AUTOMATIC SENTENCE BOUNDARY DETECTION IN CONVERSATIONAL SPEECH: A CROSS-LINGUAL EVALUATION ON ENGLISH AND CZECH

Jáchym Kolář

Faculty of Applied Sciences, Dept. of Cybernetics,
Univ. of West Bohemia in Pilsen, Czech Republic

Yang Liu

Department of Computer Science,
Univ. of Texas at Dallas, TX, USA

ABSTRACT

Automatic sentence segmentation of speech is important for enriching speech recognition output and aiding downstream language processing. This paper focuses on automatic sentence segmentation of speech in two different languages – English and Czech. For this task, we compare and combine three statistical models – HMM, maximum entropy, and a boosting-based model BoosTexter. All these approaches rely on both textual and prosodic information. We evaluate these methods on a corpus of multiparty meetings in English, and on a corpus of broadcast conversations in Czech, using both manual and speech recognition transcripts. The experiments show that superior results are achieved when all the three models are combined via posterior probability interpolation. We observe differences in terms of model performance between English and Czech, as well as the feature usage difference in prosodic models between the two languages. Overall, the analysis is important for porting sentence segmentation approaches from one language to another.

Index Terms— spoken language understanding, sentence boundary detection, prosody, machine learning

1. INTRODUCTION

Automatic sentence segmentation of speech is important to make speech recognition (ASR) output more readable and easier for downstream language processing modules. Various techniques have been studied for automatic sentence boundary detection in speech, including hidden Markov models (HMMs), maximum entropy, neural networks, and Gaussian mixture models, utilizing both textual and prosodic information [1, 2, 3, 4].

In this paper, we focus on the evaluation of different approaches for sentence segmentation in two different languages: English and Czech. Czech is different from English in many aspects that make it generally more difficult for this task. Czech belongs to the family of Slavic languages, which are highly inflectional and derivational, and thus have an extremely large number of distinct word forms. In addition, colloquial Czech has a different morphology than standard Czech – prefixes and endings are often changed in the former. As a result, the number of word forms is even higher in conversational Czech, where standard and colloquial forms are usually mixed. Another problem is relatively free word order in Czech. Furthermore, there are also differences in prosody, for example, less emphatic pre-boundary lengthening and less steep pitch movements in Czech than in English. The first published work about sentence segmentation of spoken Czech [5] described an HMM-based system evaluated on broadcast news speech. Later, Kolorenč proposed a system based on rules automatically induced by genetic algorithms [6]. The latest work on sentence segmentation of Czech focused on studying genre effects on this task [7].

Unlike previous work, the goal of this study is a cross-lingual comparison of English and Czech in the task of sentence segmentation of conversational speech, as well as evaluation of different statistical models, namely, HMM, maximum entropy, and boosting. All of these methods rely on both textual and prosodic information. We not only examine performance of individual models, but also consider their combination via posterior probability interpolation. For the evaluation, we use a corpus of English multiparty meetings and a corpus of Czech broadcast conversations. We conduct experiments using both human transcripts and speech recognition output. Our analysis shows cross-lingual differences in both model performance, and textual and prosodic feature usage.

2. METHOD

For a given word sequence, our task is to determine the location of sentence-like unit (SU) boundaries using textual information (recognized words) and acoustic information (prosody). This is represented as a classification or tagging problem, and we use statistical models to incorporate different information sources.

2.1. Knowledge sources

Because of the space limit, we briefly describe the textual and prosodic features used in this work in the following two subsections. In general, the information sources we use for Czech and English are similar. See [7] for a more detailed description of the features, along with an analysis of their contribution to sentence segmentation of Czech, and [8] for more information about English.

2.1.1. *Textual features*

To alleviate the problem of data sparsity, we not only use information about word identities but also employ automatically induced classes (AIC) and part-of-speech (POS) tags. AICs were induced using a well-known clustering algorithm minimizing perplexity of the induced class n-gram model based on the word bigram counts. The SU boundary token was excluded from merging, however, its statistics still affected the clustering.

The POS tags for the English corpus were obtained using the TnT tagger trained on conversational data. Unlike English, Czech POS tags are positional. Each tag is represented as a string of 15 subtags that approximately fit the categories of the formal Czech morphology. The total number of possible tags is high – over 1,500.

This work was supported by the Ministry of Education of the Czech Republic under project No. 1M0567, and by the NSF grant IIS-0845484. The views expressed are those of the authors, and not the funding agencies.

The tags for our data were generated by a state-of-the-art Czech tagger [9]. Based on our previous experiments, we chose not to use the POS tags directly, but rather in a combination with frequent words. This can be viewed as a form of back off – we back off from words to tags for rare words but keep word identities for frequent words.

2.1.2. Prosodic features

Our prosodic features are designed to reflect breaks in pause, temporal, intonational, or energy contours. The features are extracted from speech signal using word-level and phone-level time alignment information from an automatic speech recognizer. The features are associated with interword boundaries. In order to capture local prosodic dynamics, we also use features associated with the previous and the following word boundaries. In addition, we added information capturing phenomena such as speaker changes. Finally, we performed feature selection to identify a small set of useful prosodic features for both corpora.

2.2. Models

We examine three statistical approaches to sentence segmentation – HMM, maximum entropy (MaxEnt), and a model based on adaptive boosting called BoosTexter. All three approaches rely on both textual and prosodic information, but combine the two knowledge sources in different fashions. The HMM approach uses an independent language model and prosody model that are combined at the score level during testing. In MaxEnt, the learning algorithm combines textual features with thresholded prosodic posteriors, obtained from an independent prosodic classifier. The BoosTexter approach builds one integral model that combines the two information sources at the feature level. It would be interesting to compare these models because of their different views on the knowledge source combination. In addition, these different models are likely to be complementary, and thus their combination may yield better performance than individual models.

2.2.1. HMM

The HMM model [1] describes the joint distribution of word sequence W, prosodic features P, and SU boundaries S, $P(W, P, S)$. The model assumes that prosodic features depend only on the events (SU boundary or not), and not on the words. The transition probability is based on an n-gram language model (LM), which is trained by explicitly including the SU boundary as a token in the vocabulary. We used trigram LMs with modified Kneser-Ney smoothing. The observation likelihood comes from the prosodic model, for which we used decision tree classifiers. To overcome the problem of data skew (SU boundaries are much less frequent than non-boundaries) and to decrease classifier variance, we employ a combination of ensemble sampling with bagging [10]. During testing, we perform forward-backward decoding to find the boundaries (hidden states) given the word sequence and corresponding prosodic features (observations).

2.2.2. Maximum entropy

Unlike the generative HMM, Maximum Entropy (MaxEnt) is a discriminative model trained to directly discriminate among the target classes. Textual features for our MaxEnt model included n-grams of words, AIC, and POS tags. We used up to trigrams spanning across or neighboring with the inter-word boundary in question. To capture word repetitions, we also employed a binary feature indicating whether the word before the boundary is identical to the following word. Similar to HMM, prosodic information is used via the decision tree prosody model; however, unlike in HMM, the prosodic probabilities in the MaxEnt model were not used directly since this model usually does not perform well dealing with many real-valued features. Therefore we encoded the posteriors via thresholding to yield binary features. Because the presence of each feature in a MaxEnt model raises or lowers the final probability by a constant factor, it is reasonable to encode the posteriors in a cumulative fashion. This setup is more robust than using interval-based bins since small changes in prosodic scores may still result in matched features. We experimented with various gaps between adjacent thresholds and found 0.1 to be a convenient value. Thus, we obtained the following binary features: $p > 0.1, p > 0.2, \ldots, p > 0.9$. To avoid overfitting in MaxEnt, we used smoothing with Gaussian priors. Since MaxEnt does not have a separate LM, and assumes that all features are available during training, it does not allow to directly use additional data from text corpora that do not have any prosodic features associated with words. To overcome this problem, we used the additional LM in an HMM framework (without prosodic model) to estimate posterior SU probabilities for each boundary, and these posteriors were subsequently used as an extra feature during training and testing. Similar to the prosodic posterior probabilities, we thresholded the additional LM probabilities and used them as binary features. For all our experiments with MaxEnt, we employed the `MegaM` toolkit.[1]

2.2.3. BoosTexter

The third approach, BoosTexter, is based on boosting, which combines many weak learning algorithms to produce an accurate classifier. Each weak classifier is built based on the outputs of previous classifiers, focusing on the samples that were formerly classified incorrectly. The BoosTexter algorithm [11] was originally designed for the task of text categorization, and combines weak classifiers having a basic form of one-level decision trees (stumps) using confidence-rated predictions. In BoosTexter, we used the same textual features as in the MaxEnt model. The prosodic features were used directly in BoosTexter, unlike in HMM or MaxEnt that use the output from the decision tree prosody model. Each weak classifier checks for the presence or absence of an n-gram, or for a value of a continuous or categorical feature. The number of training iterations is optimized using a development set. BoosTexter allows a natural combination of textual and real-valued prosodic features at the feature level. Similar to MaxEnt, we also could not use additional text data directly, but used LM posteriors from the HMM model (in this case non-discretized). In our experiments, the ICSI implementation of the boosting algorithm was employed.[2]

3. EXPERIMENTS

3.1. Data and experimental setup

We evaluate our methods using two corpora – the ICSI meeting corpus [12] (English, EN) and the Czech (CZ) broadcast conversation corpus [13]. Both corpora are publicly available from LDC. All experiments were evaluated using both human-generated reference transcripts (REF) and automatic speech recognition (ASR) transcripts. The English corpus contains multichannel conversational speech recorded by headworn microphones. The data were split into a training set (51 meetings, 539k words), a development set (11 meetings, 110k words), and a test set (11 meetings, 102k words).

[1] `http://hal3.name/megam`

[2] `http://code.google.com/p/icsiboost/`

For this corpus, SU boundaries are defined as dialog act boundaries annotated based on a set of strict segmentation rules. Recognition results were obtained using the state-of-the-art SRI speech recognition system with word error rate of about 38.2%. For Czech, the data included 159.1k words for training, 24.1k words for development, and 24.6k words for testing. The corpus is annotated based on LDC's Metadata Extraction (MDE) standard [13]. The MDE annotation included labeling of SU boundaries, which were used in this work. The ASR output was obtained from the UWB LVCSR system tailored for real-time recognition of highly inflective languages [14]. The overall word error rate was 29.3%. For LM training, we also used an additional text corpus of Czech broadcast transcripts (107M words).

We must note here that there is not a perfect match in speaking styles between the English and the Czech corpus – although both contain multiparty conversational speech, meetings are more interactive and less formal than broadcast conversations. However, the Czech corpus we use is the only publicly available conversational speech database, and there is no available English broadcast conversation corpus annotated for SU boundaries.

We measure SU segmentation performance using F-measure, which is the harmonic mean of Precision (P) and Recall (R):

$$F = \frac{2PR}{P + R} \tag{1}$$

To generate the "reference" SU boundaries for the ASR words in both corpora, the reference setup was aligned to the recognition output with the constraint that two aligned words should occur within a fixed time threshold.

3.2. Results and discussion

The comparison of the three modeling techniques is visualized for all the evaluation test sets (EN REF, EN ASR, CZ REF, and CZ ASR) in Fig. 1 and 2, according to information sources used (textual vs. both textual and prosodic). The bars in both figures display F-measures for sentence boundary detection (thus higher is better). Fig. 1 shows results for models based only on textual information. HMM was the most successful approach for the EN corpus, but the difference among the three approaches is small. In contrast, BoosTexter was the best performing method for the CZ corpus, and the superiority of BoosTexter over the others (especially HMM) was greater than the differences in EN. This difference in model performance may be explained by lexical differences between English and Czech. Since Czech uses a much larger number of word forms than English, and it does not have a fixed word order, the LM performance depends highly on low order n-gram features which are more efficiently modeled in the discriminative models, MaxEnt and BoosTexter. On the other hand, higher order n-gram probabilities important for English are better modeled in the HMM approach which uses a more powerful smoothing technique. Regarding usefulness of individual textual knowledge sources, there are significant contrasts between English and Czech. First, while AIC information significantly improved results on English (relative error reduction 3.0%), it did not provide any gain for Czech. POS information was useful for both languages, but it was significantly more beneficial for Czech (relative error reduction 3.7%) than for English (1.3%).

The results of the models relying on both information sources are visualized in Fig. 2. For all the four test sets, the best results were achieved by the HMM model, however, the gaps between HMM and other models were statistically significant only for the English corpus ($p < .05$ using Sign test). There is some difference comparing the patterns shown in Fig 1 and 2. The winning method (HMM) for EN is the same for both conditions, whereas for CZ the best performing model in Fig. 2 is different from that in Fig 1 (only using textual information). It suggests that prosodic information is better modeled in the HMM approach. Also note that although the EN corpus is more spontaneous and thus may seem to be more difficult, we achieved higher F-scores for this than for the CZ corpus. This is possibly because of more difficult language modeling for Czech, which overbalances more difficult speaking style in the EN corpus. This hypothesis is also supported by the fact that the relative differences between EN and CZ are higher when we do not use prosodic information.

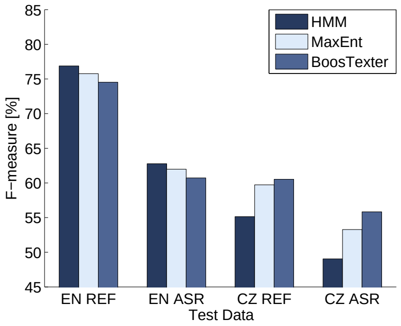

Fig. 1. Sentence boundary detection scores for individual models when only textual information is used.

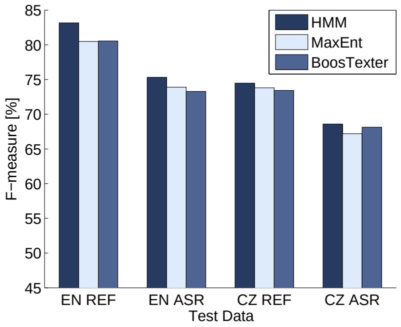

Fig. 2. Sentence boundary detection scores for individual models when both textual and prosodic information is used.

Table 1 summarizes the results for the three models using both textual and prosodic information. The table also presents results of a model that combines HMM, MaxEnt, and BoosTexter via posterior probability interpolation. The interpolation weights for the three models were estimated from development data using the EM algorithm. The results indicate that the combination improves SU segmentation accuracy in all the test conditions. The Sign test showed that the improvements over the best single models are significant at $p < .05$ for EN ASR, CZ REF, and CZ ASR. For EN REF, the difference is not significant.

We also analyzed differences between EN and CZ in terms of prosodic feature usage in decision trees. The usage metric reflects the number of times a feature is queried in a decision tree, weighted by the number of samples it affects at each node. The total feature usage within a tree sums to 1. The numbers are based on averaging over multiple trees generated in bagging. The feature usage distributions are displayed in Fig. 3. For both corpora, the most used groups

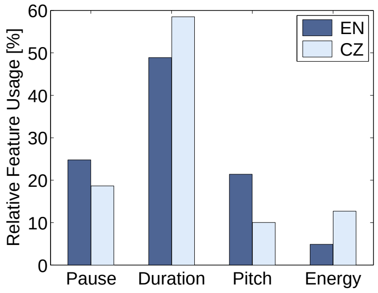

Fig. 3. Usage of different prosodic feature groups for English and Czech

are duration and pause, however, their proportions differ. Pause features have relatively higher usage in EN than CZ, and duration is used more in CZ than EN. This indicates that duration information is a relatively better cue in broadcast conversation, while in more spontaneous meetings, pause features are more reliable. Another difference between the two distributions is in the proportion of pitch and energy features – EN prefers pitch features, while energy features are used more in CZ. The low usage of energy in meetings may partly be explained by a higher variance in channel and inter-utterance loudness which affect extraction of the energy-based features (despite some feature post-processing we used).

4. SUMMARY AND CONCLUSIONS

In this paper, we have evaluated automatic sentence segmentation of conversational speech in two languages – English and Czech. We examined three different modeling approaches relying on both textual and prosodic cues: HMM, MaxEnt, and BoosTexter, and evaluated them using both reference and automatic transcripts. The results indicate that the HMM model showed most consistently good results. It produced best results in the majority of our tests. The only exceptions were the tests on Czech when using only textual information, where HMM was the worst approach. This indicates that the discriminative models, BoosTexter and MaxEnt, are more robust to lexical irregularities frequent in conversational Czech. Overall, the best results for all our test sets were achieved by a model that combines HMM, MaxEnt, and BoosTexter via posterior probability interpolation.

The results also indicate that Czech, as a rich morphology language, is more difficult than English for this task. Although the English meeting corpus is more spontaneous, contains more noise, and has a higher word recognition error rate, we achieved higher F-scores for this than for the Czech broadcast conversation corpus. The relative differences in F-scores between Czech and English were higher when we did not use prosodic information. Between English and Czech, our comparison has also shown some differences in feature usage in the prosodic models. Overall, this cross-lingual analysis is important to port sentence boundary detection approaches from one language to another.

Table 1. Overall SU boundary detection scores for HMM, MaxEnt, BoosTexter, and their combination via linear interpolation [F %]

	EN		**CZ**	
Model	**REF**	**ASR**	**REF**	**ASR**
HMM	83.17	75.33	74.48	68.59
MaxEnt	80.49	73.90	73.81	67.20
BoosTexter	80.55	73.29	73.43	68.14
Combination	**83.20**	**75.85**	**75.65**	**69.12**

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] J. Huang and G. Zweig, "Maximum entropy model for punctuation annotation from speech," in *Proc. of ICSLP 2002*, Denver, CO, USA, 2002.

[3] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, "Enriching speech recognition with automatic detection of sentence boundaries and disfluencies," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1526–1540, 2006.

[4] M. Tomalin and P. C. Woodland, "Discriminatevely trained Gaussian mixture models for sentence boundary detection," in *Proc. ICASSP*, Toulouse, France, 2006, pp. 549–552.

[5] J. Kolář, J. Švec, and J. Psutka, "Automatic punctuation annotation in Czech broadcast news speech," in *Proc. SPECOM*, St. Petersburg, Russia, 2004.

[6] J. Kolorenč, "Automatic punctuation of automatically recognized speech," in *Proc. ESSV 2005*, Prague, Czech Republic, 2005, pp. 291–297.

[7] J. Kolář, Y. Liu, and E. Shriberg, "Genre effects on automatic segmentation of speech: A comparison of broadcast news and broadcast conversations," in *Proc. ICASSP 2009*, Taipei, Taiwan, 2009.

[8] J. Kolář, E. Shriberg, and Y. Liu, "Using prosody for automatic sentence segmentation of multi-party meetings," in *Text, Speech and Dialogue*, ser. Lecture Notes in Artificial Intelligence, vol. 4188, 2006, pp. 629–636.

[9] D. Spoustová, J. Hajič, J. Votrubec, P. Krbec, and P. Květoň, "The best of two worlds: Cooperation of statistical and rule-based taggers for Czech," in *Proc. of the ACL Workshop on Balto-Slavonic NLP*, Prague, Czech Republic, 2007.

[10] Y. Liu, N. Chawla, M. Harper, E. Shriberg, and A. Stolcke, "A study in machine learning from imbalanced data for sentence boundary detection in speech," *Computer Speech and Language*, vol. 20, pp. 468–494, 2006.

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

[12] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The ICSI meeting corpus," in *Proc. ICASSP*, Hong Kong, 2003.

[13] J. Kolář and J. Švec, "The Czech broadcast conversation corpus," in *Text, Speech and Dialogue*, ser. Lecture Notes in Artificial Intelligence, vol. 5729, 2009, pp. 101–108.

[14] A. Pražák, L. Müller, J. V. Psutka, and J. Psutka, "Live TV subtitling: Fast 2-pass LVCSR system for online subtitling," in *Proc. SIGMAP 2007*, Barcelona, Spain, 2007.