions of word types (Richmond *et al.*, 1997).

Beeferman, Berger and Lafferty used the relative performance of two statistical language models and cue words to identify topic boundaries (Beeferman *et al.*, 1997).

3 New Clues for Topic Segmentation

Prior work on topic segmentation has exploited many different hints about where topic boundaries lie. The algorithms we present use many cues from the literature as well as novel ones. Our approach is statistical in nature and weights evidence based on its utility in segmenting a training corpus. As a result, we do not use clues to form hard and fast rules. Instead, they all contribute evidence used to either increase or decrease the likelihood of proposing a topic boundary between two regions of text.

3.1 Domain-specific Cue Phrases

Many discourse segmentation techniques (e.g. Hirschberg and Litman, 1993) as well as some topic segmentation algorithms rely on cue words and phrases (e.g. Beeferman *et al., 1997)*, but the types of cue words used vary greatly. Those we employ are highly domain specific. Taking an

example from the broadcast news domain where we will demonstrate the effectiveness of our algorithms, the phrase *joining us* is a good indicator that a topic shift has just occurred because news anchors frequently say things such as *joining us to discuss the crisis in Kosovo is Congressman...* when beginning new stories. Consequently, our algorithms use the presence of phrases such as this one to boost the probability of a topic boundary having occurred.

joining us
good evening
brought to you by
this just in
welcome back
<person name> <station>
this is <person name>

Table 1: A sampling of domain-specific cue phrases we employ.

Some cue phrases are more complicated and contain word sequences of particular types. Not surprisingly, the phrase *this is* is common in broadcast news. When it is followed by a person's name, however, it serves as a good clue that a topic is about to end. *This is* <person name> is almost always said when a reporter is signing off after finishing an on-location report. Generally such signoffs are followed by the start of new news stories. A sampling of the cue phrases we use is found in Table 1. Since our training corpus was relatively small we identified these by hand, but on a different corpus we induced them automatically (Reynar, 1998). The results we present later in the paper rely solely on manually identified cues phrases.

Identifying complex cue phrases involves pattern matching and determining whether particular word sequences belong to various classes. To address this, we built a named entity recognition system in the spirit of those used for the Message Understanding Conference evaluations (e.g. Bikel *et al.*, 1997). Our named entity recognizer used a maximum entropy model built with Adwait Ratnaparkhi's tools (Ratnaparkhi, 1996) to label word sequences as either person, place, company or none of the above based on local cues including the surrounding words and whether honorifics (e.g. *Mrs.* or *Gen.*) or corporate designators (e.g. *Corp.* or *Inc.*) were present. Our algorithm's labelling accuracy of 96.0% by token was sufficient for our purposes, but performance is not directly comparable to the MUC competitors'. Though we trained from the same data, we preprocessed the data to remove punctuation and capitalization so the model could be applied to broadcast news data that lacked these helpful clues. We separately identified television network acronyms using simple regular expressions.

3.2 Word Bigram Frequency

Many topic segmentation algorithms in the literature use word frequency (e.g. Hearst, 1994; Reynar, 1994; Beeferman *et al.*, 1997). An obvious extension to using word frequency is to use the frequency of multi-word phrases. Such phrases are useful because they approximate word sense disambiguation techniques. Algorithms that rely exclusively on word frequency might be fooled into suggesting that two stretches of text containing the word *plant* were part of the same story simply because of the rarity of *plant* and the low odds that two adjacent stories contained it due to chance. However, if *plant* in one section participated in bigrams such as *wild plant, native plant* and *woody plant* but in the other section was only in the bigrams *chemical plant, manufacturing plant* and *processing plant,* the lack of overlap between sets of bigrams could be used to decrease the probability that the two sections of text were in the same story. We limited the bigrams we used to those containing two content words.

3.3 Repetition of Named Entities

The named entities we identified for use in cue phrases are also good indicators of whether two sections are likely to be in the same story or not. Companies, people and places figure prominently in many documents, particularly those in the domain of broadcast news. The odds that different stories discuss the same entities are generally low. There are obviously exceptions—the President of the U.S. may figure in many stories in a single broadcast—but nonetheless the presence of the same entities in two blocks of text suggest that they are likely to be part of the same story.

3.4 Pronoun Usage

In her dissertation, Levy described a study of the impact of the type of referring expressions used, the location of first mentions of people and the gestures speakers make upon the cohesiveness of

discourse (Levy, 1984). She found a strong correlation between the types of referring expressions people used, in particular how explicit
they were, and the degree of cohesiveness with the preceding context. Less cohesive utterances generally contained more explicit referring expressions, such as definite noun phrases or phrases consisting of a possessive followed by a noun, while more cohesive utterances more frequently contained zeroes and pronouns.

We will use the converse of Levy's observation about pronouns to gauge the likelihood of a topic shift. Since Levy generally found pronouns in utterances that exhibited a high degree of cohesion with the prior context, we assume that the presence of a pronoun among the first words immediately following a putative topic boundary provides some evidence that no topic boundary actually exists there.

4 Our Algorithms

We designed two algorithms for topic segmentation. The first is based solely on word frequency and the second combines the results of the first with other sources of evidence. Both of these algorithms are applied to text following some preprocessing including tokenization, conversion to lowercase and the application of a lemmatizer (Karp *et al.*, 1992).

4.1 Word Frequency Algorithm

Our word frequency algorithm uses Katz's G model (Katz, 1996). The G model stipulates that words occur in documents either topically or non-topically. The model defines topical words as those that occur more than 1 time, while non-topical words occur only once. Counterexamples of these uses of topical and nontopical, of course, abound.

We use the G model, shown below, to determine the probability that a particular word, w, occurred k times in a document. We trained the model from a corpus of 78 million words of *Wall Street Journal* text and smoothed the parameters using Dan Melamed's implementation of Good-Turing smoothing (Gale and Sampson, 1995) and additional *ad hoc* smoothing to account for unknown words.

$$\Pr(k,w) = (1-\alpha_w)\delta_{k,0} + \alpha_w(1-\gamma_w)\delta_{k,1} + (\frac{\alpha_w\gamma_w}{B_w-1}(1-\frac{1}{B_w-1})^{k-2})(1-\delta_{k,0}-\delta_{k,1})$$

α_w is the probability that a document contains at least 1 occurrence of word w.

γ_w is the probability that w is used topically in a document given that it occurs at all.

B_w is the average number of occurrences in documents with more than 1 occurrence of w.

$\delta_{x,y}$ is a function with value 1 if x = y and 0 otherwise.

The simplest way to view the G model is to decompose it into 3 separate terms that are summed. The first term is the probablility of zero occurrences of a word, the second is the probability of one occurrence and the third is the probability of any number of occurrences greater than one.

To detect topic boundaries, we used the model to answer this simple question. Is it more or less likely that the words following a putative topic boundary were generated independently of those before it?

Given a potential topic boundary, we call the text before the boundary region 1 and the text after it region 2. For the sake of our algorithm, the size of these regions was fixed at 230 words—the average size of a topic segment in our training corpus, 30 files from the HUB-4 Broadcast News Corpus annotated with topic boundaries by the LDC (HUB-4, 1996). Since the G model, unlike language models used for speech recognition, computes the probability of a bag of words rather than a word sequence, we can use it to compute the probability of some text given knowledge of what words have occurred before that text. We computed two probabilities with the model. P_{one} is the probability that region 1 and region 2 discuss the same subject matter and hence that there is no topic boundary between them. P_{two} is the probability that they discuss different subjects and are separated by a topic boundary. P_{one}, therefore, is the probability of seeing the words in region 2 given the context, called C, of region 1. P_{two} is the

probability of seeing the words in region 2 independent of the words in region 1. Formulae for P_{one} and P_{two} are shown below. Boundaries were placed where P_{two} was greater than P_{one} by a certain threshold. The threshold was used to trade precision for recall and vice versa when identifying topic boundaries. The most natural threshold is a very small nonzero value, which is equivalent to placing a boundary wherever P_{two} is greater than P_{one}.

$$P_{one} = \prod_{w} \Pr(k, w \mid C) \qquad P_{two} = \prod_{w} \Pr(k, w)$$

Computing P_{two} is straightforward, but P_{one} requires computing conditional probabilities of the number of occurrences of each word in region 2 given the number in region 1. The formulae for the conditional probabilities are shown in Table 2. We do not have space to derive these formulae here, but they can be found in (Reynar, 1998). M is a normalizing term required to make the conditional probabilities sum to 1. In the table, x+ means x occurrences or more.

Occurrences in region 1	Occurrences in region 2	Conditional probability
0	0	$1 - \alpha$
0	1	$\alpha(1-\gamma)$
0	2+	$\frac{\alpha-\gamma}{B-1}(1-\frac{1}{B-1})^{k-2}$
1	0	$1-\gamma$
1	1+	$\frac{\gamma}{B-1}(1-\frac{1}{B-1})^{k-2}$
2+	0+	$\frac{1}{M(B-1)}(1-\frac{1}{B-1})^{k-2}$

Table 2: Conditional probabilities used to compute P_{one}.

4.2 A Maximum Entropy Model

Our second algorithm is a maximum entropy model that uses these features:

- Did our word frequency algorithm suggest a topic boundary?
- Which domain cues (such as *Joining us* or *This is* <person>) were present?
- How many content word bigrams were common to both regions adjoining the putative topic boundary?
- How many named entities were common to both regions?
- How many content words in both regions were synonyms according to WordNet (Miller *et al.*, 1990)?
- What percentage of content words in the region after the putative boundary were first uses?
- Were pronouns used in the first five words after the putative topic boundary?

We trained this model from 30 files of HUB-4 data that was disjoint from our test data.

5 Evaluation

We will present results for broadcast news data and for identifying chapter boundaries labelled by authors.

5.1 HUB-4 Corpus Performance

Table 3 shows the results of segmenting the test portion of the HUB-4 corpus, which consisted of transcribed broadcasts divided into segments by the LDC. We measured performance by comparing our segmentation to the gold standard annotation produced by the LDC.

The row labelled Random guess shows the performance of a baseline algorithm that randomly guessed boundary locations with probability equal to the fraction of possible boundary sites that were boundaries in the gold standard. The row TextTiling shows the performance of the publicly available version of that algorithm (Hearst, 1994). Optimization is the algorithm we proposed in (Reynar, 1994). Word frequency and Max. Ent. Model are the algorithms we described above. Our word frequency algorithm does better than chance, TextTiling and our previous work and our maximum entropy model does better still. See (Reynar, 1998) for graphs showing the effects of trading precision for recall with these models.

Algorithm	Precision	Recall
Random guess	0.16	0.16
TextTiling	0.21	0.41
Optimization	0.36	0.20
Word Frequency	0.55	0.52
Max. Ent. Model	**0.59**	**0.60**

Table 3: Performance on the HUB-4 English corpus.

We also tested our models on speech-recognized broadcasts from the 1997 TREC spoken document retrieval corpus. We did not have sufficient data to train the maximum entropy model, but our word frequency algorithm achieved precision of 0.36 and recall of 0.52, considerably better than the baseline of 0.19 precision and recall. Using manually produced transcripts of the same data naturally yielded better performance—precision was 0.50 and recall 0.58.

Our performance on broadcast data was surprisingly good considering we trained the word frequency model from newswire data. Given a large corpus of broadcast data, we expect our algorithms would perform even better.

We were curious, however, how much of the performance was attributable to having numerous parameters (3 per word) in the G model and how much comes from the nature of the model. To address this, we discarded the α, γ and B parameters particular to each word and instead used the same parameter values for each word—namely, those assigned to unknown words through our smoothing process. This reduced the number of parameters from 3 per word to only 3 parameters total. Performance of this hobbled version of our word frequency algorithm was so good on the HUB-4 English corpus—achieving precision of 0.42 and recall of 0.50—that we tested it on Spanish broadcast news data from the HUB-4 corpus. Even for that corpus we found much better than baseline performance. Baseline for Spanish was precision and recall of 0.28, yet our 3-parameter word frequency model achieved 0.50 precision and recall of 0.62. To reiterate, we used our word frequency model with a total of 3 parameters trained from English newswire text to segment Spanish broadcast news data

We believe that the G model, which captures the notion of burstiness very well, is a good model for segmentation. However, the more important lesson from this work is that the concept of burstiness alone can be used to segment texts. Segmentation performance is better when models have accurate measures of the likelihood of 0, 1 and 2 or more occurrences of a word. However, the mere fact that content words are bursty and are relatively unlikely to appear in neighboring regions of a document unless those two regions are about the same topic is sufficient to segment many texts. This explains our ability to segment Spanish broadcast news using a 3 parameter model trained from English newswire data.

5.2 Recovering Authorial Structure

Authors endow some types of documents with structure as they write. They may divide documents into chapters, chapters into sections, sections into subsections and so forth. We exploited these structures to evaluate topic segmentation techniques by comparing algorithmic determinations of structure to the author's original divisions. This method of evaluation is especially useful because numerous documents are now available in electronic form.

We tested our word frequency algorithm on four randomly selected texts from Project Gutenberg. The four texts were Thomas Paine's pamphlet *Common Sense* which was published in 1791, the first volume of *Decline and Fall of the Roman Empire* by Edward Gibbon, G.K. Chesterton's book *Orthodoxy* and Herman Melville's classic *Moby Dick*. We permitted the algorithm to guess boundaries only between paragraphs, which were marked by blank lines in each document.

To assess performance, we set the number of boundaries to be guessed to the number the authors themselves had identified. As a result, this evaluation focuses solely on the algorithm's ability to rank candidate boundaries and not on its adeptness at determining how many boundaries to select. To evaluate performance, we computed the accuracy of the algorithm's guesses compared to the chapter boundaries the authors identified. The documents we used for this evaluation may have contained legitimate topic boundaries which did not correspond to chapter boundaries, but we scored guesses at those boundaries incorrect.

Table 4 presents results for the four works. Our algorithm performed better than randomly assigning boundaries for each of the documents except the pamphlet *Common Sense*. Performance on the other three works was significantly better than chance and ranged from an improvement of a factor of three in accuracy over the baseline to a factor of nearly 9 for the lengthy *Decline and Fall of the Roman Empire*.

Work	# of Boundaries	Word Frequency	Random
Common Sense	7	0.00	0.36
Decline and Fall	53	0.21	0.0024
Moby Dick	132	0.55	0.173
Orthodoxy	8	0.25	0.033
Combined	200	0.43	0.059

Table 4: Accuracy of the Word Frequency algorithm on identifying chapter boundaries.

5.3 IR Task Performance

The data from the HUB-4 corpus was also used for the TREC Spoken document retrieval task. We tested the utility of our segmentations by comparing IR performance when we indexed documents, the segments annotated by the LDC and the segments identified by our algorithms. We modified SMART (Buckley, 1985) to perform better normalization for variations in document length (Singhal *et al.*, 1996) prior to conducting our IR experiments.

This IR task is atypical in that there is only 1 relevant document in the collection for each query. Consequently, performance is measured by determining the average rank determined by the IR system for the document relevant to each query. Perfect performance would be an average rank of 1, hence lower average ranks are better. Table 5 presents our results. Note that indexing the segments identified by our algorithms was better than indexing entire documents and that our best algorithm even outperformed indexing the gold standard annotation produced by the LDC.

Method	Average Rank
Documents	9.52
Annotator segments	8.42
Word frequency model	9.48
Max. Ent. Model	**7.54**

Table 5: Performance on an IR task. Lower numbers are better.

Conclusion

We described two new algorithms for topic segmentation. The first, based solely on word frequency, performs better than previous algorithms on broadcast news data. It performs well on speech recognized English despite recognition errors. Most surprisingly, a version of our first model that requires little training data could segment Spanish broadcast news documents as well—even with parameters estimated from English documents. Our second technique, a statistical model that combined numerous clues about segmentation, performs better than the first, but requires segmented training data.

We showed an improvement on a simple IR task to demonstrate the potential of topic segmentation algorithms for improving IR. Other potential uses of these algorithms include better language modeling by building topic-based language models, improving NLP algorithms (e.g. coreference resolution), summarization, hypertext linking (Salton and Buckley, 1992), automated essay grading (Burstein *et al.*, 1997) and topic detection and tracking (TDT program committee, 1998). Some of these are discussed in (Reynar, 1998), and others will be addressed in future work.

Acknowledgements

My thanks to the anonymous reviewers and the members of my thesis committee, Mitch Marcus, Aravind Joshi, Mark Liberman, Julia Hirschberg and Lyle Ungar for useful feedback. Thanks also to Dan Melamed for use of his smoothing tools and to Adwait Ratnaparkhi for use of his maximum entropy modelling software.

References

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*, pages 35-46, Providence, Rhode Island.

Bikel, D.M., Miller, S., Schwartz, R., and Weischedel, R. (1997). Nymble: a high-performance learning name-finder. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 194-201, Washington, D.C.

Buckley, C. (1985). Implementation of the SMART information retrieval system. Technical Report Technical Report 85-686, Cornell University.

Burstein, J., Wolff, S., Lu, C., and Kaplan, R. (1997). An automatic scoring system for advanced placement biology essays. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 174-181, Washington, D.C.

Church, K.W. and Gale, W.A. (1995). Inverse document frequency (IDF): A measure of deviations from Poisson. In Yarowsky, D. and Church, K., editors, *Proceedings of the Third Workshop on Very Large Corpora*, pages 121-130. Association for Computational Linguistics.

Gale, W. and Sampson, G. (1995). Good-Turing smoothing without tears. *Journal of Quantitative Linguistics*, 2.

Grosz, B. J. and Sidner, C.L. (1986). Attention, Intentions and the Structure of Discourse. *Computational Linguistics,* 12 (3): 175-204.

Halliday, M. and Hasan, R. (1976). *Cohesion in English*. Longman Group, New York.

Hearst, M.A. (1994). Multi-paragraph segmentation of expository text. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 9-16, Las Cruces, New Mexico.

Helfman, J.I. (1994). Similarity patterns in language. In *IEEE Symposium on Visual Languages*.

Hirschberg, J. and Grosz, B. (1992). Intonational features of local and global discourse. In *Proceedings of the Workshop on Spoken Language Systems*, pages 441-446. DARPA.

Hirschberg, J. and Litman, D. (1993). Empirical studies on the disambiguation of cue phrases. *Computational Linguistics*, 19(3):501-530.

HUB-4 Program Committee (1996). The 1996 HUB-4 annotation specification for evaluation of speech recognition on broadcast news, version 3.5.

Karp, D., Schabes, Y., Zaidel, M. and Egedi, D. (1992). A Freely Available Wide Coverage Morphological Analyzer for English. *Proceedings of the 15th International Conference on Computational Linguistics.* Nantes, France.

Katz, S.M. (1996). Distribution of content words and phrases in text and language modeling. *Natural Language Engineering*, 2(1):15-59.

Kozima, H. (1993). Text segmentation based on similarity between words. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Student Session, pages 286-288.

Levy, E.T. (1984). Communicating Thematic Structure in Narrative Discourse: The Use of Referring Terms and Gestures. Ph.D. thesis, University of Chicago.

Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. (1990). Five papers on WordNet. Technical report, Cognitive Science Laboratory, Princeton University.

Morris, J. and Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1):21-42.

Ponte, J.M. and Croft, W.B. (1997). Text segmentation by topic. In *European Conference on Digital Libraries*, pages 113-125, Pisa, Italy.

Ratnaparkhi, A. (1996). A maximum entropy model for part-of-speech tagging. In *Proceedings of the First Conference on Empirical Methods in Natural Language Processing*, pages 133-142, University of Pennsylvania.

Reynar, J.C. (1994). An automatic method of finding topic boundaries. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, Student Session, pages 331-333, Las Cruces, New Mexico.

Reynar, J.C. (1998). Topic Segmentation: Algorithms and Applications. Ph.D. thesis, University of Pennsylvania, Department of Computer Science.

Richmond, K., Smith, A., and Amitay, E. (1997). Detecting subject boundaries within text: A language independent statistical approach. In *Exploratory Methods in Natural Language Processing*, pages 47-54, Providence, Rhode Island.

Salton, G. and Buckley, C. (1992). Automatic text structuring experiments. In Jacobs, P.S., editor, *Text-Based Intelligent Systems: Current Research and Practice in Information Extraction and Retrieval*, pages 199-210. Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Singhal, A., Buckley, C., and Mitra, M. (1996). Pivoted document length normalization. In *Proceedings of the ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 21-29, Zurich, Switzerland. ACM.

TDT Program Committee (1998). Topic Detection and Tracking Phase 2 Evaluation Plan, version 2.1.

Youmans, G. (1991). A new tool for discourse analysis: The vocabulary management profile. *Language*, 67(4):763-789.