International Journal of Semantic Computing
Vol. 1, No. 3 (2007) 335–346

World Scientific
www.worldscientific.com

CROSS-GENRE FEATURE COMPARISONS FOR SPOKEN SENTENCE SEGMENTATION

SEBASTIEN CUENDET[*,‡], DILEK HAKKANI-TUR[*,§],
ELIZABETH SHRIBERG[*,†], JAMES FUNG[*,¶] and BENOIT FAVRE[*,‖]

[*]*International Computer Science Institute*
Berkeley, CA 94704, USA
www.icsi.berkeley.edu
[†]*SRI International, Menlo Park, CA 94025, USA*
www.speech.sri.com
[‡]*cuendet@icsi.berkeley.edu*
[§]*dilek@icsi.berkeley.edu*
[†]*ees@icsi.berkeley.edu*
[¶]*jgf@icsi.berkeley.edu*
[‖]*favre@icsi.berkeley.edu*

Automatic sentence segmentation of spoken language is an important precursor to downstream natural language processing. Previous studies combine lexical and prosodic features, but can impose significant computational challenges because of the large size of feature sets. Little is understood about which features most benefit performance, particularly for speech data from different speaking styles. We compare sentence segmentation for speech from broadcast news versus natural multi-party meetings, using identical lexical and prosodic feature sets across genres. Results based on boosting and forward selection for this task show that (1) features sets can be reduced with little or no loss in performance, and (2) the contribution of different feature types differs significantly by genre. We conclude that more efficient approaches to sentence segmentation and similar tasks can be achieved, especially if genre differences are taken into account.

Keywords: Sentence segmentation; prosody; feature selection.

1. Introduction

Recent speech processing tasks have focused on a range of genres that differ in speaking style — including news broadcasts, telephone conversations, lectures, and meetings. Such genres differ in many aspects, including vocabulary, syntax, turn-taking, discourse phenomena, disfluencies, paralinguistic effects, and prosody [1]. A typical approach to language processing tasks is to apply features and approaches developed for one genre, to another genre, using genre-specific training data where available. Other approaches use explicit adaptation techniques [2].

However, when matched training data is not available or in contexts in which speed and computational expense are important, it can be worthwhile to investigate

which features contribute most to a task and whether or not feature utility depends on the speaking style.

In this study we investigate the role of identically defined lexical and prosodic features, when applied to the same task across two very different speaking styles — broadcast news and face-to-face multi-party meetings. We focus on the task of automatic sentence segmentation, or finding boundaries of sentence units in the otherwise unannotated (devoid of punctuation, capitalization, or formatting) stream of words output by a speech recognizer. Sentence segmentation is of particular importance for speech understanding applications, because techniques aimed at semantic processing of speech input — such as parsing, machine translation, and information extraction — are often developed for text-based applications and thus assume the presence of overt sentence boundaries in their input [6, 5]. Sentence boundary annotation is also important for aiding human readability of the output of automatic speech recognition systems [3].

Previous approaches to sentence boundary detection in speech have combined lexical with prosodic features (such as pause, pitch, and energy features), using various machine learning techniques. One practical concern with such approaches is that although the gain from inclusion of prosodic features is considerable, these prosodic features require additional initial human effort and computational expense. Thus, research is needed to determine which features are most useful, and how feature utility differs for different styles of speech.

The goal of this study is to explore these questions for the task of sentence segmentation in news versus meeting speech. Specifically, we ask

(1) Can we achieve sentence boundary classification performance that is similar to an all-features performance result using only a small set of prosodic features?
(2) Do the different speaking styles differ in terms of which prosodic features are most useful for this task?

Results have implications not only for the task of sentence boundary detection, but more generally for prosodic modeling for natural language understanding across genres.

The following section describes the data set, features, and approach. Section 3 reports on experiments with lexical and prosodic features using boosting and forward selection of features, and provides further analysis of lexical and prosodic feature usage differences in the two different corpora. A summary and conclusions are provided in Section 4.

2. Method

2.1. *Data*

To study the differences between the meetings and BN speech for the task of sentence segmentation, we use the ICSI Meetings [9] and the TDT4 English Broadcast News [11] corpora. The ICSI Meeting Corpus is a collection of 75 meetings, including

simultaneous multi-channel audio recordings and word-level orthographic transcriptions. The meetings range in length from 17 to 103 minutes, but generally run just under an hour each, totalling 72 hours. We use a 73-meeting subset of this corpus that was also used in the previous research [9] with the same split into training, held-out, and test sets. The TDT4 Corpus was collected by the Linguistic Data Consortium (LDC) and includes multilingual raw material, newswire and other electronic text, web audio, broadcast radio and television. We use a subset of TDT4 English broadcast radio and television data in this study.

In the experiments to follow, classification models are trained on a set of data, tuned on a held-out set, and tested on an unseen test set, within each genre. Statistics on these data sets are shown in Table 1. The statistics in the tables are computed using the forced alignments between audio and reference transcriptions.

As shown in Table 1, the two different speaking styles differ significantly in mean sentence length, with sentences in meetings being only about half the length on average as those in broadcast news. Meetings (and conversational speech in general) tend to contain syntactically simpler sentences and significant pronominalization. News speech is typically read from a transcript, and more closely resembles written text. It contains for example appositions, center embeddings, and proper noun compounds, among other characteristics that contribute to longer sentences. Such differences are the result of both the more formal context of news speech and the speakers being professional speakers, as opposed to meetings where speakers are "common" people. Discourse phenomena also obviously differ across corpora, with meetings containing more turn exchanges, incomplete sentences, and higher rates of short backchannels (such as "yeah" and "uhhuh") than speech in news broadcasts.

2.2. *Features*

Sentence segmentation can be seen as a binary classification problem, in which each word boundary must be labeled as a sentence boundary or as a non-sentence boundary.[a] We define a large set of lexical and prosodic features, computed automatically based on the output of a speech recognizer, as described further, below.

Table 1. Data set statistics. Values are given in number of words, based on forced alignments.

	MRDA	TDT4
Training set size	90,000	150,000
Test set size	88,537	50,116
Held-out set size	110,851	23,363
Vocabulary size	10,887	18,697
Mean sentence length	6.54	14.69

[a]More detailed models may distinguish questions from statements, or complete from incomplete sentences.

Automatic speech recognition. Automatic speech recognition results for the ICSI Meetings data and TDT4 data were obtained using the state-of-the-art SRI CTS system [15] and SRI BN system [12], respectively. The meetings recognizer was trained using no acoustic data or transcripts from the analyzed meetings corpus. The word error rate for the recognizer output of the complete meetings corpus is 38.2%.

Recognition scores for the TDT4 corpus are not easily definable as only closed captions are available that frequently do not match well the actual words of the broadcast news shows. The estimated word error rate lies between 17% and 19%.

Lexical features. Previous work on sentence segmentation in broadcast news speech and in telephone conversations has used lexical and prosodic information [10, 4]. Additional work has studied the contribution of syntactic information [7]. Lexical features are usually represented as N-grams of words. In this work, lexical information is represented by 6 N-gram features for each word boundary: 3 unigrams, 2 bigrams and 1 trigram. Naming the word preceding the word boundary of interest as the *current* word, and the preceding and following words as the *previous* and *next* word respectively, the 6 lexical features are as follows:

- unigrams: {previous}, {current}, {next},
- bigrams: {current, next}, {previous, current},
- trigram: {previous, current, next}.

Prosodic features. Prosodic information is represented using mainly continuous values. We use 59 prosodic features, defined for and extracted from the regions around each inter-word boundary. The features include the pause duration at the boundary, normalized phone durations of the word preceding the boundary, and a variety of speaker-normalized pitch features and energy features preceding, following, and across the boundary. Features are an extension of similar features described in [10]. The extraction region around the boundary focuses on either one-word windows or brief time windows around the boundary. Measures include the maximum, minimum or average value in this time range. Pitch features are normalized by speaker, using the method to estimate a speaker's baseline pitch values described in [10]. Duration features, which measure the duration of the last vowel and the last rhyme in the word before the word boundary of interest, are normalized by statistics on the relevant phones in the training data. We also include "turn" features based on speaker changes. The turn features are computed differently on the two corpora. In TDT4, the speaker turns are determined by an alignment between the output of an external diarization system [13] and the words. Meetings are already broken up by channel with one channel per speaker and thus do not need diarization. A turn is added within a channel when the pause between two words is greater than 0.5 second. The reason for this is to match with the diarization system, where even if there is no speaker change, a non-speech region greater than 0.5 second segments the speaker turn.

Boosting and forward selection. For classification of word boundaries, we use the AdaBoost algorithm [8], which has been shown to be one of the best classifiers for this task [16]. Boosting aims to combine weak base classifiers to come up with a strong classifier. The learning algorithm is iterative. In each iteration, a different distribution or weighting over the training examples is used to give more emphasis to examples that are often misclassified by the preceding weak classifiers. For this approach we use the BoosTexter tool described in [8]. BoosTexter handles both discrete and continuous features, which allows for a convenient incorporation of the prosodic features described above (no binning is needed). The weak learners are one-level decision trees (stumps).

To analyze the difference in prosodic feature importance to sentence segmentation in the two genres, we rank features according to the forward selection algorithm (FSA). The FSA is an iterative algorithm that begins with an empty set of features. At each iteration, every feature that has not yet been selected is evaluated together with the previously selected features. The feature that yields the best performance is then added to the set of selected features and a new iteration, which considers the remaining features, begins. Although computationally expensive the FSA has the advantage of being intuitive and of capturing the correlation between two similar features. Indeed, once a feature has been selected, features with which it is highly correlated are less likely to be picked, since they would bring little additional knowledge to the classifier.

3. Experiments and Results

3.1. *Overall results*

Sentence segmentation quality is usually computed using one of two measures — F-measure or NIST error. F-measure is the harmonic mean of the recall and precision measures of the sentence boundaries hypothesized by the classifier to those assigned by human labelers. The NIST error rate is the ratio of the number of incorrect hypotheses made by the classifier to the number of reference sentence boundaries. If no boundaries are marked by sentence segmentation, this metric is 100%, but it can exceed 100%; the maximum error rate metric is the ratio of number of words to the number of correct boundaries. In this work, we report performance using only F-Measure.

3.2. *Lexical N-grams*

To characterize lexical differences across the two genres, we follow the comparative study reported in [14] in the context of text categorization, and utilize the widely used information gain (IG) metric. Given a term, *information gain* measures the amount of information obtained for the class prediction from the presence/absence of the term. In the case of a binary classification, the definition of the information

gain of a term t is a simplification of the definition presented in [14]:

$$\begin{aligned} G(t) = & -p(N)\log p(N) - p(S)\log p(S) \\ & + p(t)\left[p(N|t)\log p(N|t) + p(S|t)\log p(S|t)\right] \\ & + p(\bar{t})\left[p(N|\bar{t})\log p(N|\bar{t}) + p(S|\bar{t})\log p(S|\bar{t})\right] \end{aligned}$$

where S and N are the classes that designate a sentence boundary and a non-sentence boundary, respectively. Note that the IG score takes into account both classes, and we therefore do not need to take the average of the two classes. The $\chi 2$ statistic described in [14] is also useful to isolate the information of a term together with a particular class. However, in a two-class problem such as that examined here, the computation is symmetric, and therefore results are similar to those obtained using the IG score.

We consider each feature separately and compute the IG for each term that occurs in the feature vector (a term being a word in the case of the unigram feature, a bigram for the features represented by bigrams, etc.). For each genre and each of 6 lexical features, we extract the 10 terms that have the highest score. By doing that, we isolate the words that have a strong correlation with the occurrence of sentence boundaries. The underlying assumption is that if two genres are similar, the terms that are the best indicators of the beginning or the end of sentences should be similar in both genres.

Tables 2, 3 show the top 8 terms according to their IG score for 2 of the 3 unigram features and the bigram feature. IG values for the lexical features were computed on the held-out sets.

The tables show clear differences in word associations with sentence boundaries across genres. In meeting speech, as noted earlier, there are high rates of single-word backchannels such as `yeah`, `uhhuh`, and `right`. Since backchannels are treated as individual sentences in the annotation of this corpus, the presence of a backchannel word is a strong indicator for both a preceding and a following sentence boundary. (Note that not all cases of, for example, `right` are backchannels, since the word can

Table 2. Most frequent words for pre- and post-boundary unigram features.

Pre-boundary unigram		Post-boundary unigram	
MRDA	TDT4	MRDA	TDT4
yeah	the	yeah	i
uhhuh	to	so	of
okay	and	uhhuh	uh
right	of	and	to
the	a	but	but
huh	in	okay	he
i	washington	right	we
um	for	oh	i'm

Table 3. Most frequent words for pre-boundary — post-boundary bigram feature.

Pre-boundary — Post-boundary bigram	
MRDA	TDT4
yeah yeah	of the
uhhuh uhhuh	in the
yeah so	com the
yeah uhhuh	to the
yeah i	for the
okay so	court the
uhhuh yeah	glascoff coming
right so	at the

be used in other contexts. But in this data most backchannels use words that are more frequent in backchannels than in other contexts).

A quite different pattern is observed for the TDT4 corpus. In this case, words that have the highest IG score show no obvious correlation with the sentence boundary class. The explanation is that in BN, given the size of the data sets used, very few specific words appears repeatedly at sentence boundaries. Backchannels, fillers, and discourse markers are relatively rare, and a much larger set of words (including proper nouns) appear at sentence edges. As a consequence, words that obtain the best IG score for the TDT4 corpus are those that are highly correlated with the *non*-sentence boundary class distribution, i.e., words that are unlikely to end a sentence, such as `the`, `to`, or `a`. Note that this analysis does not hold for the *next word* feature, since words that begin a sentence have a pattern, even in the case of BN, as shown by the presence of `i`, `and`, and `but`.

Comparing the two lists (Tables 2 and 3) for the *current word* and the *next word* feature in the case of MRDA reveals the double usage of certain words like `yeah` or `okay`. In conversational speech, such words can be used as backchannels, which make the rank high in the *previous word* table. On the other hand, they are also used to start new sentences, which explains why they are so well ranked in the *next word* table.

The symmetry between the two classes in the IG computation allows some bigrams highly correlated with non-sentence boundary to have a high score for TDT4. For example, for the *current word — next word* feature, the bigram `of the` has the highest score, since it appears 536 times, but only twice with a sentence boundary, and thus 534 times with a non-sentence boundary.

3.3. *Prosodic features*

To rank the prosodic features according to their importance, we ran the FSA for 20 iterations. We used the BoosTexter tool [8] to train a classifier on the training data and evaluated the performance of the sentence segmentation on the held-out set. The

Table 4. Features selected for MRDA and TDT4; columns 2 and 3 show the F-Measure on the held-out set and the relative improvement from one feature to the next one, respectively. Column 4 shows the F-Measure on the test set. The last column is the F-Measure when using the feature alone.

Feature name	Held-out	Rel. impr.	Test	Alone
	MRDA			
PAU-DUR	60.0	—	62.2	60.0
F0K-WRD-DIFF-LOLO-N	61.0	1.7%	62.8	22.8
LAST-RHYME-NORM-DUR-PH	61.8	1.3%	63.6	12.6
PAU-DUR-PREV	62.5	1.2%	64.3	11.1
CROSS-SPKR PAUSE	63.0	0.7%	64.6	47.8
ENERGY-WIN-DIFF-HIHI-N	63.2	0.4%	65.3	20.4
LAST-RHYME-DUR-PH	63.8	0.5%	65.3	11.8
LAST-VOW-DUR-Z	63.8	0.4%	65.6	6.7
	TDT4			
PAU-DUR	56.0	—	55.4	56.0
F0K-DIFF-LAST-KBASELN	58.2	3.8%	57.0	34.4
F0K-WIN-DIFF-LOHI-N	59.4	2.1%	57.9	18.9
TURN-F	60.0	1.0%	58.5	35.5
PAU-DUR-PREV	60.2	0.4%	58.7	0.0
F0K-LR-MEAN-KBASELN	60.4	0.3%	58.7	0.0
F0K-DIFF-MNMN-N	60.5	0.2%	59.0	15.9
SLOPE-LAST-N	60.5	0.1%	58.9	0.8

feature with the best F-Measure was selected at each iteration. A classifier was then built on the training set and evaluated on the test set for each feature set. Table 4 reports the features that were selected until the F-Measure stopped increasing, and the corresponding performance on the development and the test sets.

The two sets each make significant use of the pause duration feature. The "pau-dur-prev" or duration of the pause one boundary earlier than the boundary of consideration is useful in MRDA in part because of the prevalence of single-word sentences such as backchannels, as described in Sec. 2.1. Both corpora make ample use of pitch features. TDT4 makes more use of "baseline" normalized pitch features that compare the location of a particular preboundary word in a speaker's pitch range to the value of a speaker's estimated baseline pitch. The closer the local pitch value is to the speaker's baseline, the more likely it is that the speaker is near a sentence end. This makes sense in that news (and read) speech is more careful and regular in intonation, whereas meeting speech is more informal and involves paralinguistic variation that can shift ending pitch values. MRDA makes use of pitch in the second feature selected, but this feature compares the pitch in words before and after the boundary, rather than the current pitch value to the speaker's estimated pitch floor. In this case, a large value of the feature indicates a sentence boundary, consistent with a large pitch reset. One interesting finding, perhaps counterintuitive at first, is that meetings make more use of duration features (of vowels or syllable rhymes) than do news broadcasts. Typically, there is a correlation between ends of major phrases and pre-boundary lengthening. Separate analysis revealed that

durational lengthening is indeed present for sentence boundaries in both corpora, but that in the case of news speech, lengthening occurs frequently elsewhere as well. That is, the register used in news broadcasts tends to insert frequent prominences and sub-sentential breaks, perhaps to keep the attention of the listener. Thus duration features may cause considerable false alarms in the case of broadcast news and are therefore less useful than they are for conversational speech. Finally, energy features do not appear to be as useful as pause, pitch, and duration features, across genres.

In MRDA, the previous pause feature, measuring the pause duration before the current word on the same channel, brings a relative improvement three times as large as in TDT4. The previous pause feature captures information about short utterances. When it is high and the current pause is high too, it suggests that the current sentence is only one word long. This is especially appropriate for conversational speech in which many utterances are backchannels, often one-word long, as already mentioned earlier in the study of lexical features. For TDT4, the improvement over the pause for the second feature selected is larger than for MRDA (3.8% vs. 1.7%). The smaller pauses at non-boundaries in more formal speech, as well as longer pauses between the end of a sentence and the beginning of the next one, explain this.

The cross-speaker pause feature is used only in MRDA by construction, since it measures not only the pause on a single channel (as the normal pause feature), but takes into account all the channels. In the case of MRDA, where every speaker has a microphone, the cross-speaker pause is not equivalent to the pause feature, whereas it is equivalent to that feature in TDT4. In addition to the pause and the two first pitch features, the turn feature provides significant benefit for TDT4. The turn feature is a binary feature that indicates a change of speaker. In BN, the speaker turn is automatically estimated by a diarization system, whereas on MRDA a turn is introduced every time there is a pause longer than 0.5 second. Thus in MRDA, the turn is highly correlated with the pause feature, whereas in TDT4 it is an independent input.

Further differences between MRDA and TDT4 are shown in Table 5. In MRDA, using only the lexical features results in significantly better performance than using only the prosodic features (+4.3% absolute). On the contrary, in TDT4, the prosodic model performs better than the lexical model (+1.4% absolute). The higher performance of the prosodic model reflects the more formal speech of TDT4, both because

Table 5. Comparison of the F-Measure with lexical features only, prosodic features only, and prosodic and lexical features together for MRDA and TDT4.

Corpus	Lexical	Prosodic	Both
MRDA	69.8	65.6	73.7
TDT4	58.0	59.4	61.8

Table 6. F-Measure with all 59 prosodic features and after 20 iterations of the FSA algorithm.

Corpus	Chance	All features (59)	FSA (20 iterations)
MRDA	15.8	65.6	66.0
TDT4	6.9	59.4	59.2

speakers make better use of prosody and because the lexical model is less strongly correlated with sentence boundaries than in conversational speech, as explained earlier.

Table 6 shows the performance of sentence segmentation for both corpora, when the classifier makes use of all the features and when it uses only the first 20 features selected by the FSA algorithm. While performance with all the features is expected to be better than that with only a subset, one can observe that performance is very close. On MRDA, performance with the reduced set of 20 features is actually better than when using all features. Going back to Table 4, one can see that performance with all the prosodic features is already reached by the reduced set of prosodic features after eight iterations of the FSA. In the case of TDT4, performance after eight iterations is 0.5% absolute less than that with all the features, but after four iterations only, the F-Measure score is less than 1% less than that with all the features. The score of the full set is reached at iteration 22 of the FSA on TDT4. Thus on TDT4 and MRDA, the same performance is reached by using 37% and 14% of the features, respectively.

Table 6 also shows the "chance performance" on one corpus. The chance performance assumes no knowledge about the data and simply classifies every example of the test set with respect to prior probability of each class in the training set. The performance reported is an average of the F-Measure over 10 runs. Comparing the chance performance with the score of the features when used alone (Table 4) shows that the performance with the first feature selected for both corpora (pause) is already four times as high as the chance performance. Some features picked later by the FSA have a performance worse than chance, but together with the previous features chosen they are able to improve the performance.

Reducing the set of features is important in terms of memory and CPU usage, as well as for computation time. For example, on the same machine, the training time is reduced by a factor of seven when using only eight features versus using all 59 features (2 h vs. 14 h).

4. Summary and Conclusions

We have compared lexical and prosodic sentence segmentation features for broadcast news and meeting speech, using identical feature sets and definitions for both genres. Analysis of sentence distributions in the two corpora shows significant differences in average sentence length, lexical, and prosodic features. For example, sentences in meetings are on average only half as long as those in broadcast conversations.

Whereas important lexical N-grams for meetings are positive cues associated with backchannels and various discourse phenomena, lexical N-grams for news speech are negative cues, i.e., N-grams in which a sentence boundary is highly unlikely.

Experiments on prosodic features using forward selection show that similar or even better performance can be achieved by using fewer features. Useful feature types, however, depend on the corpus. While both genres make use of pause and pitch information, pitch features contribute relatively more information in news speech. News speech makes use of local range information, in the form of features relative to the speaker's baseline, whereas pitch features in meetings capture pitch resets across inter-word boundaries. Interestingly, duration features, while correlated with sentence boundaries in both genres, are relatively more useful in meetings. Inspection reveals that in news speech, a problem for duration features is that they indicate many other locations, including prominent syllables and subsentential boundaries. Energy features appear to be less important than pause, pitch, and duration features in both genres.

Sentence segmentation is one of a number of tasks in which lexical and prosodic features can be combined for better performance. Based on results found here, we conclude that feature selection can produce similar or even better performance results, but that the particular features depend on the speech genre. Although in this case training data was available for both genres, information about which features benefit which genre should be even more important when adapting models to data for which little or no matched training data is available.

Acknowledgments

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under contract No. HR0011-06-C-0023 and contract No. NBCHD030010 and IM2, a research network funded by the Swiss National Science Foundation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the view of DARPA. The authors thank Matthias Zimmermann, Yang Liu, and Mathew Magimai Doss for their help and suggestions.

References

[1] D. Biber, *Variation across Speech and Writing*, Cambridge University Press, Cambridge, 1988.

[2] S. Cuendet, D. Hakkani-Tür and G. Tur, Model adaptation for sentence unit segmentation from speech, in *Proceedings of SLT*, Aruba, 2006.

[3] D. Jones, W. Shen, E. Shriberg, A. Stolcke, T. Kamm and D. Reynolds, Two experiments comparing reading with listening for human processing of conversational telephone speech, in *Proceedings of EUROSPEECH*, 2005, pp. 1145–1148.

[4] Y. Liu, E. Shriberg, A. Stolcke, B. Peskin, J. Ang, D. Hillard, M. Ostendorf, M. Tomalin, P. Woodland and M. Harper, Structural metadata research in the EARS program, in *Proceedings of ICASSP*, 2005.

[5] J. Makhoul, A. Baron, I. Bulyko, L. Nguyen, L. Ramshaw, D. Stallard, R. Schwartz and B. Xiang, The effects of speech recognition and punctuation on information extraction performance, in *Proceedings of Interspeech*, Lisbon, 2005, pp. 57–60.
[6] J. Mrozinski, E. W. D. Whittaker, P. Chatain and S. Furui, Automatic sentence segmentation of speech for automatic summarization, in *Proceedings of ICASSP*, Vol. 1, Philadelphia, PA, 2005, pp. I–I.
[7] B. Roark, Y. Liu, M. Harper, R. Stewart, M. Lease, M. Snover, I. Shafran, B. Dorr, J. Hale, A. Krasnyanskaya and L. Yung, Reranking for sentence boundary detection in conversational speech, in *Proceedings of ICASSP*, Toulouse, France, 2006.
[8] R. E. Schapire and Y. Singer, Boostexter: A boosting-based system for text categorization, *Machine Learning* **39**(2/3) (2000) 135–168.
[9] E. Shriberg, R. Dhillon, S. Bhagat, J. Ang and H. Carvey, The ICSI meeting recorder dialog act (MRDA) corpus, in *Proceedings of SigDial Workshop*, Boston, MA, 2004.
[10] E. Shriberg, A. Stolcke, D. Hakkani-Tür and G. Tur, Prosody-based automatic segmentation of speech into sentences and topics, *Speech Communication* **3** (2000) 2037–2040.
[11] S. Strassel and M. Glenn, Creating the annotated TDT-4 Y2003 evaluation corpus, in *TDT 2003 Evaluation Workshop*, NIST, 2003.
[12] A. Venkataraman, A. Stolcke, W. Wang, D. Vergyri, V. Gadde and J. Zheng, SRI's 2004 broadcast news speech to text system, in *EARS Rich Transcription 2004 Workshop*, Palisades, NY, 2004.
[13] C. Wooters, J. Fung, B. Peskin and X. Anguera, Towards robust speaker segmentation: ICSI-SRI Fall 2004 diarization system, in *RT-04F Workshop*, 2004.
[14] Y. Yang and J. Pedersen, A comparative study on feature selection in text categorization, in *Proceedings of ICML*, Nashville, US, 1997, pp. 412–420.
[15] Q. Zhu, A. Stolcke, B. Chen and N. Morgan, Using MLP features in SRI's conversational speech recognition system, in *Proceedings of INTERSPEECH*, Lisbon, Portugal, 2005, pp. 2141–2144.
[16] M. Zimmermann, D. Hakkani-Tür, J. Fung, N. Mirghafori, E. Shriberg and Y. Liu, The ICSI+ multi-lingual sentence segmentation system, in *Proceedings of ICSLP*, Pittsburgh, PA, 2006.