The ICSI+ Multilingual Sentence Segmentation System

M. Zimmerman [1], *D. Hakkani-Tür* [1], *J. Fung* [1], *N. Mirghafori* [1], *L. Gottlieb* [1], *E. Shriberg* [1,2], *Y. Liu* [3]

[1]International Computer Science Institute, [2]SRI International, [3]University of Texas, Dallas

Abstract

The ICSI+ multilingual sentence segmentation with results for English and Mandarin broadcast news automatic speech recognizer transcriptions represents a joint effort involving ICSI, SRI, and UT Dallas. Our approach is based on using hidden event language models for exploiting lexical information, and maximum entropy and boosting classifiers for exploiting lexical, as well as prosodic, speaker change and syntactic information. We demonstrate that the proposed methodology including pitch- and energy-related prosodic features performs significantly better than a baseline system that uses words and simple pause features only. Furthermore, the obtained improvements are consistent across both languages, and no language-specific adaptation of the methodology is necessary. The best results were achieved by combining hidden event language models with a boosting-based classifier that to our knowledge has not previously been applied for this task.

Index Terms: maximum entropy, boosting, hidden event language models, prosody

1. Introduction

In the context of the DARPA GALE program broadcast news (BN), broadcast conversations, newswire, and so on in languages other than English (i.e. Arabic and Mandarin) are to be translated by machine into English, summarized, and transformed into a format suitable for a number of different information retrieval techniques. To break the task into manageable units, a network of modules is implemented beginning with automatic speech recognition (ASR) in the corresponding language. Later, machine translation and further natural language processing techniques are applied. For many of these steps the ASR output needs to be enriched with information additional to words, such as speaker diarization, sentence segmentation, or story segmentation.

The role of sentence segmentation is to detect sentence boundaries in the stream of words provided by the ASR module for further downstream processing. This is helpful for various language processing tasks, such as parsing, machine translation and question answering. We formulate sentence segmentation as a binary classification task. For each position between two consecutive words the system must decide if the position marks a boundary between two sentences or if the two neighboring words belong to the same sentence.

This work concentrates on our first attempt to improve the sentence segmentation for Mandarin and English over the baseline method that includes hidden event language models (HELMs) and decision trees, where the HELM takes into account the sequence of words and the output of the decision tree that is based on pause durations. The new approach combines the HELMs for exploiting lexical information, with maximum entropy and boosting classifiers that tightly integrate lexical, as well as prosodic, speaker change and syntactic features. The boosting-based classifier (using words and all prosodic features as input) alone performs better than all the other classification schemes. When combined with a hidden event language model the improvement is even more pronounced. Furthermore, these results are consistent across both English and Mandarin data.

2. Related Work

Sentence boundary detection (and similarly adding punctuation mark) in speech has been studied in an attempt to enrich speech recognition output [1, 2, 3, 4]. In the previous approaches for this task, different classifiers have been evaluated (e.g. hidden Markov model (HMM), maximum entropy), utilizing both textual and prosodic information. In the DARPA EARS program, special efforts were made for rich transcription of speech with automatically generated structural information, including sentence boundaries, disfluencies, and filler words. For example, [4] evaluated different modeling approaches (HMM, maximum entropy, conditional random fields) and various prosodic and textual features, in both conversational telephone speech and broadcast news speech. A reranking technique [5] further improved sentence boundary detection performance upon the baseline of [4]. Sentence segmentation has also been investigated in the multiparty meeting corpus [6, 7], with observations similar to those in conversational telephone speech.

There is also related work for sentence boundary detection in other languages, for example, in Czech [8] where an HMM approach was used, and in Chinese [9, 10] where a maximum entropy classifier was used with mostly textual features.

3. Approach

For sentence segmentation, different information sources are taken into account. The most important information sources are the sequence of words from the ASR module and the duration of the pause between neighboring words. In addition, a number of prosodic features derived from measurements of duration, pitch, and energy are extracted at each inter-word boundary, and the output of a speaker diarization is considered as well. We first detail the extraction of the prosodic features, and then describe the classification techniques involved and explain the experimental setup.

3.1. Prosodic Features

Our prosodic features were calculated using Algemy, a Java-based program developed at SRI [11]. Algemy contains a graphical user interface that allows users to easily read and program scripts for the calculation of prosodic features using modular algorithms as building blocks. These blocks are strung together in directed acyclic graphs (DAGs) to extract the desired feature (see Fig. 1 for an example). In batch mode, Algemy produces features in computation time comparable to traditional scripts, but has the advantage of

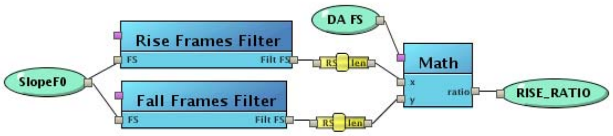

Figure 1: An example of the computation of the RISE RATIO feature from the F0 contours in the graphical user interface of Algemy.

being much easier to comprehend, change, and apply to other corpora. Another advantage is that by creating a unified set of basic, reusable, and robust modules, the overall development time for prosodic feature design can be decreased significantly.

So far we have used Algemy to implement a number of word-based features related to pitch, pitch slope, energy, and pause duration between words. The pitch and pitch slope features are based on piecewise-linear segments fit to extracted F0 values. These, and the energy features, are composed of features comparing the high, mean, and low values as well as slope patterns across word boundaries over both 20ms and word-length windows.

Currently, to process a BN show in Algemy, the only files that are needed are pitch values calculated from the waveforms using the ESPS `get_f0` program or similar, and word alignments in a modified RTTM file format (see under Evaluation Plan [12]). This is language independent, and Algemy DAGs have been made to work with different corpora in different languages quickly with minimal modifications.

3.2. Classification Approach

Hidden-event language models (HELMs) for segmentation were introduced in [13]. They can be considered a variant of the widely used statistical n-gram language models [14]. The difference arises from the fact that during the training of the hidden-event language models the events to detect (sentence boundary tokens <s> in our case) are explicitly present, while they are missing (or hidden) during the recognition phase. For the experiments in this paper we used word based 4-gram language models with interpolated Kneser-Ney smoothing [15, 16].

Decision trees (DTs) based on the C4.5 algorithm [17] are used in combination with HELM for the baseline segmentation system. The decision trees are trained on the pause durations between two consecutive words that either correspond to a sentence boundary, or occur between two words of the same sentence. This is in contrast to other work, for example [1, 4], where decision trees are trained on a large set of different prosodic features. The motivation for a pause-only system lies in the simplicity and low computational overhead of such an approach, as all the necessary information can be extracted from the ASR output alone.

Maximum entropy (MaxEnt) models have been successfully used in a wide variety of applications as they offer discriminative training and can easily handle thousands of features and the model training procedure is proved to be able to converge to the uniquely defined global optimum. See [18] for an excellent introduction. In its standard form, all features in a maximum entropy model are of a binary form indicating either the presence or absence of a feature. In our experiments two different feature sets are used. For the first model, MaxEnt(1), both words and pause durations are considered where the pause durations are binned into 10 classes. As features we use word and pause unigrams, word and pause bigrams, and bigrams of word and pause combinations.[1] The second model, MaxEnt(2), also takes into account the speaker turns that are estimated by the diarization system. In addition to the MaxEnt(1) model speaker turn unigrams, trigram, as well as bigrams of turn and word combinations are utilized.

The boosting-based method we applied is derived from a text categorization task. 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* algorithm described in [19]. In contrast to the implementation of the maximum entropy model, BoosTexter handles both discrete and continuous features, which allows for a convenient incorporation of the prosodic features described above (no binning is needed). For comparison, two boosting-based classifiers were implemented. The first classifier, BoosTexter(1), relies on words and the pause feature alone (making it comparable to MaxEnt(1) and the combination of the HELM with the pause-based decision trees). BoosTexter(2), uses the features from BoosTexter(1) plus all the pitch- and energy-related prosodic features.

For the combination of the hidden-event language model with decision trees and the sentence boundary detection approaches mentioned above, the integrated HMM scheme described in [1] is used. The original task of finding the optimal sequence T^* for a given word sequence W is extended to take into account additional information X related to the input word sequence.

$$T^* = \underset{T}{argmax}\, p(T|W, X) \qquad (1)$$

In contrast to an HMM-based HELM the states of the integrated model do emit not only words, but information of additional knowledge sources in the form of likelihoods $p(X_i|T_i, W)$ where X_i represents the additional information emitted at the position of word W_i and $T_i \in \{\text{<s>}, \emptyset\}$ depends on the presence of a sentence boundary token <s>. In [1] the required likelihoods are obtained from the outputs of decision trees computed from the prosodic features extracted around word boundaries. In our case the required likelihoods are derived from the posterior probabilities estimated by either decision trees, maximum entropy models, or the BoosTexter algorithm.

3.3. Experimental Setup

For testing our approaches, we have used English and Mandarin TDT-4 corpora. The properties of the training, development, and test sets are summarized in Table 1. Note that the data represented in Table 1 represents the subset of the TDT-4 corpora for which we had all the necessary information sources available.

As a primary input into the sentence segmentation system, the 1-best word sequence from the ASR is used, including pause durations between words as well as the phone durations for the words. In addition, speaker turn changes are estimated by a diarization system. Two independent speech recognition systems developed by two different sites are used. The English ASR system is described in [20] while a description of the Mandarin system can

[1] The features are computed using a 5-word window context (for the current, preceding two, and following two words).

Language	Training	Dev.	Test	Length
English	1,313k (267)	97k (20)	97k (20)	14.6
Mandarin	530k (131)	80k (17)	85k (17)	29.3

Table 1: Number of words (shows) for the subsets of the English and Mandarin TDT-4 corpora used in the experiments for training, development (Dev.) and tests. The last column refers to the average number of words per sentence.

be found in [21]. Recognition scores for the TDT-4 corpora used in our experiments 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 for the English TDT-4 corpus lies between 17 and 19%. In the case of the Mandarin TDT-4 word error rates between 20 and 25% were estimated. As the estimation procedure included a sub-optimal gold standard (close captions), the numbers given above most likely under-estimate the performance of the ASR systems. For the definition of the sentence segmentation gold standard, the transcriptions (i.e., the close captions of the shows) were aligned with the corresponding ASR output word sequence. Depending on the alignment, the sentence boundaries derived from the available punctuation (periods and question marks) in the transcriptions were then inserted into the ASR output.

To extract the speaker turn features the speaker segmentation system described in [22] is used for both English and Mandarin. It is based on an agglomerative clustering technique where an initial number of clusters greater than the optimum amount of speakers is iteratively merged. The stopping criterion is based on a modified Bayesian Information Criterion (BIC) metric to compare all cluster pairs where the clustering is terminated when there are no more pairs that are similar enough according to the BIC metric.

For performance evaluation, we report NIST error rate and F-measure on automatic speech recognizer output. The NIST error rate corresponds to the average number of misclassified word boundaries per reference sentence boundaries. The F-measure is the harmonic mean of the computed precision and recall given the reference sentence boundaries and the boundaries hypothesized by the segmentation system.

4. Results and Discussion

Table 2 shows the NIST error and F-measure results on the English and Mandarin test sets using various methods and features. For segmenting the test set into sentences, we use the parameters (model combination weights and probability thresholds for selecting sentence boundaries) optimized on the development set.

We obtain substantial improvements from the use of the simple pause-based features over the HELM that only includes (word based) lexical information. This finding is consistent with previous work and holds for the HELM+DT, the MaxEnt(1), and the BoosTexter(1) system. However, both MaxEnt(1) and BoosTexter(1) classifiers perform significantly better alone than the HELM+DT system, and can be even further improved when combined with HELMs. The difference between the MaxEnt(1) and the BoosTexter(1) results can potentially be attributed to the binning process of the pause durations as the BoosTexter directly works on the continuous pause durations while the MaxEnt model relies on discretized pause durations. The improvement from MaxEnt(1) to MaxEnt(2) that also includes the speaker turn features is more pronounced for the English case. This can be attributed to the fact that in our English data roughly twice as many speaker changes are detected by the speaker diarization system compared to our Mandarin data. The improvements from the use of the additional prosodic features in BoosTexter(2) over BoosTexter(1) seem consistent for both English or Mandarin. It is also interesting to see that the performance gain of BoosTexter(2) over BoosTexter(1) still holds when these models are combined with the HELM.[2]

We presented a multilingual sentence segmentation system. In contrast to previous work we provided results for the same methodology using both English and Mandarin broadcast news data and demonstrate that the proposed methodology performs very competitively and ports robustly across the two languages. Furthermore, our best system applied a boosting-based classifier originally developed for text categorization that has previously not been used for sentence segmentation. In our experiments, we have achieved a large gain in performance by using pause duration in between words as a feature. Additional gains were found for pitch- and energy-related prosodic features, as well as features derived from speaker turns obtained from a speaker diarization system.

In future work we will try to optimize segmentation system parameters according to different downstream processing tasks. For example, preliminary experiments indicate that in the case of a machine translation system it is beneficial to slightly oversegment the texts compared to the segmentation ground truth obtained from punctuation. For prosodic feature extraction it will be important to add the currently missing speaker-specific normalization of the pitch features. Currently, the probability of pitch halving and doubling is calculated over an entire show. It is desirable to utilize speaker-diarization segment information to calculate such features over speaker-specific regions. This will allow us to more accurately fit piecewise-linear segments as well as more accurately normalize the existing pitch features. We also would like to investigate syntactically motivated features and further classification schemes, such as support vector machines and conditional random fields, as well classification combination approaches.

Acknowledgments: We thank Mei-Yuh Hwang, Wen Wang, and Dimitra Vergyri for helping us by generating the ASR output and Harry Bratt, Luciana Ferrer, and Martin Garciarena for their help with Algemy. The development of Algemy was separately funded by SRI. This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-C-0023. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of DARPA.

5. References

[1] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tür, “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 *Proc. of ISCA Workshop: Automatic Speech Recognition: Challenges for the new Millennium ASR-2000*, 2000, pp. 228–235.

[2] For the results reported in Table 2 data a difference of 1% absolute is mostly highly significant (> 99%) whereas a difference of 0.5% absolute is significant around the 95% percentage level using a simple sign test.

					English		Mandarin	
Method	Words	Pause	F0+Energy	Speaker	NIST Error	F-Measure	NIST Error	F-Measure
HELM	√				85.6%	51.1%	80.0%	55.8%
MaxEnt(1)	√	√			71.9%	62.0%	70.1%	65.7%
MaxEnt(2)	√	√		√	70.6%	62.7%	65.4%	67.3%
BoosTexter(1)	√	√			70.1%	63.6%	66.3%	69.3%
BoosTexter(2)	√	√	√		68.8%	65.5%	64.9%	69.8%
HELM + DT	√	√			74.1%	61.7%	71.2%	65.4%
HELM + MaxEnt(1)	√	√			69.8%	63.4%	66.2%	67.3%
HELM + MaxEnt(2)	√	√		√	69.0%	63.9%	64.7%	68.0%
HELM + BoosTexter(1)	√	√			67.0%	64.9%	60.7%	70.2%
HELM + BoosTexter(2)	√	√	√		62.4%	67.3%	58.7%	70.8%

Table 2: Investigated configurations of classifiers and features. Classifiers are Hidden-event LMs (HELM), decision trees (DT), maximum entropy (MaxEnt), and boosting (BoosTexter). The feature groups include language model and pause (LM+Pause), Pitch and Energy (F0+Energy), and speaker turns estimated by a diarization module (Speaker).

[3] J. Huang and G. Zweig, "Maximum entropy model for punctuation annotation from speech," in *Proc. of ICSLP*, 2002, pp. 917–920.

[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 *Proc. of ICASSP*, 2005, vol. 5, pp. 957–960.

[5] 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 *Proc. of ICASSP*, 2006, vol. 1, pp. 545–548.

[6] J. Ang, Y. Liu, and E. Shriberg, "Automatic dialog act segmentation and classification in multiparty meetings," in *Proc. of ICASSP*, 2005, vol. 1, pp. 1061–1064.

[7] M. Zimmermann, Y. Liu, E. Shriberg, and A. Stolcke, "A* based segmentation and classification of dialog acts in multiparty meetings," in *Proc. of ASRU*, 2005, pp. 215–219.

[8] J. Kolar, J. Svec, and J. Psutka, "Automatic punctuation annotation in Czech broadcast news speech," in *Proc. of 9th Conference Speech and Computer*, 2004, pp. 319–325.

[9] C. Zong and F. Ren, "Chinese utterance segmentation in spoken language translation," in *The 4th International Conference on Computational Linguistics and Intelligent Text Processing*, 2003, pp. 516–525.

[10] D. Liu and C. Zong, "Utterance segmentation using combined approach based on bi-directional n-gram and maximum entropy," in *Proc. of ACL-2003 Workshop: The Second SIGHAN Workshop on Chinese Language Processing*, 2003, pp. Pages 16–23.

[11] H. Bratt, "Algemy, a tool for prosodic feature analysis and extraction," *Personal Communication*, 2006.

[12] NIST Website, "Rich transcription 2004 spring meeting recognition evaluation," http://www.nist.gov/speech/tests/rt/rt2004/spring/.

[13] A. Stolcke and E. Shriberg, "Automatic linguistic segmentation of conversational speech," in *Proc. of ICSLP*, Philadelphia, USA, 1996, vol. 2, pp. 1005–1008.

[14] F. Jelinek, "Self-organized language modeling for speech recognition," in *Readings in Speech Recognition*, A. Waibel and K.-F. Lee, Eds., pp. 450–506. Morgan Kaufmann Publishers, Inc., 1990.

[15] J. T. Goodman, "A bit of progress in language modeling," Msr-tr-2001-72, Machine Learning and Applied Statistics Group, Microsoft, Redmond, USA, 2001.

[16] R. Kneser and H. Ney, "Improved backing-off for m-gram language models," in *Proc. of ICASSP*, Massachusetts, USA, 1995, vol. 1, pp. 181–184.

[17] R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Diego, 1993.

[18] A. L. Berger, S. A. Della Pietra, and V. J. Della Pietra, "A maximum entropy approach to natural language processing," *Computational Linguistics*, vol. 22, no. 1, pp. 39–71, 1996.

[19] R. E. Schapire and Y. Singer, "BoosTexter: A boosting-based system for text categorization," *Machine Learning*, vol. 39, no. 2-3, pp. 135–168, 2000.

[20] A. Venkataraman, A. Stolcke, W. Wang, D. Vergyri, V. Ramana, R. Gadde, and J. Zheng, "SRI's 2004 broadcast news speech to text system," in *EARS Rich Transcription 2004 workshop, Palisades*, Nov 2004.

[21] M.-Y. Hwang, X. Lei, W. Wang, and T. Shinozaki, "Investigation on Mandarin broadcast news speech recognition," in *Proc. of ICSLP*, 2006.

[22] C. Wooters, J. Fung, B. Peskin, and X. Anguera, "Towards robust speaker segmentation: The ICSI-SRI fall 2004 diarization system," in *RT-04F Workshop*, 2004.