Using Linguistically Motivated Features for Paragraph Boundary Identification

Katja Filippova and **Michael Strube**
EML Research gGmbH
Schloss-Wolfsbrunnenweg 33
69118 Heidelberg, Germany
http://www.eml-research.de/nlp

Abstract

In this paper we propose a machine-learning approach to paragraph boundary identification which utilizes linguistically motivated features. We investigate the relation between paragraph boundaries and discourse cues, pronominalization and information structure. We test our algorithm on German data and report improvements over three baselines including a reimplementation of Sporleder & Lapata's (2006) work on paragraph segmentation. An analysis of the features' contribution suggests an interpretation of what paragraph boundaries indicate and what they depend on.

1 Introduction

Our work is concerned with multi-document summarization, namely with the merging of multiple documents about the same topic taken from the web. We view summarization as extraction of important sentences from the text. As a consequence of the merging process the layout of the documents is lost. In order to create the layout of the output, the document structure (Power et al., 2003) has to be regenerated. One aspect of this structure is of particular importance for our work: the paragraph structure. In web documents paragraph boundaries are used to anchor figures and illustrations, so that the figures are always aligned with the same paragraph even when the font size or the window size is changed. Since we want to include figures in the generated summaries, paragraph segmentation is an important subtask in our application.

Besides multi-document summarization of web documents, paragraph boundary identification (PBI) could be useful for a number of different applications, such as producing the layout for transcripts provided by speech recognizers and optical character recognition systems, and determining the layout of documents generated for output devices with different screen size.

Though related to the task of topic segmentation which stimulated a large number of studies (Hearst, 1997; Choi, 2000; Galley et al., 2003, inter alia), paragraph segmentation has not been thoroughly investigated so far. We explain this by the fact that paragraphs are considered a stylistic phenomenon and that there is no unanimous opinion on what the function of the paragraph is. Some authors (Irmscher (1972) as cited by Stark (1988)) suggest that paragraph structure is arbitrary and can not be determined based solely on the properties of the text. Still, psycholinguistic studies report that humans agree, at least to some extent, on placing boundaries between paragraphs. These studies also note that paragraph boundaries are informative and make the reader perceive paragraph-initial sentences as being important (Stark, 1988). In contrast to topic segmentation, paragraph segmentation has the advantage that large amounts of annotated data are readily availabe for supervised learning.

In this paper we describe our approach to paragraph segmentation. Previous work (Sporleder & Lapata, 2004; 2006) mainly focused on superficial and easily obtainable surface features like punctuation, quotes, distance and words in the sentence. Their approach was claimed to be domain- and language-independent. Our hypothesis, however, is that linguistically motivated features, which we compute automatically, provide a better paragraph segmentation than Sporleder & Lapata's surface ones, though our approach may loose some of the

Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006), pages 267–274,
Sydney, July 2006. ©2006 Association for Computational Linguistics

domain-independence. We test our hypothesis on a corpus of biographies downloaded from the German Wikipedia[1]. The results we report in this paper indicate that linguistically motivated features outperform surface features significantly. It turned out that pronominalization and information structure contribute to the determination of paragraph boundaries while discourse cues have a negative effect.

The paper is organized as follows: First, we describe related work in Section 2, then in Section 3 our data is introduced. The baselines, the machine learners, the features and the experimental setup are given in Section 4. Section 5 reports and discusses the results.

2 Related Work

Compared to other text segmentation tasks, e.g. topic segmentation, PBI has received relatively little attention. We are aware of three studies which approach the problem from different perspectives. Bolshakov & Gelbukh (2001) assume that splitting text into paragraphs is determined by text cohesion: The link between a paragraph initial sentence and the preceding context is weaker than the links between sentences within a paragraph. They evaluate text cohesion using a database of collocations and semantic links and insert paragraph boundaries where the cohesion is low.

The algorithm of Sporleder & Lapata (2004, 2006) uses surface, syntactic and language model features and is applied to three different languages and three domains (fiction, news, parliament). This study is of particular interest to us since one of the languages the algorithm is tested on is German. They investigate the impact of different features and data size, and report results significantly better than a simple baseline. However, their results vary considerably between the languages and the domains. Also, the features determined important is different for each setting. So, it may be possible that Sporleder & Lapata do not provide conclusive results.

Genzel (2005) considers lexical and syntactic features and reports accuracy obtained from English fiction data as well as from the WSJ corpus. He points out that lexical coherence and structural features turn out to be the most useful for his algorithm. Unfortunately, the only evaluation measure he provides is accuracy which, for the PBI task, does not describe the performance of a system sufficiently.

In comparison to the mentioned studies, our goal is to examine the influence of cohesive features on the choice of paragraph boundary insertion. Unlike Bolshakov & Gelbukh (2001), who have similar motivation but measure cohesion by collocations, we explore the role of discourse cues, pronominalization and information structure.

The task of topic segmentation is closely related to the task of paragraph segmentation. If there is a topic boundary, it is very likely that it coincides with a paragraph boundary. However, the reverse is not true and one topic can extend over several paragraphs. So, if determined reliably, topic boundaries could be used as high precision, low recall predictors for paragraph boundaries. Still, there is an important difference: While work on topic segmentation mainly depends on content words (Hearst, 1997) and relations between them which are computed using lexical chains (Galley et al., 2003), paragraph segmentation as a stylistic phenomenon may depend equally likely on function words. Hence, paragraph segmentation is a task which encompasses the traditional borders between content and style.

3 Data

The data we used is a collection of biographies from the German version of Wikipedia. We selected all biographies under the Wikipedia categories of physicists, chemists, mathematicians and biologists and obtained 970 texts with an average length of 20 sentences and 413,776 tokens in total.

Although our corpus is substantially smaller than the German corpora of Sporleder & Lapata (2006), it should be big enough for a fair comparison between their algorithm and the algorithm proposed here. Having investigated the effect of the training size, Sporleder & Lapata (2006) came to the conclusion that their system performs well being trained on a small data set. In particular, the learning curve for German shows an improvement of only about 2% when the amount of training data is increased from 20%, which in case of German fiction approximately equals 370,000 tokens, to 100%.

Fully automatic preprocessing in our system comprises the following stages: First, a list of people of a certain Wikipedia category is taken and for every person an article is extracted The text

[1] http://de.wikipedia.org

	training	development	test
tokens	347,763	39,228	19,943
sentences	15,583	1,823	922
paragraphs	5,323	654	362

Table 1: Number of tokens and sentences per set

is purged from Wiki tags and comments, the information on subtitles and paragraph structure is preserved. Second, sentence boundaries are identified with a Perl CPAN module[2] whose performance we improved by extending the list of abbreviations and modifying the output format. Next, the sentences are split into tokens. The TnT tagger (Brants, 2000) and the TreeTagger (Schmid, 1997) are used for tagging and lemmatizing. Finally, the texts are parsed with the CDG dependency parser (Foth & Menzel, 2006). Thus, the text is split on three levels: paragraphs, sentences and tokens, and morphological and syntactic information is provided.

A publicly available list of about 300 discourse connectives was downloaded from the Internet site of the Institute for the German Language[3] (Institut für Deutsche Sprache, Mannheim) and slightly extended. These are identified in the text and annotated automatically as well. Named entities are classified according to their type using information from Wikipedia: *person, location, organization* or *undefined*. Given the peculiarity of our corpus, we are able to identify all mentions of the biographee in the text by simple string matching. We also annotate different types of referring expressions (*first, last, full name*) and resolve anaphora by linking personal pronouns to the biographee provided that they match in number and gender.

The annotated corpus is split into training (85%), development (10%) and testing (5%) sets. Distribution of data among the three sets is presented in Table 1. Sentences which serve as subtitles in a text are filtered out because they make identifying a paragraph boundary for the following sentence trivial.

4 Experiments

4.1 Machine Learners

The PBI task was reformulated as a binary classification problem: every training instance representing a sentence was classified either as paragraph-initial or not.

We used two machine learners: BoosTexter (Schapire & Singer, 2000) and TiMBL (Daelemans et al., 2004). BoosTexter was developed for text categorization, and combines simple rules (decision stumps) in a boosting manner. Sporleder & Lapata used this learner because it has the ability to combine many only moderately accurate hypotheses. TiMBL is a memory-based learner which classifies every test instance by finding the most similar examples in the training set, hence it does not abstract from the data and is well suited to handle features with many values, e.g. the list of discourse cues. For both classifiers, all experiments were run with the default settings.

4.2 Baselines

We compared the performance of our algorithm against three baselines. The first one (**distance**) trivially inserts a paragraph break after each third sentence, which is the average number of sentences in a paragraph. The second baseline (**Galley**) hypothesizes that paragraph breaks coincide with topic boundaries and utilizes Galley et al.'s (2003) topic boundary identification tool LCseg. The third baseline (**Sporleder**) is a reimplementation of Sporleder & Lapata's 2006 algorithm with the following features:

Word and **Sentence Distances** from the current sentence to the previous paragraph break;

Sentence Length and **Relative Position (relPos)** of the sentence in a text;

Quotes encodes whether this and the previous sentences contain a quotation, and whether the quotation is continued in the current sentence or not;

Final Punctuation of the previous sentence;

Words – the first **(word1)**, the first two **(word2)**, the first three and all words from the sentence;

Parsed has positive value in case the sentence is parsed, negative otherwise;

Number of S, VP, NP and **PP** nodes in the sentence;

Signature is the sequence of PoS tags with and without punctuation;

[2] http://search.cpan.org/˜holsten/Lingua-DE-Sentence-0.07/Sentence.pm

[3] http://hypermedia.ids-mannheim.de/index.html

Children of Top-Level Nodes are two features representing the sequence of syntactic labels of the children of the root of the parse tree and the children of the highest S-node;

Branching Factor features express the average number of children of S, VP, NP and PP nodes in the parse;

Tree Depth is the average length of the path from the root to the leaves;

Per-word Entropy is a feature based on Genzel & Charniak's (2003) observation that paragraph-initial sentences have lower entropy than non-initial ones;

Sentence Probability according to a language model computed from the training data;

Character-level *n*-gram models are built using the CMU toolkit (Clarkson & Rosenfeld, 1997).

Since the parser we used produces dependency trees as an output, we could not distinguish between such features as **children of the root of the tree** and **children of the top-level S-node**. Apart from this minor change, we reimplemented the algorithm in every detail.

4.3 Our Features

For our algorithm we first selected the features of Sporleder & Lapata's (2006) system which performed best on the development set. These are relative position, the first and the first two words (**relPos, word1, word2**). Quote and final punctuation features, which were particularly helpful in Sporleder & Lapata's experiments on the German fiction data, turned out to be superfluous given the infrequency of quotations and the prevalent use of the period as sentence delimiter in our data.

We experimented with *text cohesion* features assuming that the paragraph structure crucially depends on cohesion and that paragraph breaks are likely to occur between sentences where cohesive links are weak. In order to estimate the degree of cohesion, we looked at lexical cohesion, pronominalization, discourse cues and information structure.

4.3.1 Lexical Cohesion

nounOver, verbOver: Similar to Sporleder & Lapata (2006), we introduced an overlap feature, but measured the degree of overlap as a number of common noun and verb lemmas between two adjacent sentences. We preferred lemmas over words in order to match all possible forms of the same word in German.

LCseg: Apart from the overlap, a boolean feature based on LCseg (Galley et al., 2003) marked whether the tool suggests that a new topic begins with the current sentence. This feature, relying on lexical chains, was supposed to provide more fine-grained information on the degree of similarity between two sentences.

4.3.2 Pronominalization

As Stark (1988) points out, humans tend to interpret over-reference as a clue for the beginning of a new paragraph: In a sentence, if a non-pronominal reference is preferred over a pronominal one where the pronoun would be admissible, humans are likely to mark this sentence as a paragraph-initial one. In order to check whether over-reference indeed correlates with paragraph-initial sentences, we described the way the biographee is referred to in the current and the previous sentences.

prevSPerson, currSPerson: This feature[4] with the values *NA, biographee, other* indicates whether there is a reference to the biographee or some other person in the sentence.

prevSRE, currSRE: This feature describes the biographee's referring expression and has three possible values: *NA, name, pronoun*.

Although our annotation distinguishes between first, last and full names, we found out that, for the PBI task, the distinction is spurious and unifying these three under the same category improves the results.

REchange: Since our classifiers assume feature independence and can not infer the information on the change in referring expression, we explicitly encoded that information by merging the values of the previous feature for the current and the preceding sentences into one, which has nine possible values (*name-name, NA-name, pronoun-name*, etc.).

[4] Prefixes **prevS-**, **currS-** stand for the previous and the current sentences respectively.

4.3.3 Discourse Cues

The intuition behind these features is that cue words and phrases are used to signal the relation between the current sentence and the preceding sentence or context (Mann & Thompson, 1988). Such connectives as *endlich (finally), abgesehen davon (apart from that), danach (afterwards)* explicitly mark a certain relation between the sentence they occur in and the preceding context. We hypothesize that the relations which hold across paragraph boundaries should differ from those which hold within paragraphs and that the same is true for the discourse cues. Absence of a connective is supposed to be informative as well, being more typical for paragraph-initial sentences.

Three features describe the connective of the current sentence. Another three features describe the one from the preceding sentence.

prevSCue, currSCue: This feature is the connective itself (*NA* in case of none).

prevSCueClass, currSCueClass: This feature represents the semantic class of the cue word or phrase as assigned by the IDS Mannheim. There are 25 values, including *NA* in case of no connective, altogether, with the most frequent values being *temporal, concessive, conclusive*, etc.

prevSProCue, currSProCue: The third binary feature marks whether the connective is proadverbial or not (*NA* if there is no connective). Being anaphors, proadverbials, such as *deswegen (because of that), darüber (about that)* explicitly link a sentence to the preceding one(s).

4.3.4 Information Structure

Information structure, which is in German to a large extent expressed by word order, provides additional clues to the degree of connectedness between two sentences. In respect to the PBI task, Stark (1988) reports that paragraph-initial sentences are often *theme-marking* which means that the subject of such sentences is not the first element. Given the lower frequency of paragraph-initial sentences, this feature can not be considered reliable, but in combination with others it provides an additional clue. In German, the first element best corresponds to the *prefield* (Vorfeld) – normally, the single constituent placed before the finite verb in the main clause.

currSVF encodes whether the constituent in the prefield is a *NP, PP, ADV, CARD*, or *Sub.Clause*. Values different from *NP* unambiguously represent theme-marking sentences, whereas the *NP* value may stand for both: theme-marking as well as not theme-marking sentence.

4.4 Discussion

Note, that we did not exclude text-initial sentences from the study because the encoding we used does not make such cases trivial for classification. Although some of the features refer to the previous sentence, none of them has to be necessarily realized and therefore none of them explicitly indicates the absence of the preceding sentence. For example, the label *NA* appears in cases where there is no discourse cue in the preceding sentence as well as in cases where there is no preceding sentence. The same holds for all other features prefixed with **prevS-**.

Another point concerns the use of pronominalization-based features. Sporleder & Lapata (2006) waive using such features because they consider pronominalization dependent on the paragraph structure and not the other way round. At the same time they mention speech and optical character recognition tasks as possible application domains for the PBI. There, pronouns are already given and need not be regenerated, hence for such applications features which utilize pronouns are absolutely appropriate. Unlike the recognition tasks, for multi-document summarization both decisions have to be made, and the order of the two tasks is not self-evident. The best decision would probably be to decide simultaneously on both using optimization methods (Roth & Yih, 2004; Marciniak & Strube, 2005). Generating pronouns before inserting boundaries seems as reasonable as doing it the other way round.

4.5 Feature Selection

We determine the relevant feature set and evaluate which features from this set contribute most to the performance of the system by the following procedures.

First, we follow an iterative algorithm similar to the wrapper approach for feature selection (Kohavi & John, 1997) using the development data and TiMBL. The feature subset selection algorithm performs a hill-climbing search along the

Feature set	F-measure
all	58.85%
–prevSCue	0.78%
–currSCue	0.32%
–currSCueClass	0.38%
–prevSCueClass	0.37%
–prevSProCue	1.02%
best	61.72%

Table 2: Removed features

Feature set	F-measure
relPos, word1, word2	48.06%
+currSRE	+10.50%
+currSVF	+0.49%
+currSPerson	+0.57%
+prevSPerson	+1.32%
best	60.94%

Table 3: Best features

feature space. We start with a model based on all available features. Then we train models obtained by removing one feature at a time. We choose the worst performing feature, namely the one whose removal gives the largest improvement based on the F-measure, and remove it from the model. We then train classifiers removing each of the remaining features separately from the enhanced model. The process is iteratively run as long as significant improvement is observed.

To measure the contribution of the relevant features we start with the three best features from Sporleder & Lapata (2006) (see Section 4.3) and train TiMBL combining the current feature set with each feature in turn. We then choose the best performing feature based on the F-measure and add it to the model. We iterate the process until all features are added to the three-feature system.

Thus, we optimize the default setting and obtain the information on what the paragraph structure crucially depends.

5 Results

Having trained our algorithm on the development data, we then determined the optimal feature combination and finally evaluated the performance on the previously unseen test data.

Table 2 and Table 3 present the ranking of the least and of the most beneficial features respectively. Somewhat surprising to us, Table 2 shows that basically *all* features capturing information on discourse cues actually worsened the performance of the classifier. The bad performance of the *prevSCue* and *currSCue* features may be caused by their extreme sparseness. To test these features reasonably, we plan to increase the data set size by an order of magnitude. Then, at least, it should be possible to determine which discourse cues, if any, are correlated with paragraph boundaries. The bad performance of the *prevSCueClass* and *currSCueClass* features may be caused by the categorization provided by the IDS. This question also requires further investigation, maybe with a different categorization.

Table 3 also provides interesting insights in the feature set. First, with only the three features *relPos, word1* and *word2* the baseline performs almost as well as the full feature set used by Sporleder & Lapata. Then, as expected, *currSRE* provides the largest gain in performance, followed by *currSVF*, *currSPerson* and *prevSPerson*. This result confirms our hypothesis that linguistically motivated features capturing information on pronominalization and information structure play an important role in determining paragraph segmentation.

The results of our system and the baselines for different classifiers (BT stands for BoosTexter and Ti for TiMBL) are summarized in Table 4. Accuracy is calculated by dividing the number of matches over the total number of test instances. Precision, recall and F-measure are obtained by considering true positives, false positives and false negatives. The latter metric, WindowDiff (Pevzner & Hearst, 2002), is supposed to overcome the disadvantage of the F-measure which penalizes near misses as harsh as more serious mistakes. The value of WindowDiff varies between 0 and 1, where a lesser count corresponds to better performance.

The significance of our results was computed using the χ^2 test. All results are significantly better (on the $p < 0.01$ level or below) than both baselines and the reimplemented version of Sporleder & Lapata's (2006) algorithm whose performance on our data is comparable to what the authors reported on their corpus of German fiction. Interestingly, TiMBL does much better than BoosTexter on Sporleder & Lapata's feature set. Apparently, Sporleder & Lapata's presupposition, that they would rely on many weak hypotheses,

	Accuracy	Precision	Recall	F-measure	WindowDiff
distance	52.16	37.98	31.88	34.66	.426
Galley	56.83	43.04	26.15	32.54	.416
development					
Sporleder_BT	71.96	80.15	30.46	44.15	.327
Sporleder_Ti	62.36	48.65	62.89	54.86	.338
all_BT	74.93	72.10	50.67	59.52	.286
all_Ti	70.54	59.81	57.91	58.85	.302
best_Ti	73.39	64.73	58.97	61.72	.280
test					
Sporleder_BT	68.76	80.15	28.61	42.16	.341
Sporleder_Ti	60.62	50.46	59.67	54.68	.345
all_BT	72.12	71.31	50.13	58.88	.286
all_Ti	67.13	59.14	56.40	57.74	.303
best_Ti	68.00	60.46	56.67	58.50	.302

Table 4: Results for the development and test sets with the two classifiers

does not hold. This is also confirmed by the results reported in Table 3 where only three of their features perform surprisingly strong. In contrast, on our feature set TiMBL and BoosTexter perform almost equally. However, BoosTexter achieves in all cases a much higher precision which is preferable over the higher recall provided by TiMBL.

6 Conclusion

In this paper, we proposed a novel approach to paragraph boundary identification based on linguistic features such as pronominalization, discourse cues and information structure. The results are significantly higher than all baselines and a reimplementation of Sporleder & Lapata's (2006) system and achieve an F-measure of about 59%.

We investigated to what extent the paragraph structure is determined by each of the three factors and came to the conclusion that it crucially depends on the use of pronouns and information structure. Surprisingly, discourse cues did not turn out to be useful for this task and even negatively affected the results which we explain by the extremely sparseness of the cues in our data.

It turned out that the best results could be achieved by a combination of surface features (*relPos, word1, word2*) and features capturing text cohesion. This indicates that paragraph boundary identification requires features usually used for style analysis and ones describing cohesive relations. Therefore, paragraph boundary identification is in fact a task which crosses the borders between content and style.

An obvious limitation of our study is that we trained and tested the algorithm on one-genre domain where pronouns are used extensively. Experimenting with different genres should shed light on whether our features are in fact domain-dependent. In the future, we also want to experiment with a larger data set for determining whether discourse cues really do not correlate with paragraph boundaries. Then, we will move on towards multi-document summarization, the application which motivates the research described here.

Acknowledments: This work has been funded by the Klaus Tschira Foundation, Heidelberg, Germany. The first author has been supported by a KTF grant (09.009.2004). We would also like to thank the three anonymous reviewers for their comments.

References

Bolshakov, Igor A. & Alexander Gelbukh (2001). Text segmentation into paragraph based on local text cohesion. In *Text, Speech and Dialogue*, pp. 158–166.

Brants, Thorsten (2000). TnT – A statistical Part-of-Speech tagger. In *Proceedings of the 6th Conference on Applied Natural Language Processing*, Seattle, Wash., 29 April – 4 May 2000, pp. 224–231.

Choi, Freddy Y. Y. (2000). Advances in domain independent linear text segmentation. In *Pro-*

ceedings of the 1st Conference of the North American Chapter of the Association for Computational Linguistics, Seattle, Wash., 29 April – 3 May, 2000, pp. 26–33.

Clarkson, Philip & Roni Rosenfeld (1997). Statistical language modeling. In *Proceedings of ESCA, EuroSpeech'97*. Rhodes, pp. 2707–2710.

Daelemans, Walter, Jakub Zavrel, Ko van der Sloot & Antal van den Bosch (2004). *TiMBL: Tilburg Memory Based Learner, version 5.1, Reference Guide*. Technical Report ILK 04-02: ILK Tilburg.

Foth, Kilian & Wolfgang Menzel (2006). Robust parsing: More with less. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics,* Trento, Italy, 3–7 April 2006, pp. 25–32.

Galley, Michel, Kathleen R. McKeown, Eric Fosler-Lussier & Hongyan Jing (2003). Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics,* Sapporo, Japan, 7–12 July 2003, pp. 562–569.

Genzel, Dmitriy (2005). A paragraph boundary detection system. In *Proceedings of the Sixth International Conference on Intelligent Text Processing and Computational Linguistics,* Mexico City, Mexico.

Genzel, Dmitriy & Eugene Charniak (2003). Variation of entropy and parse trees of sentences as a function of the sentence number. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing,* Sapporo, Japan, 11–12 July 2003, pp. 65–72.

Hearst, Marti A. (1997). TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

Irmscher, William F. (1972). *The Holt Guide to English*. New-York: Holt, Rinehart Winston.

Kohavi, Ron & George H. John (1997). Wrappers for feature subset selection. *Artificial Intelligence Journal,* 97(1-2):273–324.

Mann, William C. & Sandra A. Thompson (1988). Rhetorical structure theory. Toward a functional theory of text organization. *Text*, 8(3):243–281.

Marciniak, Tomacz & Michael Strube (2005). Beyond the pipeline: Discrete optimization in NLP. In *Proceedings of the 9th Conference on Computational Natural Language Learning,* Ann Arbor, Mich., USA, 29–30 June 2005, pp. 136–145.

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

Power, Richard, Donia Scott & Nadjet Bouayad-Agha (2003). Document structure. *Computational Linguistics*, 29(2):211–260.

Roth, Dan & Wen-tau Yih (2004). A linear programming formulation for global inference in natural language tasks. In *Proceedings of the 8th Conference on Computational Natural Language Learning,* Boston, Mass., USA, 6–7 May 2004, pp. 1–8.

Schapire, Robert E. & Yoram Singer (2000). BoosTexter: A boosting-based system for text categorization. *Machine Learning,* 39(2/3):135–168.

Schmid, Helmut (1997). Probabilistic part-of-speech tagging using decision trees. In Daniel Jones & Harold Somers (Eds.), *New Methods in Language Processing,* pp. 154–164. London, UK: UCL Press.

Sporleder, Caroline & Mirella Lapata (2004). Automatic paragraph identification: A study across languages and domains. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing,* Barcelona, Spain, 25–26 July 2004, pp. 72–79.

Sporleder, Caroline & Mirella Lapata (2006). Broad coverage paragraph segmentation across languages and domains. *ACM Transactions in Speech and Language Processing*. To appear.

Stark, Heather (1988). What do paragraph markings do? *Discourse Processes,* (11):275–303.