Word Distributions for Thematic Segmentation in a Support Vector Machine Approach

Maria Georgescul
ISSCO/TIM, ETI
University of Geneva
1211 Geneva, Switzerland
maria.georgescul@eti.unige.ch

Alexander Clark
Department of Computer Science
Royal Holloway University of London
Egham, Surrey TW20 0EX, UK
alexc@cs.rhul.ac.uk

Susan Armstrong
ISSCO/TIM, ETI
University of Geneva
1211 Geneva, Switzerland
susan.armstrong@issco.unige.ch

Abstract

We investigate the appropriateness of using a technique based on support vector machines for identifying thematic structure of text streams. The thematic segmentation task is modeled as a binary-classification problem, where the different classes correspond to the presence or the absence of a thematic boundary. Experiments are conducted with this approach by using features based on word distributions through text. We provide empirical evidence that our approach is robust, by showing good performance on three different data sets. In particular, substantial improvement is obtained over previously published results of word-distribution based systems when evaluation is done on a corpus of recorded and transcribed multi-party dialogs.

1 Introduction

(Todd, 2005) distinguishes between "local-level topics (of sentences, utterances and short discourse segments)" and "discourse topics (of more extended stretches of discourse)".[1] (Todd, 2005) points out that "discourse-level topics are one of the most elusive and intractable notions in semantics". Despite this difficulty in giving a rigorous definition of *discourse topic*, the task of discourse/dialogue segmentation into thematic episodes can be described by invoking an "intuitive notion of topic" (Brown and Yule, 1998). Thematic segmentation also relates to several notions such as speaker's intention, topic flow and cohesion.

In order to find out if thematic segment identification is a feasible task, previous state-of-the-art works appeal to experiments, in which several human subjects are asked to mark thematic segment boundaries based on their intuition and a minimal set of instructions. In this manner, previous studies, e.g. (Passonneau and Litman, 1993; Galley et al., 2003), obtained a level of inter-annotator agreement that is statistically significant.

Automatic thematic segmentation (TS), i.e. the segmentation of a text stream into topically coherent segments, is an important component in applications dealing with large document collections such as information retrieval and document browsing. Other tasks that could benefit from the thematic textual structure include anaphora resolution, automatic summarisation and discourse understanding.

The work presented here tackles the problem of TS by adopting a supervised learning approach for capturing linear document structure of non-overlapping thematic episodes. A prerequisite for the input data to our system is that texts are divided into sentences or utterances.[2] Each boundary between two consecutive utterances is a potential thematic segmentation point and therefore, we model the TS task as a binary-classification problem, where each utterance should be classified as marking the

[1] In this paper, we make use of the term *topic* or *theme* as referring to the discourse/dialogue topic.

[2] Occasionally within this document we employ the term utterance to denote either a sentence or an utterance in its proper sense.

presence or the absence of a topic shift in the discourse/dialogue based only on observations of patterns in vocabulary use.

The remainder of the paper is organised as follows. The next section summarizes previous techniques, describes how our method relates to them and presents the motivations for a support vector approach. Sections 3 and 4 present our approach in adopting support vector learning for thematic segmentation. Section 5 outlines the empirical methodology and describes the data used in this study. Section 6 presents and discusses the evaluation results. The paper closes with Section 7, which briefly summarizes this work and offers some conclusions and future directions.

2 Related Work

As in many existing approaches to the thematic segmentation task, we make the assumption that the thematic coherence of a text segment is reflected at lexical level and therefore we attempt to detect the correlation between word distribution and thematic changes throughout the text. In this manner, (Hearst, 1997; Reynar, 1998; Choi, 2000) start by using a similarity measure between sentences or fixed-size blocks of text, based on their word frequencies in order to find changes in vocabulary use and therefore the points at which the topic changes. Sentences are then grouped together by using a clustering algorithm. (Utiyama and Isahara, 2001) models the problem of TS as a problem of finding the minimum cost path in a graph and therefore adopts a dynamic programming algorithm. The main advantage of such methods is that no training time and corpora are required.

By modeling TS as binary-classification problem, we introduce a new technique based on support vector machines (SVMs). The main advantage offered by SVMs with respect to methods such as those described above is related to the distance (or similarity) function used. Thus, although (Choi, 2000; Hearst, 1997) employ a distance function (i.e. *cosine distance*) to detect thematic shifts, SVMs are capable of using a larger variety of similarity functions.

Moreover, SVMs can employ distance functions that operate in extremely high dimensional feature spaces. This is an important property for our task, where handling high dimensionality data representation is necessary (see section 4).

An alternative to dealing with high dimension data may be to reduce the dimensionality of the data representation. Therefore, linear algebra dimensionality reduction methods like singular value decomposition have been adopted by (Choi et al., 2001; Popescu-Belis et al., 2004) in Latent Semantic Analysis (LSA) for the task of thematic segmentation. A Probabilistic Latent Semantic Analysis (PLSA) approach has been adopted by (Brants et al., 2002; Farahat and Chen, 2006) for the TS task. (Blei and Moreno, 2001) proposed a TS approach, by embedding a PLSA model in an extended Hidden Markov Model (HMM) approach, while (Yamron et al., 1998) have previously proposed a HMM approach for TS.

A shortcoming of the methods described above is due to their typically generative manner of training, i.e. using the maximum likelihood estimation for a joint sampling model of observation and label sequences. This poses the challenge of finding more appropriate *objective functions*, i.e. alternatives to the log-likelihood that are more closely related to application-relevant performance measures. Secondly, efficient inference and learning for the TS task often requires making questionable conditional independence assumptions. In such cases, improved performance may be obtained by using methods with a more discriminative character, by allowing direct dependencies between a label and past/future observations and by efficient handling higher-order combinations of input features. Given the discriminative character of SVMs, we expect our model to attain similar benefits.

3 Support Vector Learning Task and Thematic Segmentation

The theory of Vapnik and Chervonenkis (Vapnik, 1995) motivated the introduction of support vector learning. SVMs have originally been used for classification purposes and their principles have been extended to the task of regression, clustering and feature selection. (Kauchak and Chen, 2005) employed SVMs using features (derived for instance from information given by the presence of paragraphs, pronouns, numbers) that can be reliably used for topic

segmentation of narrative documents. Aside from the fact that we consider the TS task on different datasets (not only on narrative documents), our approach is different from the approach proposed by (Kauchak and Chen, 2005) mainly by the data representation we propose and by the fact that we put the emphasis on deriving the thematic structure merely from word distribution, while (Kauchak and Chen, 2005) observed that the 'block similarities provide little information about the actual segment boundaries' on their data and therefore they concentrated on exploiting other features.

An excellent general introduction to SVMs and other kernel methods is given for instance in (Cristianini and Shawe-Taylor, 2000). In the section below, we give some highlights representing the main elements in using SVMs for thematic segmentation.

The support vector learner $\mathcal{L}$ is given a *training set* of n *examples*, usually denoted by S_{train}= (($\vec{u}_1$, y_1),...,($\vec{u}_n$, y_n))$\subseteq$ $(U \times Y)^n$ drawn independently and identically distributed according to a fixed distribution $Pr(u, y) = Pr(y|u)Pr(u)$. Each training example consists of a high-dimensional vector $\vec{u}$ describing an utterance and the class label y. The utterance representations we chose are further described in Section 4. The class label y has only two possible values: 'thematic boundary' or 'non-thematic boundary'. For notational convenience, we replace these values by +1 and -1 respectively, and thus we have $y \in \{$-1, 1$\}$. Given a hypothesis space $\mathcal{H}$, of functions $h : U \rightarrow \{-1, +1\}$ having the form $h(\vec{u}) = sign(< \vec{w}, \vec{u} > +b)$, the inductive support vector learner $\mathcal{L}_{ind}$ seeks a decision function h_{ind} from $\mathcal{H}$, using S_{train} so that the expected number of erroneous predictions is minimized. Using the structural risk minimization principle (Vapnik, 1995), the support vector learner gets the optimal decision function h by minimizing the following cost function:

$$\mathcal{W}^{ind}(\vec{w}, b, \xi_1, \xi_2, ..., \xi_n) = \tfrac{1}{2} < \vec{w}, \vec{w} > +$$

$$+ C^+ \sum_{i=0, y_i=1}^{n} \xi_i + C^- \sum_{i=0, y_i=-1}^{n} \xi_i,$$

subject to:

$$y_i[< \vec{w} \cdot \vec{u}_i > +b] \ \leq 1 - \xi_i \ for \ i = 1, 2, ..., n;$$

$$\xi_i \geq 0 \ for \ i = 1, 2, ..., n.$$

The parameters $\vec{w}$ and b follow from the optimisation problem, which is solved by applying Lagrangian theory. The so-called *slack variables* ξ_i, are introduced in order to be able to handle non-separable data. The positive parameters C^+ and C^- are called *regularization parameters* and determine the amount up to which errors are tolerated. More exactly, training data may contain noisy or outlier data that are not representative of the underlying distribution. On the one hand, fitting exactly to the training data may lead to overfitting. On the other hand, dismissing true properties of the data as sampling bias in the training data will result in low accuracy. Therefore, the regularization parameter is used to balance the trade-off between these two competing considerations. Setting the regularization parameter too low can result in poor accuracy, while setting it too high can lead to overfitting. In the TS task, we used an automated procedure to select the regularization parameters, as further described in section 5.3.

In cases where non-linear hypothesis functions should be optimised, each $\vec{u}_i$ can be mapped into $\varphi(\vec{u}_i) \in F$, where F is a higher dimensional space usually called *feature space*, in order to make linear the relation between $\vec{u}_i$ and y_i. Thus the original linear learning machine can be adopted in finding the classification solution in the feature space.

When using a mapping function $\varphi : U \rightarrow F$, if we have a way of computing the inner product $\langle \varphi(\vec{u}_i), \varphi(\vec{u}_j) \rangle$ directly as a function of the original input point, then the so-called kernel function $K(\vec{u}_i, \vec{u}_j) = \langle \varphi(\vec{u}_i), \varphi(\vec{u}_j) \rangle$ is proved to simplify the computational complexity implied by the direct use of the mapping function φ. The choice of appropriate kernels and its specific parameters is an empirical issue. In our experiments, we used the Gaussian radial basis function (RBF) kernel:

$$K_{RBF}(\vec{u}_i, \vec{u}_j) = \exp(-\gamma^2 ||\vec{u}_i - \vec{u}_j||^2).$$

For the SVM calculations, we used the *LIBSVM* library (Chang and Lin, 2001).

4 Representation of the information used to determine thematic boundaries

As presented in section 3, in the thematic segmentation task, an input $\vec{u}_i$ to the support vector classifier is a vectorial representation of the utterance to

be classified and its context. Each dimension of the input vector indicates the value of a certain feature characterizing the utterance. All input features here are indicator functions for a word occurring within a fixed-size window centered on the utterance being labeled. More exactly, the input features are computed in the following steps:

1. The text has been pre-processed by tokenization, elimination of stop-words and lemmatization, using *TreeTagger* (Schmid, 1996).

2. We make use of the so-called *bag of words* approach, by mapping each utterance to a *bag*, i.e. a set that contains word frequencies. Therefore, word frequencies have been computed to count the number of times that each term (i.e. word lemma) is used in each utterance. Then a transformation of the raw word frequency counts is applied in order to take into account both the local (i.e. for each utterance) word frequencies as well as the overall frequencies of their occurrences in the entire text collection. More exactly, we made experiments in parallel with three such transformations, which are very commonly used in information retrieval domain (Dumais, 1991): *tf.idf*, *tf.normal* and *log.entropy*.

3. Each i-th utterance is represented by a vector $\vec{u}_i$, where a j-th element of $\vec{u}_i$ is computed as:

$$u_{i,j} = \left(\sum_{t=i-winSize}^{i} f_{t,j} \right) \left(\sum_{k=i+1}^{i+winSize} f_{k,j} \right),$$

   where $winSize \geq 1$ and $f_{i,j}$ is the weighted frequency (determined in the previous step) of the j-th word from the vocabulary in the i-th utterance. In this manner, we will have $u_{i,j} > 0$ if and only if at least two occurrences of the j-th term occur within $(2 \cdot winSize)$ utterances on opposite sides of a boundary candidate. That is, each $u_{i,j}$ is capturing how many word co-occurrences appear across the candidate utterance in an interval (of $(2 \cdot winSize)$ utterances) centered in the boundary candidate utterance.

4. Each attribute value from the input data is scaled to the interval $[0, 1]$.

Note that the vector space representation adopted in the previous steps will result in a sparse high dimensional input data for our system. More exactly, table 1 shows the average number of non-zero features per example corresponding to each data set (further described in section 5.1).

Data set	Non zero features
ICSI	3.67%
TDT	0.40%
Brown	0.12%

Table 1: The percentage of non-zero features per example.

5 Experimental Setup

5.1 Data sets used

In order to evaluate how robust our SVM approach is, we performed experiments on three English data sets of approximately the same dimension (i.e. containing about 260,000 words).

The first dataset is a subset of the ICSI-MR corpus (Janin et al., 2004), where the gold standard for thematic segmentations has been provided by taking into account the agreement of at least three human annotators (Galley et al., 2003). The corpus consists of high-quality close talking microphone recordings of multi-party dialogues. Transcriptions at word level with utterance-level segmentations are also available. A test sample from this dataset consists of the transcription of an approximately one-hour long meeting and contains an average of about seven thematic episodes.

The second data set contains documents randomly selected from the Topic Detection and Tracking (TDT) 2 collection, made available by (LDC, 2006). The TDT collection includes broadcast news and newswire text, which are segmented into topically cohesive stories. We use the story segmentation provided with the corpus as our gold standard labeling. A test sample from our subset contains an average of about 24 segments.

The third dataset we use in this study was originally proposed in (Choi, 2000) and contains artificial thematic episodes. More precisely, the dataset is built by concatenating short pieces of texts that

Data set	Weighting schema	winSize	γ	*C*
ICSI	log.entropy	57	0.0625	0.01
TDT	tf.idf	17	0.0625	0.1
Brown	tf.idf	5	0.0625	0.001

Table 2: The optimal settings found for the SVM model, using the RBF kernel.

have been randomly extracted from the Brown corpus. Any test sample from this dataset consists of ten segments. Each segment contains at least three sentences and no more than eleven sentences.

While the focus of our paper is not on the method of evaluation, it is worth pointing out that the performance on the synthetic data set is a very poor guide to the performance on naturally occurring data (Georgescul et al., 2006). We include the synthetic data for comparison purposes.

5.2 Handling unbalanced data

We have a small percentage of positive examples relative to the total number of training examples. Therefore, in order to ensure that positive points are not considered as being noisy labels, we change the penalty of the minority (positive) class by setting the parameter C^+ of this class to:

$$C^+ = \lambda \cdot \left(\frac{n}{n^+ - 1} - 1 \right) \cdot C^-,$$

where n^+ is the number of positive training examples, n is the total number of training examples and λ is the scaling factor. In the experiments reported here, we set the value for the scale factor λ to $\lambda = 1$ and we have: $C^+ = 7 \cdot C^-$ for the synthetic data derived from Brown corpus; $C^+ = 18 \cdot C^-$for the TDT data and $C^+ = 62 \cdot C^-$ for the ICSI meeting data.

5.3 Model selection

We used 80% of each dataset to determine the best model settings, while the remaining 20% is used for testing purposes. Each training set (for each dataset employed) was divided into disjoint subsets and five-fold cross-validation was applied for model selection.

In order to avoid too many combinations of parameter settings, model selection is done in two phases, by distinguishing two kinds of parameters. First, the parameters involved in data representation (see section 4) are addressed. We start with choosing an appropriate term weighting scheme and a good value for the $winSize$ parameter. This choice is based on a systematic grid search over 20 different values for $winSize$ and the three variants *tf.idf*, *tf.normal* and *log.entropy* for term weighting. We ran five-fold cross validation, by using the RBF kernel with its parameter γ fixed to $\gamma = 1$. We also set the regularization parameter C equal to C = 1.

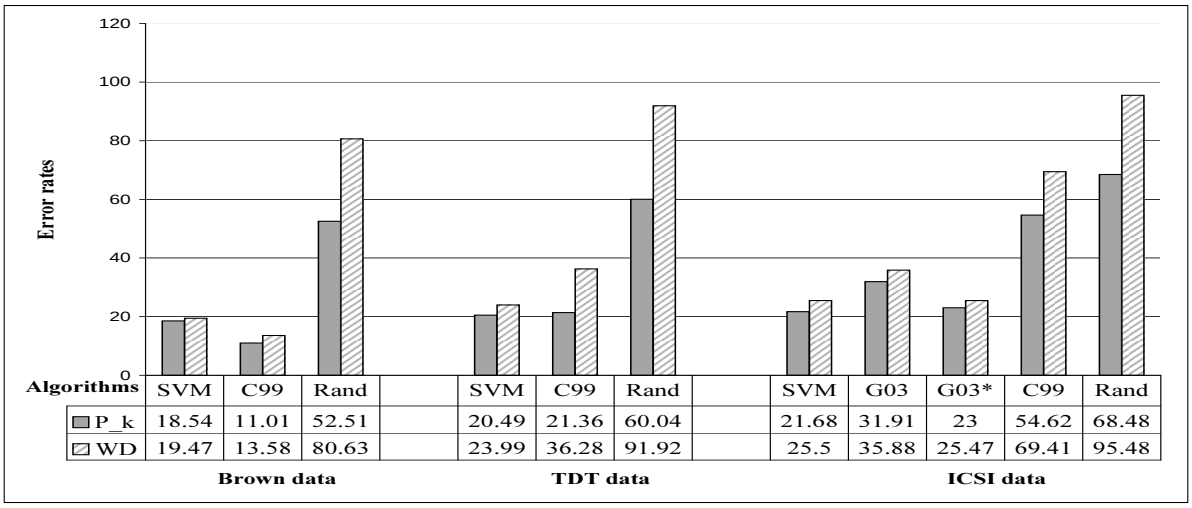
In the second phase of model selection, we take the optimal parameter values selected in the previous phase as a constant factor and search the most appropriate values for C and γ parameters. The range of values we select from is: $C \in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$ and $\gamma \in \{2^{-6}, 2^{-5}, 2^{-4}, ..., 2^4, 2^6\}$ and for each possible value we perform five-fold cross validation. Therefore, we ran the algorithm five times for the $91 = 7 \times 13$ parameter settings. The most suitable model settings found are shown in Table 2. For these settings, we show the algorithm's results in section 6.

6 Evaluation

6.1 Evaluation Measures

Beeferman et al. (1999) underlined that the standard evaluation metrics of *precision* and *recall* are inadequate for thematic segmentation, namely by the fact that these metrics did not account for how far away a hypothesized boundary (i.e. a boundary found by the automatic procedure) is from the reference boundary. On the other hand, for instance, an algorithm that places a boundary just one utterance away from the reference boundary should be penalized less than an algorithm that places a boundary ten (or more) utterances away from the reference boundary.

Hence the use of two other evaluation metrics is favored in thematic segmentation: the P_k metric (Beeferman et al., 1999) and the *WindowDiff* error metric (Pevzner and Hearst, 2002). In con-



Algorithms	SVM	C99	Rand		SVM	C99	Rand		SVM	G03	G03*	C99	Rand
P_k	18.54	11.01	52.51		20.49	21.36	60.04		21.68	31.91	23	54.62	68.48
WD	19.47	13.58	80.63		23.99	36.28	91.92		25.5	35.88	25.47	69.41	95.48
	Brown data				TDT data				ICSI data				

Figure 1: Error rates of the segmentation systems.

trast to precision and recall, these metrics allow for a slight vagueness in where the hypothesized thematic boundaries are placed and capture "the notion of nearness in a principled way, gently penalizing algorithms that hypothesize boundaries that aren't quite right, and scaling down with the algorithm's degradation" (Beeferman et al., 1999). That is, computing both P_k and *WindowDiff* metrics involves the use of a fixed-size (i.e. having a fixed number of either words or utterances) window that is moved step by step over the data. At each step, P_k and *WindowDiff* are basically increased (each metric in a slightly different way) if the hypothesized boundaries and the reference boundaries are not within the same window.

During the model selection phase, we used precision and recall in order to measure the system's error rate. This was motivated by the fact that posing the TS task as a classification problem leads to a loss of the sequential nature of the data, which is an inconvenient in computing the P_k and *WindowDiff* measures. However, during the final testing phase of our system, as well as for the evaluation of the previous systems, we use both the P_k and the *WindowDiff* error metric.

The relatively small size of our datasets does not allow for dividing our test set into multiple sub-test sets for applying statistical significance tests. This would be desirable in order to indicate whether the differences in system error rates are statistically significant over different data sets. Nevertheless, we believe that measuring differences in error rates obtained on the test set is indicative of the relative performance. Thus, the experimental results shown in this paper should be considered as illustrative rather than exhaustive.

6.2 Results

In order to determine the adequacy of our SVM approach over different genres, we ran our system over three datasets, namely the ICSI meeting data, the TDT broadcast data and the Brown written genre data.

By measuring the system error rates using the P_k and the *WindowDiff* metrics, Figure 1 summarizes the quantitative results obtained in our empirical evaluation. In Figure 1, our SVM approach is labeled as *SVM* and we abbreviate *WindowDiff* as *WD*. The results of our *SVM* system correspond to the parameter values detected during model selection (see Table 2). We compare our system against an existing thematic segmenter in the literature: *C99* (Choi, 2000). We also give for comparison the error rates of a naive algorithm, labeled as *Rand* algorithm, which randomly distributes boundaries throughout the text.

The *LCseg* system (Galley et al., 2003), labeled here as *G03*, is to our knowledge the only word distribution based system evaluated on ICSI meeting data. Therefore, we replicate the results reported by (Galley et al., 2003) when evaluation of *LCseg* was done on ICSI data. The so-labeled *G03** algorithm

indicates the error rates obtained by (Galley et al., 2003) when extra (meeting specific) features have been adopted in a decision tree classifier. However, note that the results reported by (Galley et al.) are not directly comparable with our results because of a slight difference in the evaluation procedure: (Galley et al.) performed 25-fold cross validation and the average P_k and WD error rates have been computed on the held-out sets.

Figure 1 illustrates the following interesting results. For the ICSI meeting data, our SVM approach provides the best performance relative to the competing word distribution based state-of-the-art methods. This proves that our SVM-based system is able to build a parametric model that leads to a segmentation that highly correlates to a human thematic segmentation. Furthermore, by taking into account the relatively small size of the data set we used for training, it can be concluded that the SVM can build qualitatively good models even with a small training data. The work of (Galley et al., 2003) shows that the *G03** algorithm is better than *G03* by approximately 10%, which indicates that on meeting data the performance of our word-distribution based approach could possibly be increased by using other meeting-specific features.

By examining the error rates given by P_k metric for the three systems on the TDT data set, we observe that our system and *C99* performed more or less equally. With respect to the *WindowDiff* metric, our system has an error rate approximately 10% smaller than *C99*.

On the synthetic data set, the *SVM* approach performed slightly worse than *C99*, avoiding however catastrophic failure, as observed with the *C99* method on ICSI data.

7 Conclusions

We have introduced a new approach based on word distributions for performing thematic segmentation. The thematic segmentation task is modeled here as a binary classification problem and support vector machine learning is adopted. In our experiments, we make a comparison of our approach versus existing linear thematic segmentation systems reported in the literature, by running them over three different data sets. When evaluating on real data, our approach either outperformed the other existing methods or performs comparably to the best. We view this as a strong evidence that our approach provides a unified and robust framework for the thematic segmentation task. The results also suggest that word distributions themselves might be a good candidate for capturing the thematic shifts of text and that SVM learning can play an important role in building an adaptable correlation.

Our experiments also show the sensitivity of a segmentation method to the type of a corpus on which it is tested. For instance, the C99 algorithm which achieves superior performance on a synthetic collection performs quite poorly on the real-life data sets.

While we have shown empirically that our technique can provide considerable gains by using single word distribution features, future work will investigate whether the system can be improved by exploiting other features derived for instance from syntactic, lexical and, when available, prosodic information. If further annotated meeting data becomes available, it would be also interesting to replicate our experiments on a bigger data set in order to verify whether our system performance improves.

Acknowledgments This work is partially supported by the Interactive Multimodal Information Management project (http://www.im2.ch/). Many thanks to the reviewers for their insightful suggestions. We are grateful to the International Computer Science Institute (ICSI), University of California for sharing the data with us. The authors also thank Michael Galley who kindly provided us the thematic annotations of the ICSI data.

References

Doug Beeferman, Adam Berger, and John Lafferty. 1999. Statistical Models for Text Segmentation. *Machine Learning*, 34(1-3):177–210.

David M. Blei and Pedro J. Moreno. 2001. Topic Segmentation with an Aspect Hidden Markov Model. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 343–348. ACM Press.

Thorsten Brants, Francine Chen, and Ioannis Tsochantaridis. 2002. Topic-Based Document Segmentation with Probabilistic Latent Semantic Analysis. In *Proceedings of the Eleventh International Conference on*

Information and Knowledge Management, pages 211–218, McLean, Virginia, USA. ACM Press.

Gillian Brown and George Yule. 1998. *Discourse Analysis*. Cambridge Textbooks in Linguistics, Cambridge.

Chih-Chung Chang and Chih-Jen Lin. 2001. LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

Freddy Choi, Peter Wiemer-Hastings, and Johanna Moore. 2001. Latent Semantic Analysis for Text Segmentation. In *Proceedings of the 6th Conference on Empirical Methods in Natural Language Processing*, Seattle, WA.

Freddy Choi. 2000. Advances in Domain Independent Linear Text Segmentation. In *Proceedings of the 1st Conference of the North American Chapter of the Association for Computational Linguistics*, pages 26–33, Seattle, USA.

Nello Cristianini and John Shawe-Taylor. 2000. *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, Cambridge, UK.

Susan Dumais. 1991. Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments and Computers*, 23(2):229–236.

Ayman Farahat and Francine Chen. 2006. Improving Probabilistic Latent Semantic Analysis with Principal Component Analysis. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy.

Michael Galley, Kathleen McKeown, Eric Fosler-Luissier, and Hongyan Jing. 2003. Discourse Segmentation of Multy-Party Conversation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 562–569.

Maria Georgescul, Alexander Clark, and Susan Armstrong. 2006. An Analysis of Quantitative Aspects in the Evaluation of Thematic Segmentation Algorithms. *To appear*.

Marti Hearst. 1997. TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages. *Computational Linguistics*, 23(1):33–64.

Adam Janin, Jeremy Ang, Sonali Bhagat, Rajdip Dhillon, Jane Edwards, Javier Macias-Guarasa, Nelson Morgan, Barbara Peskin, Elizabeth Shriberg, Andreas Stolcke, Chuck Wooters, and Britta Wrede. 2004. The ICSI Meeting Project: Resources and Research. In *ICASSP 2004 Meeting Recognition Workshop (NIST RT-04 Spring Recognition Evaluation)*, Montreal.

David Kauchak and Francine Chen. 2005. Feature-Based Segmentation of Narrative Documents. In *Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing*, pages 32–39, Ann Arbor; MI; USA.

LDC. 2006. The Linguistic Data Consortium. Available from World Wide Web: http://www.ldc.upenn.edu.

Rebecca J. Passonneau and Diane J. Litman. 1993. Intention-based Segmentation: Human Reliability and Correlation with Linguistic Cues. In *Proceedings of the 31st conference on Association for Computational Linguistics*, pages 148 – 155, Columbus, Ohio.

Lev Pevzner and Marti Hearst. 2002. A Critique and Improvement of an Evaluation Metric for Text Segmentation. *Computational Linguistics*, 16(1):19–36.

Andrei Popescu-Belis, Alexander Clark, Maria Georgescul, Sandrine Zufferey, and Denis Lalanne. 2004. Shallow Dialogue Processing Using Machine Learning Algorithms (or Not). In Bourlard H. and Bengio S., editors, *Multimodal Interaction and Related Machine Learning Algorithms*, pages 277–290. LNCS 3361, Springer-Verlag, Berlin.

Jeffrey Reynar. 1998. *Topic Segmentation: Algorithms and Applications*. Ph.D. thesis, University of Pennsylvania.

Helmut Schmid. 1996. Probabilistic Part-of-Speech Tagging Using Decision Trees. Technical report, Institute for Computational Linguistics of the University of Stuttgart.

Richard Watson Todd. 2005. A fuzzy approach to discourse topics. *Journal of the International Association for Semiotic Studies*, 155:93–123.

Masao Utiyama and Hitoshi Isahara. 2001. A Statistical Model for Domain-Independent Text Segmentation. In *Proceedings of the 39th Annual Meeting of the ACL joint with the 10th Meeting of the European Chapter of the ACL*, pages 491–498, Toulouse, France.

Vladimir Naumovich Vapnik. 1995. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York.

Jonathan P. Yamron, Ira Carp, Lawrence Gillick, Stewe Lowe, and Paul van Mulbregt. 1998. A Hidden Markov Model Approach to Text Segmentation and Event Tracking. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, volume 17, pages 333–336, Seattle, WA.