USING CONDITIONAL RANDOM FIELDS FOR SENTENCE BOUNDARY DETECTION IN SPEECH

Yang Liu[1,2], *Mary Harper*[2], *Elizabeth Shriberg*[1,3], *Andreas Stolcke*[1,3]

[1] International Computer Science Institute, Berkeley, CA, USA
[2] School of Electrical and Computer Engineering, Purdue University, Lafayette, IN, USA
[3] SRI International, Menlo Park, CA, USA
{yangl,ees,stolcke}@icsi.berkeley.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, hidden Markov model (HMM) and maximum entropy (Maxent) classifier approaches have been used for detecting sentence boundaries using both textual and prosodic information. A conditional random field (CRF) combines advantages of these approaches, being both discriminative and able to perform sequence decoding. We show in this paper that a CRF yields a lower error rate than the HMM and Maxent models on the NIST sentence boundary detection task. Extensive comparisons across two corpora on both human transcriptions and recognition output confirm the strength of the CRF modeling approach when applying a variety of 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, and summarization, among others.

Most prior work on sentence segmentation [1, 2, 3, 4, 5] have used an HMM approach, in which the word/tag sequences are modeled by N-gram language models (LMs) [6]. Additional features (mostly related to speech prosody) are modeled as observation likelihoods attached to the N-gram states of the HMM [1]. A forward-backward algorithm is used to find the event with the highest posterior probability for each interword boundary:

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

where W and F are the words and features for the entire test sequence, respectively. The HMM is a generative modeling approach since it describes a stochastic process with hidden variables (sentence boundary) that produces the observable data. The HMM approach has two main drawbacks. First, the standard training methods for HMMs 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 word and POS labels) without greatly increasing the number of model parameters, which in turn would make robust estimation difficult.

A maximum entropy (Maxent) posterior classification method has been evaluated in an attempt to overcome these shortcomings of the HMM approach [7]. Maxent estimates the posterior probabilities directly, replacing the generative modeling approach of the HMM. Such a model takes the familiar exponential form:

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

where $Z_\lambda(W,F)$ is a normalization term. The indicator functions $g_k(E_i,W_i,F_i)$ correspond to features defined over events, words, and prosody. The parameters in Maxent are chosen to maximize the conditional likelihood $\prod_i P(W_i|W_i,F_i)$ over the training data, better matching the classification accuracy metric. The Maxent framework provides a more principled way to combine a large number of overlapping features, as confirmed by the results of [7]; however, it uses only local information to make the decision for each boundary.

A simple combination of the Maxent and HMM was found to improve upon the performance of either model alone [7] 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 [8] combines the benefits of these two approaches. Like Maxent, a CRF can accommodate many correlated features and can be trained in a discriminative way. Like an HMM, a CRF uses a sequence decoding that is globally optimal. Hence, we will compare the performance of the CRF model to both the HMM and Maxent approaches on the sentence boundary detection task.

Section 2 of this paper 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 X. CRFs have been successfully used for a

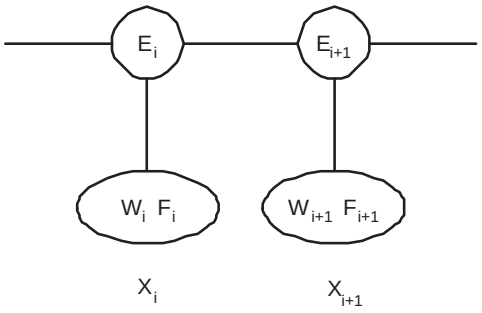

Fig. 1. A graphical representation of CRF for the sentence boundary detection problem. E represents the state tags (i.e., sentence boundary or not), while W and F are word and prosodic features respectively.

variety of text processing tasks [8, 9, 10], but this is the first time it is applied to a speech related task with both acoustic and textual knowledge sources. Figure 1 is a graphical representation of this modeling approach. The states of the model correspond to event labels E_i. The observations X_i associated with the states are the words W_i, as well as other (mainly prosodic) features F_i. The most likely sequence $\hat{E}$ for the given input sequence (observations) X is:

$$\hat{E} = \arg\max_{E} \frac{exp(\lambda * G(E,X))}{Z_\lambda(X)} \quad (3)$$

where the function G is a potential function over the events and the observations, and Z_λ is the normalization term

$$Z_\lambda = \sum_{E} exp(\lambda * F(E,X)) \quad (4)$$

The model is trained to maximize the conditional log-likelihood of a given training set. 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.[1]

A CRF differs from an HMM with respect to its training objective function (joint versus conditional likelihood) and its handling of dependent word features. HMM training does not maximize the posterior probabilities of the correct labels; whereas, the CRF directly estimates posterior boundary label probabilities $P(E|W,F)$. 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); 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 uses only local evidence (the surrounding word context and the local prosodic features), as shown in Equation (2).

We use the Mallet package [11] to implement the CRF model. To avoid overfitting, we employ a Gaussian prior with a zero mean on the parameters [12], similar to what is used for training Maxent models. The CRF takes longer to train than the HMM and Maxent models, especially when the number of features becomes large; whereas, the HMM requires less time for training than all the models.

[1]The forward-backward algorithm would likely be better here, but it is not implemented in the current software used [11].

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 [13]. SU boundaries as well as other structural metadata events were annotated by LDC according to an annotation guideline [14]. Both the transcription and the recorded speech were used by the annotators when labeling the transcriptions. We chose to evaluate on the NIST SU detection task because of the availability of annotated data and high quality scoring tools.

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. These differences are 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 DARPA Rich Transcription Fall 2003 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 from the Switchboard corpus for training and 6 hours (72 conversations) for testing. The BN data has about 20 hours of Broadcast News shows in the training set and 3 hours (6 shows) in the test set. The annotated training data is only a subset of the data used for training the speech recognizer because more effort is required to annotate the transcriptions.

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 study the SU detection task without the confounding effect of speech recognition errors. We use the speech recognition output obtained from the SRI recognizer [15].

For testing, a system determines the locations of sentence boundaries given the word sequence W and the speech. System performance is evaluated using the official NIST evaluation tools.[2] System output is scored by first finding a minimum edit distance alignment between the hypothesized word string and the reference, 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 either the number of true SU boundaries or the number of total word boundaries. The former is the **NIST SU error metric** while the latter is the **per-boundary-based metric**.

3.2. Feature Extraction

To obtain a good quality estimation of the conditional probability of the event tag given the observations $p(E_i|X_i)$, the observations should be based on features that are discriminative of the two events (SU versus not). As in [7], we utilize both textual and prosodic information for SU detection. Words and sentence

[2]See http://www.nist.gov/speech/tests/rt/rt2003/fall/ for more details.

boundaries are mutually constrained via syntactic structure. Therefore, the word identities themselves (from automatic recognition or human transcripts) constitute a primary knowledge source for the sentence segmentation task. Word N-grams are used as features in the CRF model. We also make use of various automatic taggers that map the word sequence to other representations. The tagged versions of the word stream are provided to support generalizations based on syntactic structure and to smooth out possibly undertrained word-based probability estimates. These tags include the part-of-speech tags, syntactic chunk tags, and automatically induced word classes. In addition, we use an extra text corpus, which is not annotated according to the guideline used for labeling the training and test data by LDC [14]. The hidden-event n-gram LM trained from the extra corpus is used to estimate posterior event probabilities for the LDC-annotated training and test data, and these posteriors are then thresholded to yield binary features [7].

In addition to textual features, we extract prosodic features that capture duration, pitch, and energy patterns associated with the word boundaries [1]. A decision tree classifier is used to model the prosodic features. We then encode the decision tree posteriors in a cumulative fashion through a series of binary features. We include speaker change as a feature in the observation X_i [7].

The features used for the CRF are the same as we used for a Maxent model devised for the SU detection task [7]. The same knowledge sources are used in the HMM approach, but with different representations. Keeping the knowledge sources consistent across the models enables us to focus on comparing modeling approaches. We will compare the three models (CRF, HMM, Maxent) to one another, as well as to a voting-based combination.

4. EXPERIMENTAL RESULTS AND DISCUSSION

SU detection results using the CRF, HMM, and Maxent approaches individually, using the reference transcriptions or speech recognition output, are shown in Tables 1 and 2. We present results when different knowledge sources are used: word N-gram only, word N-gram and prosodic information, and using all the features listed in Section 3. The detection error rate is reported using both 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 1 that the CRF is superior to both the HMM and Maxent 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 between 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 between 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 a performance similar to or better than other methods using all the knowledges sources. This may be useful when feature extraction is computationally expensive.

[CTS]		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)		

Table 1. CTS 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. '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.

We observe from Table 1 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 more 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, Maxent is not superior to the HMM; only when various additional textual features are included in the feature set does Maxent show its strength compared to the HMM, highlighting the benefit of Maxent's handling of many correlated features.

The combined result (using majority vote) of the three approaches in Table 1 is superior to any model alone. Previously, we found that the Maxent and HMM posteriors combine well because the two approaches have different error patterns [7]. The toolkit we use for the implementation of the CRF does not have the functionality of generating a posterior probability for a sequence; therefore, we do not combine the system output via posterior probability interpolation, which we would expect to yield a better performance.

4.2. BN Results

Table 2 shows the SU detection results for BN. Similar to the patterns as 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

[BN]		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)		

Table 2. BN 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.

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 conditions.

Comparing Table 1 and Table 2, 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 signals 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, the baseline error rate (i.e., the priors of the SUs) is lower on BN than CTS.

5. CONCLUSIONS AND FUTURE WORK

We have shown that a discriminatively trained CRF model is a competitive approach for the sentence boundary detection task in speech using textual and prosodic information. The CRF combines advantages of the generative HMM approach and the conditional Maxent approach, being discriminatively trained and able to model the entire sequence. It outperforms the HMM and Maxent approaches consistently across various testing conditions. We also find that as more features are used, the differences among the modeling approaches decrease.

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.

6. ACKNOWLEDGMENTS

The authors gratefully thank Andrew McCallum at the University of Massachusetts and Fernando Pereira at the University of Pennsylvania for their assistance with their CRF toolkit. This research has been supported by DARPA under contract MDA972-02-C-0038, ARDA under MDA904-03-C-1788, NSF-STIMULATE under IRI-9619921, NSF O229012, and NASA under NCC 2-1256. Distribution is unlimited. Any opinions expressed in this paper are those of the authors and do not necessarily reflect the views of ARDA, DARPA, NADA, or NSF. Part of this work was carried out while the second author was on leave from Purdue University and at NSF.

7. REFERENCES

[1] E. Shriberg, A. Stolcke, D. H. Tur, and G. Tur, "Prosody-based automatic segmentation of speech into sentences and topics," *Speech Communication*, vol. 32, no. 1-2, pp. 127–154, 2000.

[2] Y. Gotoh and S. Renals, "Sentence boundary detection in broadcast speech transcripts," in *ISCA Workshop: Automatic Speech Recognition: Challenges for the New Millennium ASR-2000*, 2000, pp. 228–235.

[3] H. Christensen, "Punctuation annotation using statistical prosody models," in *ISCA Workshop on Prosody in Speech Recognition and Understanding*, 2001.

[4] J. Kim and P. C. Woodland, "The use of prosody in a combined system for punctuation generation and speech recognition," in *Proc. of Eurospeech 2001*, 2001, pp. 2757–2760.

[5] National Institute of Standards and Technology, "RT-03F workshop agenda and presentations," http://www.nist.gov/speech/tests/rt/rt2003/fall/ presentations/, Nov. 2003.

[6] A. Stolcke and E. Shriberg, "Automatic linguistic segmentation of conversational speech," in *Proc. of ICSLP 1996*, 1996, pp. 1005–1008.

[7] Y. Liu, A. Stolcke, M. Harper, and E. Shriberg, "Comparing and combining generative and posterior probability models: Some advances in sentence boundary detection in speech," in *Proc. of EMNLP 2004*, 2004.

[8] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random field: Probabilistic models for segmenting and labeling sequence data," in *Prof. of ICML 2001*, 2001, pp. 282–289.

[9] F. Sha and F. Pereira, "Shallow parsing with conditional random fields," in *Proc. HLT/NAACL*, 2003.

[10] A. McCallum and W. Li, "Early results for named entity recognition with conditional random fields," in *CoNLL*, 2003.

[11] A. McCallum, "Mallet: A machine learning for language toolkit," http://mallet.cs.umass.edu, 2002.

[12] S. Chen and R. Rosenfeld, "A Gaussian prior for smoothing maximum entropy models," Tech. Rep., Carnegie Mellon University, 1999.

[13] DARPA Information Processing Technology Office, "Effective, affordable, reusable speech-to-text (EARS)," http://www.darpa.mil/ipto/programs/ears/, 2003.

[14] S. Strassel, *Simple Metadata Annotation Specification V5.0*, Linguistic Data Consortium, 2003.

[15] A. Stolcke et al., "Speech-to-text research at SRI-ICSI-UW," 2003, http://www.nist.gov/speech/tests/rt/rt2003/spring/presentations/index.htm.