

Using Conditional Random Fields For Sentence Boundary Detection In Speech

Yang Liu

ICSI, Berkeley

yangl@icsi.berkeley.edu

Andreas Stolcke

SRI and ICSI

stolcke,ees@speech.sri.com

Elizabeth Shriberg

Mary Harper

Purdue University

harper@ecn.purdue.edu

Abstract

Sentence boundary detection in speech is important for enriching speech recognition output, making it easier for humans to read and downstream modules to process. In previous work, we have developed hidden Markov model (HMM) and maximum entropy (Maxent) classifiers that integrate textual and prosodic knowledge sources for detecting sentence boundaries. In this paper, we evaluate the use of a conditional random field (CRF) for this task and relate results with this model to our prior work. We evaluate across two corpora (conversational telephone speech and broadcast news speech) on both human transcriptions and speech recognition output. In general, our CRF model yields a lower error rate than the HMM and Maxent models on the NIST sentence boundary detection task in speech, although it is interesting to note that the best results are achieved by three-way voting among the classifiers. This probably occurs because each model has different strengths and weaknesses for modeling the knowledge sources.

1 Introduction

Standard speech recognizers output an unstructured stream of words, in which the important structural features such as sentence boundaries are missing.

Sentence segmentation information is crucial and assumed in most of the further processing steps that one would want to apply to such output: tagging and parsing, information extraction, summarization, among others.

1.1 Sentence Segmentation Using HMM

Most prior work on sentence segmentation (Shriberg et al., 2000; Gotoh and Renals, 2000; Christensen et al., 2001; Kim and Woodland, 2001; NIST-RT03F, 2003) have used an HMM approach, in which the word/tag sequences are modeled by N-gram language models (LMs) (Stolcke and Shriberg, 1996). Additional features (mostly related to speech prosody) are modeled as observation likelihoods attached to the N-gram states of the HMM (Shriberg et al., 2000). Figure 1 shows the graphical model representation of the variables involved in the HMM for this task. Note that the words appear in both the states¹ and the observations, such that the word stream constrains the possible hidden states to matching words; the ambiguity in the task stems entirely from the choice of events. This architecture differs from the one typically used for sequence tagging (e.g., part-of-speech tagging), in which the “hidden” states represent only the events or tags. Empirical investigations have shown that omitting words in the states significantly degrades system performance for sentence boundary detection (Liu, 2004). The observation probabilities in the HMM, implemented using a decision tree classifier, capture the probabilities of generating the prosodic features

¹In this sense, the states are only partially “hidden”.

$P(F_i|E_i, W_i)$.² An N-gram LM is used to calculate the transition probabilities:

$$\begin{aligned} P(W_i E_i | W_1 E_1 \dots W_{i-1} E_{i-1}) = \\ P(W_i | W_1 E_1 \dots W_{i-1} E_{i-1}) \times \\ P(E_i | W_1 E_1 \dots W_{i-1} E_{i-1} E_i) \end{aligned}$$

In the HMM, the forward-backward algorithm is used to determine the event with the highest posterior probability for each interword boundary:

$$\hat{E}_i = \arg \max_{E_i} P(E_i | W, F) \quad (1)$$

The HMM is a generative modeling approach since it describes a stochastic process with hidden variables (sentence boundary) that produces the observable data. This HMM approach has two main drawbacks. First, standard training methods maximize the joint probability of observed and hidden events, as opposed to the posterior probability of the correct hidden variable assignment given the observations, which would be a criterion more closely related to classification performance. Second, the N-gram LM underlying the HMM transition model makes it difficult to use features that are highly correlated (such as words and POS labels) without greatly increasing the number of model parameters, which in turn would make robust estimation difficult. More details about using textual information in the HMM system are provided in Section 3.

1.2 Sentence Segmentation Using Maxent

A maximum entropy (Maxent) posterior classification method has been evaluated in an attempt to overcome some of the shortcomings of the HMM approach (Liu et al., 2004; Huang and Zweig, 2002). For a boundary position i , the Maxent model takes the exponential form:

$$P(E_i | T_i, F_i) = \frac{1}{Z_\lambda(T_i, F_i)} e^{\sum_k \lambda_k g_k(E_i, T_i, F_i)} \quad (2)$$

where $Z_\lambda(T_i, F_i)$ is a normalization term and T_i represents textual information. The indicator functions $g_k(E_i, T_i, F_i)$ correspond to features defined over events, words, and prosody. The parameters in

²In the prosody model implementation, we ignore the word identity in the conditions, only using the timing or word alignment information.

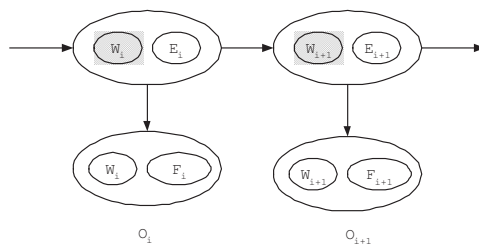


Figure 1: A graphical model of HMM for the sentence boundary detection problem. Only one word+event pair is depicted in each state, but in a model based on N-grams, the previous $N - 1$ tokens would condition the transition to the next state. O are observations consisting of words W and prosodic features F , and E are sentence boundary events.

Maxent are chosen to maximize the conditional likelihood $\prod_i P(E_i | T_i, F_i)$ over the training data, better matching the classification accuracy metric. The Maxent framework provides a more principled way to combine the largely correlated textual features, as confirmed by the results of (Liu et al., 2004); however, it does not model the state sequence.

A simple combination of the results from the Maxent and HMM was found to improve upon the performance of either model alone (Liu et al., 2004) because of the complementary strengths and weaknesses of the two models. An HMM is a generative model, yet it is able to model the sequence via the forward-backward algorithm. Maxent is a discriminative model; however, it attempts to make decisions locally, without using sequential information.

A conditional random field (CRF) model (Laferty et al., 2001) combines the benefits of the HMM and Maxent approaches. Hence, in this paper we will evaluate the performance of the CRF model and relate the results to those using the HMM and Maxent approaches on the sentence boundary detection task. The rest of the paper is organized as follows. Section 2 describes the CRF model and discusses how it differs from the HMM and Maxent models. Section 3 describes the data and features used in the models to be compared. Section 4 summarizes the experimental results for the sentence boundary detection task. Conclusions and future work appear in Section 5.

2 CRF Model Description

A CRF is a random field that is globally conditioned on an observation sequence O . CRFs have been successfully used for a variety of text processing tasks (Lafferty et al., 2001; Sha and Pereira, 2003; McCallum and Li, 2003), but they have not been widely applied to a speech-related task with both acoustic and textual knowledge sources. The top graph in Figure 2 is a general CRF model. The states of the model correspond to event labels E . The observations O are composed of the textual features, as well as the prosodic features. The most likely event sequence \hat{E} for the given input sequence (observations) O is

$$\hat{E} = \arg \max_E \frac{e^{\sum_k \lambda_k G_k(E, O)}}{Z_\lambda(O)} \quad (3)$$

where the functions G are potential functions over the events and the observations, and Z_λ is the normalization term:

$$Z_\lambda(O) = \sum_E e^{\sum_k \lambda_k G_k(E, O)} \quad (4)$$

Even though a CRF itself has no restriction on the potential functions $G_k(E, O)$, to simplify the model (considering computational cost and the limited training set size), we use a first-order CRF in this investigation, as at the bottom of Figure 2. In this model, an observation O_i (consisting of textual features T_i and prosodic features F_i) is associated with a state E_i .

The model is trained to maximize the conditional log-likelihood of a given training set. Similar to the Maxent model, the conditional likelihood is closely related to the individual event posteriors used for classification, enabling this type of model to explicitly optimize discrimination of correct from incorrect labels. The most likely sequence is found using the Viterbi algorithm.³

A CRF differs from an HMM with respect to its training objective function (joint versus conditional likelihood) and its handling of dependent word features. Traditional HMM training does not maximize the posterior probabilities of the correct labels; whereas, the CRF directly estimates posterior

³The forward-backward algorithm would most likely be better here, but it is not implemented in the software we used (McCallum, 2002).

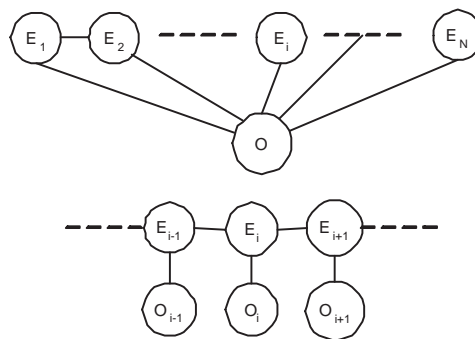


Figure 2: Graphical representations of a general CRF and the first-order CRF used for the sentence boundary detection problem. E represent the state tags (i.e., sentence boundary or not). O are observations consisting of words W or derived textual features T and prosodic features F .

boundary label probabilities $P(E|O)$. The underlying N-gram sequence model of an HMM does not cope well with multiple representations (features) of the word sequence (e.g., words, POS), especially when the training set is small; however, the CRF model supports simultaneous correlated features, and therefore gives greater freedom for incorporating a variety of knowledge sources. A CRF differs from the Maxent method with respect to its ability to model sequence information. The primary advantage of the CRF over the Maxent approach is that the model is optimized globally over the entire sequence; whereas, the Maxent model makes a local decision, as shown in Equation (2), without utilizing any state dependency information.

We use the Mallet package (McCallum, 2002) to implement the CRF model. To avoid overfitting, we employ a Gaussian prior with a zero mean on the parameters (Chen and Rosenfeld, 1999), similar to what is used for training Maxent models (Liu et al., 2004).

3 Experimental Setup

3.1 Data and Task Description

The sentence-like units in speech are different from those in written text. In conversational speech, these units can be well-formed sentences, phrases, or even a single word. These units are called SUs in the DARPA EARS program. SU boundaries, as

well as other structural metadata events, were annotated by LDC according to an annotation guideline (Strassel, 2003). Both the transcription and the recorded speech were used by the annotators when labeling the boundaries.

The SU detection task is conducted on two corpora: Broadcast News (BN) and Conversational Telephone Speech (CTS). BN and CTS differ in genre and speaking style. The average length of SUs is longer in BN than in CTS, that is, 12.35 words (standard deviation 8.42) in BN compared to 7.37 words (standard deviation 8.72) in CTS. This difference is reflected in the frequency of SU boundaries: about 14% of interword boundaries are SUs in CTS compared to roughly 8% in BN. Training and test data for the SU detection task are those used in the NIST Rich Transcription 2003 Fall evaluation. We use both the development set and the evaluation set as the test set in this paper in order to obtain more meaningful results. For CTS, there are about 40 hours of conversational data (around 480K words) from the Switchboard corpus for training and 6 hours (72 conversations) for testing. The BN data has about 20 hours of Broadcast News shows (about 178K words) in the training set and 3 hours (6 shows) in the test set. Note that the SU-annotated training data is only a subset of the data used for the speech recognition task because more effort is required to annotate the boundaries.

For testing, the system determines the locations of sentence boundaries given the word sequence W and the speech. The SU detection task is evaluated on both the reference human transcriptions (REF) and speech recognition outputs (STT). Evaluation across transcription types allows us to obtain the performance for the best-case scenario when the transcriptions are correct; thus factoring out the confounding effect of speech recognition errors on the SU detection task. We use the speech recognition output obtained from the SRI recognizer (Stolcke et al., 2003).

System performance is evaluated using the official NIST evaluation tools.⁴ System output is scored by first finding a minimum edit distance alignment between the hypothesized word string and the refer-

⁴See <http://www.nist.gov/speech/tests/rt/rt2003/fall/> for more details about scoring.

ence transcriptions, and then comparing the aligned event labels. The SU error rate is defined as the total number of deleted or inserted SU boundary events, divided by the number of true SU boundaries. In addition to this **NIST SU error metric**, we use the total number of interword boundaries as the denominator, and thus obtain results for the **per-boundary-based metric**.

3.2 Feature Extraction and Modeling

To obtain a good-quality estimation of the conditional probability of the event tag given the observations $P(E_i|O_i)$, the observations should be based on features that are discriminative of the two events (SU versus not). As in (Liu et al., 2004), we utilize both textual and prosodic information.

We extract prosodic features that capture duration, pitch, and energy patterns associated with the word boundaries (Shriberg et al., 2000). For all the modeling methods, we adopt a modular approach to model the prosodic features, that is, a decision tree classifier is used to model them. During testing, the decision tree prosody model estimates posterior probabilities of the events given the associated prosodic features for a word boundary. The posterior probability estimates are then used in various modeling approaches in different ways as described later.

Since words and sentence boundaries are mutually constraining, the word identities themselves (from automatic recognition or human transcriptions) constitute a primary knowledge source for sentence segmentation. We also make use of various automatic taggers that map the word sequence to other representations. Tagged versions of the word stream are provided to support various generalizations of the words and to smooth out possibly undertrained word-based probability estimates. These tags include part-of-speech tags, syntactic chunk tags, and automatically induced word classes. In addition, we use extra text corpora, which were not annotated according to the guideline used for the training and test data (Strassel, 2003). For BN, we use the training corpus for the LM for speech recognition. For CTS, we use the Penn Treebank Switchboard data. There is punctuation information in both, which we use to approximate SUs as defined in the annotation guideline (Strassel, 2003).

As explained in Section 1, the prosody model and

Table 1: Knowledge sources and their representations in different modeling approaches: HMM, Maxent, and CRF.

	HMM	Maxent	CRF
	generative model		conditional approach
Sequence information	yes	no	yes
LDC data set (words or tags)	LM	N-grams as indicator functions	
Probability from prosody model	real-valued	cumulatively binned	
Additional text corpus	N-gram LM	binned posteriors	
Speaker turn change	in prosodic features	a separate feature, in addition to being in the prosodic feature set	
Compound feature	no	POS tags and decisions from prosody model	

the N-gram LM can be integrated in an HMM. When various textual information is used, jointly modeling words and tags may be an effective way to model the richer feature set; however, a joint model requires more parameters. Since the training set for the SU detection task in the EARS program is quite limited, we use a loosely coupled approach:

- Linearly combine three LMs: the word-based LM from the LDC training data, the automatic-class-based LMs, and the word-based LM trained from the additional corpus.
- These interpolated LMs are then combined with the prosody model via the HMM. The posterior probabilities of events at each boundary are obtained from this step, denoted as $P_{HMM}(E_i|W, C, F)$.
- Apply the POS-based LM alone to the POS sequence (obtained by running the POS tagger on the word sequence W) and generate the posterior probabilities for each word boundary $P_{posLM}(E_i|POS)$, which are then combined from the posteriors from the previous step, i.e., $P_{final}(E_i|T, F) = P_{HMM}(E_i|W, C, F) + P_{posLM}(E_i|P)$.

The features used for the CRF are the same as those used for the Maxent model devised for the SU detection task (Liu et al., 2004), briefly listed below.

- N-grams of words or various tags (POS tags, automatically induced classes). Different N s and different position information are used (N varies from one through four).

- The cumulative binned posterior probabilities from the decision tree prosody model.
- The N-gram LM trained from the extra corpus is used to estimate posterior event probabilities for the LDC-annotated training and test sets, and these posteriors are then thresholded to yield binary features.
- Other features: speaker or turn change, and compound features of POS tags and decisions from the prosody model.

Table 1 summarizes the features and their representations used in the three modeling approaches. The same knowledge sources are used in these approaches, but with different representations. The goal of this paper is to evaluate the ability of these three modeling approaches to combine prosodic and textual knowledge sources, not in a rigidly parallel fashion, but by exploiting the inherent capabilities of each approach. We attempt to compare the models in as parallel a fashion as possible; however, it should be noted that the two discriminative methods better model the textual sources and the HMM better models prosody given its representation in this study.

4 Experimental Results and Discussion

SU detection results using the CRF, HMM, and Maxent approaches individually, on the reference transcriptions or speech recognition output, are shown in Tables 2 and 3 for CTS and BN data, respectively. We present results when different knowledge sources are used: word N-gram only, word N-gram and prosodic information, and using all the

Table 2: Conversational telephone speech SU detection results reported using the NIST SU error rate (%) and the boundary-based error rate (% in parentheses) using the HMM, Maxent, and CRF individually and in combination. Note that the ‘all features’ condition uses all the knowledge sources described in Section 3.2. ‘Vote’ is the result of the majority vote over the three modeling approaches, each of which uses all the features. The baseline error rate when assuming there is no SU boundary at each word boundary is 100% for the NIST SU error rate and 15.7% for the boundary-based metric.

Conversational Telephone Speech				
		HMM	Maxent	CRF
REF	word N-gram	42.02 (6.56)	43.70 (6.82)	37.71 (5.88)
	word N-gram + prosody	33.72 (5.26)	35.09 (5.47)	30.88 (4.82)
	all features	31.51 (4.92)	30.66 (4.78)	29.47 (4.60)
	Vote: 29.30 (4.57)			
STT	word N-gram	53.25 (8.31)	53.92 (8.41)	50.20 (7.83)
	word N-gram + prosody	44.93 (7.01)	45.50 (7.10)	43.12 (6.73)
	all features	43.05 (6.72)	43.02 (6.71)	42.00 (6.55)
	Vote: 41.88 (6.53)			

features described in Section 3.2. The word N-grams are from the LDC training data and the extra text corpora. ‘All the features’ means adding textual information based on tags, and the ‘other features’ in the Maxent and CRF models as well. The detection error rate is reported using the NIST SU error rate, as well as the per-boundary-based classification error rate (in parentheses in the table) in order to factor out the effect of the different SU priors. Also shown in the tables are the majority vote results over the three modeling approaches when all the features are used.

4.1 CTS Results

For CTS, we find from Table 2 that the CRF is superior to both the HMM and the Maxent model across all conditions (the differences are significant at $p < 0.05$). When using only the word N-gram information, the gain of the CRF is the greatest, with the differences among the models diminishing as more features are added. This may be due to the impact of the sparse data problem on the CRF or simply due to the fact that differences among modeling approaches are less when features become stronger, that is, the good features compensate for the weaknesses in models. Notice that with fewer knowledge sources (e.g., using only word N-gram and prosodic information), the CRF is able to achieve performance similar to or even better than other methods using all the knowl-

edges sources. This may be useful when feature extraction is computationally expensive.

We observe from Table 2 that there is a large increase in error rate when evaluating on speech recognition output. This happens in part because word information is inaccurate in the recognition output, thus impacting the effectiveness of the LMs and lexical features. The prosody model is also affected, since the alignment of incorrect words to the speech is imperfect, thereby degrading prosodic feature extraction. However, the prosody model is more robust to recognition errors than textual knowledge, because of its lesser dependence on word identity. The results show that the CRF suffers most from the recognition errors. By focusing on the results when only word N-gram information is used, we can see the effect of word errors on the models. The SU detection error rate increases more in the STT condition for the CRF model than for the other models, suggesting that the discriminative CRF model suffers more from the mismatch between the training (using the reference transcription) and the test condition (features obtained from the errorful words).

We also notice from the CTS results that when only word N-gram information is used (with or without combining with prosodic information), the HMM is superior to the Maxent; only when various additional textual features are included in the feature set does Maxent show its strength compared to

Table 3: Broadcast news SU detection results reported using the NIST SU error rate (%) and the boundary-based error rate (% in parentheses) using the HMM, Maxent, and CRF individually and in combination. The baseline error rate is 100% for the NIST SU error rate and 7.2% for the boundary-based metric.

Broadcast News				
		HMM	Maxent	CRF
REF	word N-gram	80.44 (5.83)	81.30 (5.89)	74.99 (5.43)
	word N-gram + prosody	59.81 (4.33)	59.69 (4.33)	54.92 (3.98)
	all features	48.72 (3.53)	48.61 (3.52)	47.92 (3.47)
	Vote: 46.28 (3.35)			
STT	word N-gram	84.71 (6.14)	86.13 (6.24)	80.50 (5.83)
	word N-gram + prosody	64.58 (4.68)	63.16 (4.58)	59.52 (4.31)
	all features	55.37 (4.01)	56.51 (4.10)	55.37 (4.01)
	Vote: 54.29 (3.93)			

the HMM, highlighting the benefit of Maxent’s handling of the textual features.

The combined result (using majority vote) of the three approaches in Table 2 is superior to any model alone (the improvement is not significant though). Previously, it was found that the Maxent and HMM posteriors combine well because the two approaches have different error patterns (Liu et al., 2004). For example, Maxent yields fewer insertion errors than HMM because of its reliance on different knowledge sources. The toolkit we use for the implementation of the CRF does not generate a posterior probability for a sequence; therefore, we do not combine the system output via posterior probability interpolation, which is expected to yield better performance.

4.2 BN Results

Table 3 shows the SU detection results for BN. Similar to the patterns found for the CTS data, the CRF consistently outperforms the HMM and Maxent, except on the STT condition when all the features are used. The CRF yields relatively less gain over the other approaches on BN than on CTS. One possible reason for this difference is that there is more training data for the CTS task, and both the CRF and Maxent approaches require a relatively larger training set than the HMM. Overall the degradation on the STT condition for BN is smaller than on CTS. This can be easily explained by the difference in word error rates, 22.9% on CTS and 12.1% on BN. Finally, the vote among the three approaches outperforms any model on both the REF and STT condi-

tions, and the gain from voting is larger for BN than CTS.

Comparing Table 2 and Table 3, we find that the NIST SU error rate on BN is generally higher than on CTS. This is partly because the NIST error rate is measured as the percentage of errors per reference SU, and the number of SUs in CTS is much larger than for BN, giving a large denominator and a relatively lower error rate for the same number of boundary detection errors. Another reason is that the training set is smaller for BN than for CTS. Finally, the two genres differ significantly: CTS has the advantage of the frequent backchannels and first person pronouns that provide good cues for SU detection. When the boundary-based classification metric is used (results in parentheses), the SU error rate is lower on BN than on CTS; however, it should also be noted that the baseline error rate (i.e., the priors of the SUs) is lower on BN than CTS.

5 Conclusion and Future Work

Finding sentence boundaries in speech transcriptions is important for improving readability and aiding downstream language processing modules. In this paper, prosodic and textual knowledge sources are integrated for detecting sentence boundaries in speech. We have shown that a discriminatively trained CRF model is a competitive approach for the sentence boundary detection task. The CRF combines the advantages of being discriminatively trained and able to model the entire sequence, and so it outperforms the HMM and Maxent approaches

consistently across various testing conditions. The CRF takes longer to train than the HMM and Maxent models, especially when the number of features becomes large; the HMM requires the least training time of all approaches. We also find that as more features are used, the differences among the modeling approaches decrease. We have explored different approaches to modeling various knowledge sources in an attempt to achieve good performance for sentence boundary detection. Note that we have not fully optimized each modeling approach. For example, for the HMM, using discriminative training methods is likely to improve system performance, but possibly at a cost of reducing the accuracy of the combined system.

In future work, we will examine the effect of Viterbi decoding versus forward-backward decoding for the CRF approach, since the latter better matches the classification accuracy metric. To improve SU detection results on the STT condition, we plan to investigate approaches that model recognition uncertainty in order to mitigate the effect of word errors. Another future direction is to investigate how to effectively incorporate prosodic features more directly in the Maxent or CRF framework, rather than using a separate prosody model and then binning the resulting posterior probabilities.

Important ongoing work includes investigating the impact of SU detection on downstream language processing modules, such as parsing. For these applications, generating probabilistic SU decisions is crucial since that information can be more effectively used by subsequent modules.

6 Acknowledgments

The authors thank the anonymous reviewers for their valuable comments, and Andrew McCallum and Aron Culotta at the University of Massachusetts and Fernando Pereira at the University of Pennsylvania for their assistance with their CRF toolkit. This work has been supported by DARPA under contract MDA972-02-C-0038, NSF-STIMULATE under IRI-9619921, NSF KDI BCS-9980054, and ARDA under contract MDA904-03-C-1788. Distribution is unlimited. Any opinions expressed in this paper are those of the authors and do not reflect the funding agencies. Part of the work was carried out while the last author was on leave from Purdue University and at NSF.

References

S. Chen and R. Rosenfeld. 1999. A Gaussian prior for smoothing maximum entropy models. Technical report, Carnegie Mellon University.

- H. Christensen, Y. Gotoh, and S. Renal. 2001. Punctuation annotation using statistical prosody models. In *ISCA Workshop on Prosody in Speech Recognition and Understanding*.
- Y. Gotoh and S. Renals. 2000. Sentence boundary detection in broadcast speech transcripts. In *Proceedings of ISCA Workshop: Automatic Speech Recognition: Challenges for the New Millennium ASR-2000*, pages 228–235.
- J. Huang and G. Zweig. 2002. Maximum entropy model for punctuation annotation from speech. In *Proceedings of the International Conference on Spoken Language Processing*, pages 917–920.
- J. Kim and P. C. Woodland. 2001. The use of prosody in a combined system for punctuation generation and speech recognition. In *Proceedings of the European Conference on Speech Communication and Technology*, pages 2757–2760.
- J. Lafferty, A. McCallum, and F. Pereira. 2001. Conditional random field: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the International Conference on Machine Learning*, pages 282–289.
- Y. Liu, A. Stolcke, E. Shriberg, and M. Harper. 2004. Comparing and combining generative and posterior probability models: Some advances in sentence boundary detection in speech. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.
- Y. Liu. 2004. *Structural Event Detection for Rich Transcription of Speech*. Ph.D. thesis, Purdue University.
- A. McCallum and W. Li. 2003. Early results for named entity recognition with conditional random fields. In *Proceedings of the Conference on Computational Natural Language Learning*.
- A. McCallum. 2002. Mallet: A machine learning for language toolkit. <http://mallet.cs.umass.edu>.
- NIST-RT03F. 2003. RT-03F workshop agenda and presentations. <http://www.nist.gov/speech/tests/rt/rt2003/fall/presentations/>, November.
- F. Sha and F. Pereira. 2003. Shallow parsing with conditional random fields. In *Proceedings of Human Language Technology Conference / North American Chapter of the Association for Computational Linguistics annual meeting*.
- E. Shriberg, A. Stolcke, D. Hakkani-Tur, and G. Tur. 2000. Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication*, pages 127–154.
- A. Stolcke and E. Shriberg. 1996. Automatic linguistic segmentation of conversational speech. In *Proceedings of the International Conference on Spoken Language Processing*, pages 1005–1008.
- A. Stolcke, H. Franco, R. Gadde, M. Graciarena, K. Precoda, A. Venkataraman, D. Vergyri, W. Wang, and J. Zheng. 2003. Speech-to-text research at SRI-ICSI-UW. <http://www.nist.gov/speech/tests/rt/rt2003/spring/presentations/index.htm>.
- S. Strassel, 2003. *Simple Metadata Annotation Specification V5.0*. Linguistic Data Consortium.