Feature-Based Segmentation of Narrative Documents

David Kauchak
Palo Alto Research Center and
University of California, San Diego
San Diego, CA 92093
dkauchak@cs.ucsd.edu

Francine Chen
Palo Alto Research Center
3333 Coyote Hill Rd.
Palo Alto, CA 94304
fchen@parc.com

Abstract

In this paper we examine topic segmentation of narrative documents, which are characterized by long passages of text with few headings. We first present results suggesting that previous topic segmentation approaches are not appropriate for narrative text. We then present a feature-based method that combines features from diverse sources as well as learned features. Applied to narrative books and encyclopedia articles, our method shows results that are significantly better than previous segmentation approaches. An analysis of individual features is also provided and the benefit of generalization using outside resources is shown.

1 Introduction

Many long text documents, such as magazine articles, narrative books and news articles contain few section headings. The number of books in narrative style that are available in digital form is rapidly increasing through projects such as Project Gutenberg and the Million Book Project at Carnegie Mellon University. Access to these collections is becoming easier with directories such as the Online Books Page at the University of Pennsylvania.

As text analysis and retrieval moves from retrieval of documents to retrieval of document passages, the ability to segment documents into smaller, coherent regions enables more precise retrieval of meaningful portions of text (Hearst, 1994) and improved question answering. Segmentation also has applications in other areas of information access, including document navigation (Choi, 2000), anaphora and ellipsis resolution, and text summarization (Kozima, 1993).

Research projects on text segmentation have focused on broadcast news stories (Beeferman et al., 1999), expository texts (Hearst, 1994) and synthetic texts (Li and Yamanishi, 2000; Brants et al., 2002). Broadcast news stories contain cues that are indicative of a new story, such as “coming up”, or phrases that introduce a reporter, which are not applicable to written text. In expository texts and synthetic texts, there is repetition of terms within a topical segment, so that the similarity of “blocks” of text is a useful indicator of topic change. *Synthetic* texts are created by concatenating stories, and exhibit stronger topic changes than the subtopic changes within a document; consequently, algorithms based on the similarity of text blocks work well on these texts.

In contrast to these earlier works, we present a method for segmenting *narrative* documents. In this domain there is little repetition of words and the segmentation cues are weaker than in broadcast news stories, resulting in poor performance from previous methods.

We present a feature-based approach, where the features are more strongly engineered using linguistic knowledge than in earlier approaches. The key to most feature-based approaches, particularly in NLP tasks where there is a broad range of possible feature sources, is identifying appropriate features. Selecting features in this domain presents a number of interesting challenges. First, features used in previous methods are not sufficient for solving this problem. We explore a number of different sources of information for extracting features, many previously unused. Second, the sparse nature of text and the high

Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in NLP, pages 32–39,
Ann Arbor, June 2005. ©2005 Association for Computational Linguistics

cost of obtaining training data requires generalization using outside resources. Finally, we incorporate features from non-traditional resources such as lexical chains where features must be extracted from the underlying knowledge representation.

2 Previous Approaches

Previous topic segmentation methods fall into three groups: similarity based, lexical chain based, and feature based. In this section we give a brief overview of each of these groups.

2.1 Similarity-based

One popular method is to generate similarities between blocks of text (such as blocks of words, sentences or paragraphs) and then identify section boundaries where dips in the similarities occur.

The cosine similarity measure between term vectors is used by Hearst (1994) to define the similarity between blocks. She notes that the largest dips in similarity correspond to defined boundaries. Brants et al. (2002) learn a PLSA model using EM to smooth the term vectors. The model is parameterized by introducing a latent variable, representing the possible "topics". They show good performance on a number of different synthetic data sets.

Kozima and Furugori (1994) use another similarity metric they call "lexical cohesion". The "cohesiveness" of a pair of words is calculated by spreading activation on a semantic network as well as word frequency. They showed that dips in lexical cohesion plots had some correlation with human subject boundary decisions on one short story.

2.2 Lexical Chains

Semantic networks define relationships between words such as synonymy, specialization/generalization and part/whole. Stokes et al. (2002) use these relationships to construct lexical chains. A lexical chain is a sequence of lexicographically related word occurrences where every word occurs within a set distance from the previous word. A boundary is identified where a large numbers of lexical chains begin and end. They showed that lexical chains were useful for determining the text structure on a set of magazine articles, though they did not provide empirical results.

2.3 Feature-based

Beeferman et al. (1999) use an exponential model and generate features using a maximum entropy selection criterion. Most features learned are cue-based features that identify a boundary based on the occurrence of words or phrases. They also include a feature that measures the difference in performance of a "long range" vs. "short range" model. When the short range model outperforms the long range model, this indicates a boundary. Their method performed well on a number of broadcast news data sets, including the CNN data set from TDT 1997.

Reynar (1999) describes a maximum entropy model that combines hand selected features, including: broadcast news domain cues, number of content word bigrams, number of named entities, number of content words that are WordNet synonyms in the left and right regions, percentage of content words in the right segment that are first uses, whether pronouns occur in the first five words, and whether a word frequency based algorithm predicts a boundary. He found that for the HUB-4 corpus, which is composed of transcribed broadcasts, that the combined feature model performed better than TextTiling.

Mochizuki et al. (1998) use a combination of linguistic cues to segment Japanese text. Although a number of cues do not apply to English (e.g., topical markers), they also use anaphoric expressions and lexical chains as cues. Their study was small, but did indicate that lexical chains are a useful cue in some domains.

These studies indicate that a combination of features can be useful for segmentation. However, Mochizuki et al. (1998) analyzed Japanese texts, and Reynar (1999) and Beeferman et al. (1999) evaluated on broadcast news stories, which have many cues that narrative texts do not. Beeferman et al. (1999) also evaluated on concatenated Wall Street Journal articles, which have stronger topic changes than within a document. In our work, we examine the use of linguistic features for segmentation of narrative text in English.

3 Properties of Narrative Text

Characterizing data set properties is the first step towards deriving useful features. The approaches in the previous section performed well on broad-

Table 1: Previous approaches evaluated on narrative data from *Biohazard*

Model	Word Error	Sent. Error	Window Diff
random	0.486	0.490	0.541
TextTiling	0.481	0.497	0.526
PLSA	0.480	0.521	0.559

cast news, expository and synthetic data sets. Many properties of these documents are not shared by narrative documents. These properties include: 1) cue phrases, such as “welcome back” and “joining us” that feature-based methods used in broadcast news, 2) strong topic shifts, as in synthetic documents created by concatenating newswire articles, and 3) large data sets such that the training data and testing data appeared to come from similar distributions.

In this paper we examine two narrative-style books: *Biohazard* by Ken Alibek and *The Demon in the Freezer* by Richard Preston. These books are segmented by the author into sections. We manually examined these author identified boundaries and they are reasonable. We take these sections as true locations of segment boundaries. We split *Biohazard* into three parts, two for experimentation (exp1 and exp2) and the third as a holdout for testing. *Demon in the Freezer* was reserved for testing. *Biohazard* contains 213 true and 5858 possible boundaries. *Demon* has 119 true and 4466 possible boundaries. Locations between sentences are considered possible boundaries and were determined automatically.

We present an analysis of properties of the book *Biohazard* by Ken Alibek as an exemplar of narrative documents (for this section, test=exp1 and train=exp2). These properties are different from previous expository data sets and will result in poor performance for the algorithms mentioned in Section 2. These properties help guide us in deriving features that may be useful for segmenting narrative text.

Vocabulary The book contains a single topic with a number of sub-topics. These changing topics, combined with the varied use of words for narrative documents, results in many unseen terms in the test set. 25% of the content words in the test set do not occur in the training set and a third of the words in the test set occur two times or less in the training set. This causes problems for those methods that learn a model of the training data such as Brants et al. (2002) and Beeferman et al. (1999) because, without outside resources, the information in the training data is not sufficient to generalize to the test set.

Boundary words Many feature-based methods rely on cues at the boundaries (Beeferman et al., 1999; Reynar, 1999). 474 content terms occur in the first sentence of boundaries in the training set. Of these terms, 103 occur at the boundaries of the test set. However, of those terms that occur *significantly* at a training set boundary (where significant is determined by a likelihood-ratio test with a significance level of 0.1), only 9 occur at test boundaries. No words occur significantly at a training boundary AND also significantly at a test boundary.

Segment similarity Table 1 shows that two similarity-based methods that perform well on synthetic and expository text perform poorly (i.e., on par with random) on *Biohazard*. The poor performance occurs because block similarities provide little information about the actual segment boundaries on this data set. We examined the average similarity for two adjacent regions within a segment versus the average similarity for two adjacent regions that cross a segment boundary. If the similarity scores were useful, the within segment scores would be higher than across segment scores. Similarities were generated using the PLSA model, averaging over multiple models with between 8 and 20 latent classes. The average similarity score within a segment was 0.903 with a standard deviation of 0.074 and the average score across a segment boundary was 0.914 with a standard deviation of 0.041. In this case, the across boundary similarity is actually higher. Similar values were observed for the cosine similarities used by the TextTiling algorithm, as well as with other numbers of latent topics for the PLSA model. For all cases examined, there was little difference between inter-segment similarity and across-boundary similarity, and there was always a large standard deviation.

Lexical chains Lexical chains were identified as synonyms (and exact matches) occurring within a distance of one-twentieth the average segment length and with a maximum chain length equal to the average segment length (other values were examined with similar results). Stokes et al. (2002) suggest that high concentrations of lexical chain beginnings and endings are indicative of a boundary location. On the narrative data, of the 219 overall chains, only 2 begin at a boundary and only 1 ends at a boundary. A more general heuristic identifies boundaries where there is an increase in the number of chains beginning and ending near a possible boundary while also minimizing chains that span boundaries. Even this heuristic does not appear indicative on this data set. Over 20% of the chains actually cross segment boundaries. We also measured the average distance from a boundary and the nearest beginning and ending of a chain if a chain begins/ends within that segment. If the chains are a good feature, then these should be relatively small. The average segment length is 185 words, but the average distance to the closest beginning chain is 39 words away and closest ending chain is 36 words away. Given an average of 4 chains per segment, the beginning and ending of chains were not concentrated near boundary locations in our narrative data, and therefore not indicative of boundaries.

4 Feature-Based Segmentation

We pose the problem of segmentation as a classification problem. Sentences are automatically identified and each boundary between sentences is a possible segmentation point. In the classification framework, each segmentation point becomes an example. We examine both support vector machines (SVMlight (Joachims, 1999)) and boosted decision stumps (Weka (Witten and Frank, 2000)) for our learning algorithm. SVMs have shown good performance on a variety of problems, including natural language tasks (Cristianini and Shawe-Taylor, 2000), but require careful feature selection. Classification using boosted decisions stumps can be a helpful tool for analyzing the usefulness of individual features. Examining multiple classification methods helps avoid focusing on the biases of a particular learning method.

4.1 Example Reweighting

One problem with formulating the segmentation problem as a classification problem is that there are many more negative than positive examples. To discourage the learning algorithm from classifying all results as negative and to instead focus on the positive examples, the training data must be reweighted.

We set the weight of positive vs. negative examples so that the number of boundaries after testing agrees with the expected number of segments based on the training data. This is done by iteratively adjusting the weighting factor while re-training and retesting until the predicted number of segments on the test set is approximately the expected number. The expected number of segments is the number of sentences in the test set divided by the number of sentences per segment in the training data. This value can also be weighted based on prior knowledge.

4.2 Preprocessing

A number of preprocessing steps are applied to the books to help increase the informativeness of the texts. The book texts were obtained using OCR methods with human correction. The text is preprocessed by tokenizing, removing stop words, and stemming using the Inxight LinguistiX morphological analyzer. Paragraphs are identified using formatting information. Sentences are identified using the TnT tokenizer and parts of speech with the TnT part of speech tagger (Brants, 2000) with the standard English *Wall Street Journal* n-grams. Named entities are identified using finite state technology (Beesley and Karttunen, 2003) to identify various entities including: person, location, disease and organization. Many of these preprocessing steps help provide salient features for use during segmentation.

4.3 Engineered Features

Segmenting narrative documents raises a number of interesting challenges. First, labeling data is extremely time consuming. Therefore, outside resources are required to better generalize from the training data. WordNet is used to identify words that are similar and tend to occur at boundaries for the "word group" feature. Second, some sources of information, in particular entity chains, do not fit into the standard feature based paradigm. This requires extracting features from the underlying information source. Extracting these features represents a tradeoff between information content and generalizability. In the case of entity chains, we extract features that characterize the occurrence distribution of the entity chains. Finally, the "word groups" and "entity groups" feature groups generate candidate features and a selection process is required to select useful features. We found that a likelihood ratio test for significance worked well for identifying those features that would be useful for classification. Throughout this section, when we use the term "significant" we are referring to significant with respect to the likelihood ratio test (with a significance level of 0.1).

We selected features both a priori and dynamically during training (i.e., word groups and entity groups are selected dynamically). Feature selection has been used by previous segmentation methods (Beeferman et al., 1999) as a way of adapting better to the data. In our approach, knowledge about the task is used more strongly in defining the feature types, and the selection of features is performed prior to the classification step. We also used mutual information, statistical tests of significance and classification performance on a development data set to identify useful features.

Word groups In Section 3 we showed that there are not consistent cue phrases at boundaries. To generalize better, we identify word *groups* that occur significantly at boundaries. A word group is all words that have the same parent in the WordNet hierarchy. A binary feature is used for each learned group based on the occurrence of at least one of the words in the group. Groups found include months, days, temporal phrases, military rankings and country names.

Entity groups For each entity group (i.e. named entities such as person, city, or disease tagged by the named entity extractor) that occurs significantly at a boundary, a feature indicating whether or not an entity of that group occurs in the sentence is used.

Full name The named entity extraction system tags persons named in the document. A rough co-reference resolution was performed by grouping together references that share at least one token (e.g., "General Yury Tikhonovich Kalinin" and "Kalinin"). The full name of a person is the longest reference of a group referring to the same person. This feature indicates whether or not the sentence contains a full name.

Entity chains Word relationships work well when the documents have disjoint topics; however, when topics are similar, words tend to relate too easily. We propose a more stringent chaining method called entity chains. Entity chains are constructed in the same fashion as lexical chains, except we consider named entities. Two entities are considered related (i.e. in the same chain) if they refer to the same entity. We construct entity chains and extract features that characterize these chains: How many chains start/end at this sentence? How many chains cross over this sentence/previous sentence/next sentence? Distance to the nearest dip/peak in the number of chains? Size of that dip/peak?

Pronoun Does the sentence contain a pronoun? Does the sentence contain a pronoun within 5 words of the beginning of the sentence?

Numbers During training, the patterns of numbers that occur significantly at boundaries are selected. Patterns considered are any number and any number with a specified length. The feature then checks if that pattern appears in the sentence. A commonly found pattern is the number pattern of length 4, which often refers to a year.

Conversation Is this sentence part of a conversation, i.e. does this sentence contain "direct speech"? This is determined by tracking beginning and ending quotes. Quoted regions and single sentences between two quoted regions are considered part of a conversation.

Paragraph Is this the beginning of a paragraph?

5 Experiments

In this section, we examine a number of narrative segmentation tasks with different segmentation methods. The only data used during development was the first two thirds from *Biohazard* (exp1 and exp2). All other data sets were only examined after the algorithm was developed and were used for testing purposes. Unless stated otherwise, results for the feature based method are using the SVM classifier.[1]

5.1 Evaluation Measures

We use three segmentation evaluation metrics that have been recently developed to account for "close but not exact" placement of hypothesized boundaries: word error probability, sentence error probability, and WindowDiff. Word error probability

[1] SVM and boosted decision stump performance is similar. For brevity, only SVM results are shown for most results.

Table 2: Experiments with *Biohazard*

	Word Error	Sent. Error	Window Diff	Sent err improv
Biohazard				
random (sent.)	0.488	0.485	0.539	——-
random (para.)	0.481	0.477	0.531	(base)
Biohazard				
exp1 → holdout	0.367	0.357	0.427	25%
exp2 → holdout	0.344	0.325	0.395	32%
3x cross validtn.	0.355	0.332	0.404	24%
Train *Biohazard*				
Test *Demon*	0.387	0.364	0.473	25%

(Beeferman et al., 1999) estimates the probability that a randomly chosen pair of words k words apart is incorrectly classified, i.e. a false positive or false negative of being in the same segment. In contrast to the standard classification measures of precision and recall, which would consider a "close" hypothesized boundary (e.g., off by one sentence) to be incorrect, word error probability gently penalizes "close" hypothesized boundaries. We also compute the sentence error probability, which estimates the probability that a randomly chosen pair of sentences s sentences apart is incorrectly classified. k and s are chosen to be half the average length of a section in the test data. WindowDiff (Pevzner and Hearst, 2002) uses a sliding window over the data and measures the difference between the number of hypothesized boundaries and the actual boundaries within the window. This metric handles several criticisms of the word error probability metric.

5.2 Segmenting Narrative Books

Table 2 shows the results of the SVM-segmenter on *Biohazard* and *Demon in the Freezer*. A baseline performance for segmentation algorithms is whether the algorithm performs better than naive segmenting algorithms: choose no boundaries, choose all boundaries and choose randomly. Choosing all boundaries results in word and sentence error probabilities of approximately 55%. Choosing no boundaries is about 45%. Table 2 also shows the results for random placement of the **correct** number of segments. Both random boundaries at sentence locations and random boundaries at paragraph locations are shown (values shown are the averages of 500 random runs). Similar results were obtained for random segmentation of the *Demon* data.

Table 3: Performance on Groliers articles

	Word Error	Sent. Error	Window Diff
random	0.482	0.483	0.532
TextTile	0.407	0.412	0.479
PLSA	0.420	0.435	0.507
features (stumps)	0.387	0.400	0.495
features (SVM)	0.385	0.398	0.503

For *Biohazard* the holdout set was not used during development. When trained on either of the development thirds of the text (i.e., exp1 or exp2) and tested on the test set, a substantial improvement is seen over random. 3-fold cross validation was done by training on two-thirds of the data and testing on the other third. Recalling from Table 1 that both PLSA and TextTiling result in performance similar to random even when given the correct number of segments, we note that all of the single train/test splits performed better than any of the naive algorithms and previous methods examined.

To examine the ability of our algorithm to perform on unseen data, we trained on the entire *Biohazard* book and tested on *Demon in the Freezer*. Performance on *Demon in the Freezer* is only slightly worse than the *Biohazard* results and is still much better than the baseline algorithms as well as previous methods. This is encouraging since *Demon* was not used during development, is written by a different author and has a segment length distribution that is different than *Biohazard* (average segment length of 30 vs. 18 in *Biohazard*).

5.3 Segmenting Articles

Unfortunately, obtaining a large number of narrative books with meaningful labeled segmentation is difficult. To evaluate our algorithm on a larger data set as well as a wider variety of styles similar to narrative documents, we also examine 1000 articles from Groliers Encyclopedia that contain subsections denoted by major and minor headings, which we consider to be the true segment boundaries. The articles contained 8,922 true and 102,116 possible boundaries. We randomly split the articles in half, and perform two-fold cross-validation as recommended by Dietterich (1998). Using 500 articles from one half of the pair for testing, 50 articles are randomly selected from the other half for training. We used

Table 4: Ave. human performance (Hearst, 1994)

	Word Error (%)	Sent. Error (%)	Window Diff (%)
Sequoia	0.275	0.272	0.351
Earth	0.219	0.221	0.268
Quantum	0.179	0167	0.316
Magellan	0.147	0.147	0.157

a subset of only 50 articles due to the high cost of labeling data. Each split yields two test sets of 500 articles and two training sets. This procedure of two-fold cross-validation is performed five times, for a total of 10 training and 10 corresponding test sets. Significance is then evaluated using the t-test.

The results for segmenting Groliers Encyclopedia articles are given in Table 3. We compare the performance of different segmentation models: two feature-based models (SVMs, boosted decision stumps), two similarity-based models (PLSA-based segmentation, TextTiling), and randomly selecting segmentation points. All segmentation systems are given the estimated number of segmentation points based based on the training data. The feature based approaches are significantly[2] better than either PLSA, TextTiling or random segmentation. For our selected features, boosted stump performance is similar to using an SVM, which reinforces our intuition that the selected features (and not just classification method) are appropriate for this problem.

Table 1 indicates that the previous TextTiling and PLSA-based approaches perform close to random on narrative text. Our experiments show a performance improvement of >24% by our feature-based system, and significant improvement over other methods on the Groliers data. Hearst (1994) examined the task of identifying the paragraph boundaries in *expository* text. We provide analysis of this data set here to emphasize that identifying segments in natural text is a difficult problem and since current evaluation methods were not used when this data was initially presented. Human performance on this task is in the 15%-35% error rate. Hearst asked seven human judges to label the paragraph boundaries of four different texts. Since no ground truth was available, true boundaries were identified by those boundaries that had a majority vote as a boundary. Table 4 shows the average human performance for each text. We show these results not for direct comparison with our methods, but to highlight that even human segmentation on a related task does not achieve particularly low error rates.

Table 5: Feature occurrences at boundary and non-boundary locations

	boundary	non-boundary
Paragraph	74	621
Entity groups	44	407
Word groups	39	505
Numbers	16	59
Full name	2	109
Conversation	0	510
Pronoun	8	742
Pronoun ≤ 5	1	330

5.4 Analysis of Features

The top section of Table 5 shows features that are intuitively hypothesized to be positively correlated with boundaries and the bottom section shows negatively correlated. For this analysis, exp1 from *Alibek* was used for training and the holdout set for testing. There are 74 actual boundaries and 2086 possibly locations. Two features have perfect recall: paragraph and conversation. Every true section boundary is at a paragraph and no section boundaries are within conversation regions. Both the word group and entity group features have good correlation with boundary locations and also generalized well to the training data by occurring in over half of the positive test examples.

The benefit of generalization using outside resources can be seen by comparing the boundary words found using word groups versus those found only in the training set as in Section 3. Using word groups triples the number of significant words found in the training set that occur in the test set. Also, the number of shared words that occur significantly in both the training and test set goes from none to 9. More importantly, significant words occur in 37 of the test segments instead of none without the groups.

[2] For both SVM and stumps at a level of 0.005 using a t-test except SVM_TextTile-WindowDiff (at 0.05) and stumps_TextTile-WindowDiff and SVM/stumps_PLSA-WindowDiff (not significantly different)

6 Discussion and Summary

Based on properties of narrative text, we proposed and investigated a set of features for segmenting narrative text. We posed the problem of segmentation as a feature-based classification problem, which presented a number of challenges: many different feature sources, generalization from outside resources for sparse data, and feature extraction from nontraditional information sources.

Feature selection and analyzing feature interaction is crucial for this type of application. The paragraph feature has perfect recall in that all boundaries occur at paragraph boundaries. Surprisingly, for certain train/test splits of the data, the performance of the algorithm was actually better without the paragraph feature than with it. We hypothesize that the noisiness of the data is causing the classifier to learn incorrect correlations.

In addition to feature selection issues, posing the problem as a classification problem loses the sequential nature of the data. This can produce very unlikely segment lengths, such as a single sentence. We alleviated this by selecting features that capture properties of the sequence. For example, the entity chains features represent some of this type of information. However, models for complex sequential data should be examined as possible better methods.

We evaluated our algorithm on two books and encyclopedia articles, observing significantly better performance than randomly selecting the correct number of segmentation points, as well as two popular, previous approaches, PLSA and TextTiling.

Acknowledgments

We thank Marti Hearst for the human subject performance data and the anonymous reviewers for their very helpful comments. Funded in part by the Advanced Research and Development Activity NIMD program (MDA904-03-C-0404).

References

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

Kenneth R. Beesley and Lauri Karttunen. 2003. *Finite State Morphology*. CSLI Publications, Palo Alto, CA.

Thorsten Brants, Francine Chen, and Ioannis Tsochantaridis. 2002. Topic-based document segmentation with probabilistic latent semantic analysis. In *Proceedings of CIKM*, pg. 211–218.

Thorsten Brants. 2000. TnT – a statistical part-of-speech tagger. In *Proceedings of the Applied NLP Conference*.

Freddy Choi. 2000. Improving the efficiency of speech interfaces for text navigation. In *Proceedings of IEEE Colloquium: Speech and Language Processing for Disabled and Elderly People*.

Nello Cristianini and John Shawe-Taylor. 2000. *An Introduction to Support Vector Machines*. Cambridge University Press.

Thomas Dietterich. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10:1895–1923.

Marti A. Hearst. 1994. Multi-paragraph segmentation of expository text. In *Meeting of ACL*, pg. 9–16.

Thorsten Joachims, 1999. *Advances in Kernel Methods - Support Vector Learning*, chapter Making large-Scale SVM Learning Practical. MIT-Press.

Hideki Kozima and Teiji Furugori. 1994. Segmenting narrative text into coherent scenes. In *Literary and Linguistic Computing*, volume 9, pg. 13–19.

Hideki Kozima. 1993. Text segmentation based on similarity between words. In *Meeting of ACL*, pg. 286–288.

Hang Li and Kenji Yamanishi. 2000. Topic analysis using a finite mixture model. In *Proceedings of Joint SIGDAT Conference of EMNLP and Very Large Corpora*, pg. 35–44.

Hajime Mochizuki, Takeo Honda, and Manabu Okumura. 1998. Text segmentation with multiple surface linguistic cues. In *COLING-ACL*, pg. 881–885.

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

Jeffrey Reynar. 1999. Statistical models for topic segmentation. In *Proceedings of ACL*, pg. 357–364.

Nicola Stokes, Joe Carthy, and Alex Smeaton. 2002. Segmenting broadcast news streams using lexical chains. In *Proceedings of Starting AI Researchers Symposium, (STAIRS 2002)*, pg. 145–154.

Ian H. Witten and Eibe Frank. 2000. *Data Mining: Practical machine learning tools with Java implementations*. Morgan Kaufmann.