Comparing and Combining Modeling Techniques for Sentence Segmentation of Spoken Czech Using Textual and Prosodic Information

Jáchym Kolář [1,2], *Yang Liu* [3]

[1]Dept. of Cybernetics, Faculty of Applied Sciences, Univ. of West Bohemia, Pilsen, Czech Republic
[2]Spoken Language Processing Group, LIMSI-CNRS, Orsay, France
[3]Department of Computer Science, University of Texas at Dallas, Richardson, TX, U.S.A.
jachym@limsi.fr, yangl@hlt.utdallas.edu

Abstract

This paper deals with automatic sentence boundary detection in spoken Czech using both textual and prosodic information. This task is important to make automatic speech recognition (ASR) output more readable and easier for downstream language processing modules. We compare and combine three statistical models – hidden Markov model, maximum entropy, and adaptive boosting. We evaluate these methods on two Czech corpora, broadcast news and broadcast conversations, using both manual and ASR transcripts. Our results show that superior results are achieved when all the three models are combined via posterior probability interpolation, and that there is substantial difference among the three methods when using different knowledge sources, as well as in different genres. Feature analysis also reveals significant differences in prosodic feature usage patterns between the two genres.

Index Terms: sentence segmentation, prosody, HMM, maximum entropy, boosting

1. Introduction

Automatic speech recognition (ASR) systems typically output only a raw stream of words, leaving out important structural information such as locations of sentence boundaries. However, sentence boundaries are crucial for many downstream language processing methods (e.g., parsing, information extraction, machine translation), which are typically trained on formatted text. Thus, our task is to determine the location of sentence unit boundaries for a given word sequence using textual information (recognized words) and acoustic information (prosody). A number of techniques have been proposed for this task, including multi-layer perceptrons, hidden Markov models (HMMs), and maximum entropy [1, 2, 3]. Subsequent work has studied the use of confusion networks, morphological and syntactic features, or speaker adaptation for this task [4, 5, 6].

In this paper, we focus on the Czech language, which is generally more difficult than English for this task. As other Slavic languages, Czech is highly inflectional and derivational, and thus has an extremely large number of distinct word forms. Furthermore, colloquial Czech has a different morphology than literary Czech – prefixes and endings are often changed. Another difficulty is that Czech does not have a fixed order of sentence constituents (subject, object, possessor, etc.), which evidently affects the predictive power of statistical models based on n-gram contexts. There are also differences in prosody between Czech and English. For example, Czech has a less emphatic preboundary lengthening and less steep intra-sentential pitch movements than English.

The first paper dealing with sentence segmentation of Czech [7] described an HMM-based system for automatic punctuation of broadcast news speech. More recent studies on Czech focused on analyzing genre effects [8] and differences between Czech and English [9]. Unlike previous work, this paper evaluates sentence segmentation of Czech in two different genres – Broadcast News (BN) and Broadcast Conversations (BC), using three different models – HMM, maximum entropy, and boosting. All methods are evaluated using both manual and speech recognition transcripts. We not only evaluate performance of individual models, but also consider their combination. The remainder of this paper is organized as follows. Section 2 describes our textual and prosodic features, Section 3 presents the three statistical models we use, Section 4 reports our experiments, and Section 5 provides conclusions.

2. Features

2.1. Textual Features

Data sparsity is a serious problem for morphologically rich languages as Czech. To mitigate this problem, we not only use information about word identities, but also utilize automatically induced classes (AIC) and part-of-speech (POS) tags. AICs were induced based on the word bigram counts to minimize perplexity of the induced class n-gram model [10]. We used 300 class for BN and 275 for BC.

Czech POS tags are positional, represented as strings of 15 subtags that correspond to the individual categories of Czech morphology. The total number of possible tags is high – over 1,500. The tags for our data were generated by a state-of-the-art Czech tagger [11]. As shown in our previous experiments [8], it is better not to use the POS tags directly, but rather in a combination with frequent words. This approach 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. Optimizing the model on development data, we ended up with keeping 1600 most frequent words for BN, and 2000 words for BC.

2.2. Prosodic Features

We developed a large database of prosodic features designed to reflect breaks in pause, temporal, intonational, or energy contours in speech. These features were extracted directly from speech signal using word-level and phone-level time alignment

This work was partially funded by the Ministry of Education of the Czech Republic under the project No. 2C06020, by OSEO under the Quaero program, and by the NSF grant IIS-0845484. The views expressed are those of the authors, and not the funding agencies.

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.

The group of pause features consisted of the pause duration after the current, the previous, and the following word. The duration features included the duration of vowels, final rhymes, and the whole word, aiming mainly to reflect the phenomenon of preboundary lengthening. We used raw durations as well as duration features normalized using phoneme duration statistics from the whole training set. The pitch features included features describing minimal, maximal and mean values, f_0 slopes, and differences and ratios of values across word boundaries. These features were extracted both from raw f_0 value and from an f_0 contour stylized by a piece-wise linear function. The energy features were represented by maximal, minimal, and mean frame-level RMS values, both raw and per-channel normalized. In addition to purely prosodic information, we added features capturing phenomena such as speaker changes.

We also performed feature selection to identify a small set of prosodic features in two steps. First, for each of the broad prosodic feature categories, we selected those features with a feature usage statistics higher than a predefined threshold. Then using these features, we performed leave-one-out feature selection and removed a feature if its deletion did not yield any performance loss. This feature reduction algorithm ended up selecting 11 features for BN and 17 features for BC. Note that for both test conditions (reference and ASR words), we trained our models on the prosodic database generated using a forced alignment of the reference transcripts. For ASR condition, we did not get any gain from training prosodic models using ASR output. In addition, generating ASR results for large databases is computationally expensive.

3. Models

We use three statistical models – HMM, maximum entropy, and a model based on adaptive boosting. All three approaches rely on both textual and prosodic information, but combine the two knowledge sources in different fashions.

3.1. HMM

The HMM model [2] has been widely studied in past work on sentence segmentation of speech. It describes the joint distribution of a word sequence W, prosodic features P, and sentence boundaries S, $P(W, P, S)$. The model assumes that prosodic features depend only on the target events (sentence boundary or not), and not on the words. Thus, we train an independent language model and prosody model, and combine them at the score level during testing. The observation likelihoods are estimated by the prosodic model, for which we use decision tree classifiers. To overcome the problem of data skew (sentence boundaries are much less frequent than non-sentence boundaries) and to decrease classifier variance, we use a combination of ensemble sampling and bagging [12]. The transition probabilities are based on an n-gram language model (LM), which is trained by explicitly including the sentence boundary as a token in the vocabulary. In this work, trigram models with modified Kneser-Ney smoothing were employed. To find the sentence boundaries, the model performs forward-backward decoding in which the word/event pairs correspond to hidden states, and the words and prosodic feature vectors to observations.

3.2. Maximum Entropy

Maximum Entropy (MaxEnt) is a discriminative model trained to directly discriminate among the target classes. It allows a natural combination of potentially overalapping features coming from multiple knowledge sources. As textual features, we used n-grams of words, AIC, and POS tags up to trigrams spanning across or neighboring with the inter-word boundary in question. To capture word repetitions, we also used a binary feature indicating whether the word before the boundary is identical with the following word or not. As with HMM, prosodic information is used via the decision tree prosody model, but unlike in HMM, the prosodic probabilities in the MaxEnt model were not used directly. Since MaxEnt usually does not perform well dealing with many real-valued features, 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 sequence of binary features: $p > 0.1$, $p > 0.2$, ..., $p > 0.9$. For all our experiments with MaxEnt, we employed the `MegaM` toolkit.[1]

3.3. Boosting

Our third approach is based on adaptive boosting, a machine learning method in which many weak learning algorithms are combined to produce an accurate classifier. Each weak classifier is built based on the outputs of previous classifiers – subsequent classifiers are tweaked in favor of those instances misclassified by previous classifiers. In this work, we use weak classifiers that have a basic form of one-level decision trees (stumps) introduced by Schapire and Freund [13]. We used the same textual features as in the MaxEnt model. The prosodic features were used directly in the boosting model, unlike the previous two approaches 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. In our experiments, the ICSIboost implementation of the boosting algorithm was employed.[2]

4. Experiments

4.1. Data and Experimental Setup

We use two Czech corpora – broadcast news (BN) that is mostly read speech, and broadcast conversations (BC) consisting of mostly spontaneous speech. Both corpora were annotated based on LDC's Metadata Extraction (MDE) standard [14]. The annotation included labeling of sentence-like unit (SU) boundaries, which were used in this work. The SU annotation guideline was designed to achieve good annotation consistency even on conversational speech. The two corpora differ in the distribution of SU and non-SU interword boundaries. The SU percentage is 8.1% for BN and 6.8% for BC, which means that SUs in BC are on average slightly longer. As also shown by the corpora analysis [14], BC is more conversational than BN – for example, the proportion of "deletable" words (fillers and edits) is 9.8% in BC but only 1.1% in BN. This may pose more difficulty for the task of sentence segmentation in BC.

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

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

The data in each corpus were split into a training set, a development set, and a test set. For BN, the data sets comprised 174.8k words for training, 28.2k for development, and 31.2k words for testing. For BC, the data included 159.1k words for training, 24.1k words for development, and 24.6k words for testing. All experiments were evaluated using both human-generated reference transcripts (REF) and automatic speech recognition (ASR) transcripts. The ASR output was obtained from the UWB LVCSR system tailored for real-time recognition of highly inflective languages [15]. The overall word error rates were 12.4% for BN and 29.3% for BC.

In addition to the MDE-annotated data, we also used a text corpus of Czech broadcast transcripts for training LMs. This corpus is much larger than the MDE corpora – it contains 107M words. Note that this is just a text corpus so we do not have any prosodic features associated with these words. This additional textual data is used in the three models differently. In the HMM approach, the auxiliary LM is incorporated with the baseline LMs during testing. However, MaxEnt and boosting do not have a separate LM, and they assume that all features are available during training. Therefore, 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. In the MaxEnt model, we thresholded the LM probabilities and used binary features; whereas for the boosting model, the auxiliary posteriors were used directly.

We measure sentence segmentation performance using a classification error rate called "Boundary Error Rate" (BER) [2]. It is defined as

$$BER = \frac{Ins + Del}{N_W} \quad [\%]$$

where Ins denotes the number of false SU boundaries, Del the number of misses, and N_W the number of words in the test set.

4.2. Results and Discussion

We display comparisons of the three modeling techniques for all the evaluation test sets (BN REF, BN ASR, BC REF, and BC ASR) in three figures differing in information sources used (textual, prosodic, and both). The bars in all the figures show sentence boundary detection error rates (thus lower is better). Figure 1 visualizes results for models based on only textual information. MaxEnt was the most successful approach for the BN corpus, while HMM was the best performing method for the BC corpus. The superiority of HMM over other models for BC was greater than that of MaxEnt over others for BN – the former is significant at $p < .05$ (REF) and $p < .01$ (ASR) using a Sign test, whereas for BN, the prevalence of MaxEnt over the second best model is not significant.

Figure 2 shows BERs achieved by models based only on prosodic information. Note that because the HMM and the MaxEnt approach share the same prosodic model based on decision trees, we only display results using decision trees versus boosting for the prosodic features in the figure. The decision tree model outperformed the boosting model in all test sets. For BN, the difference is significant at $p < .01$ for both REF and ASR. For BC, the difference is significant in ASR conditions. The results indicate that the bagged full decision trees are more powerful than boosted stumps in handling prosodic features. The power of the boosting model increases when textual features are incorporated, as will be shown below.

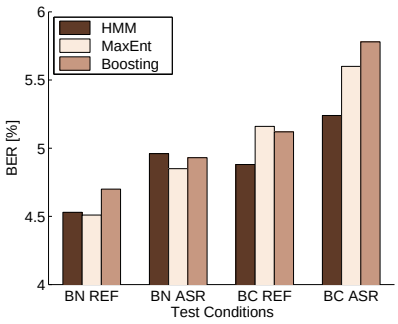

Figure 1: *SU segmentation error rates [BER%] for individual models when only textual information is used.*

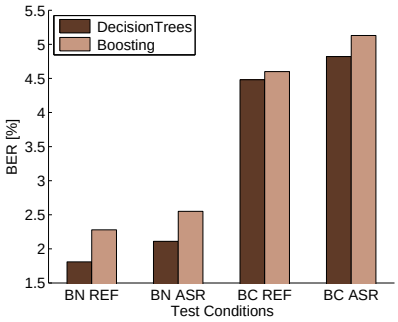

Figure 2: *SU segmentation error rates [BER%] when only prosodic information is used.*

To better understand the prosodic model, we also analyze feature usage statistics. The usage metric [2] 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 one. We observe some differences in feature usage between BN and BC. Among the duration features, normalized duration of the last vowel was the most important feature for BN (9.4% of overall usage), while the raw word duration feature was dominant for BC (16.5%). From the pitch features, BN heavily uses a feature reflecting the ratio between the last f_0 value and the speaker's f_0 baseline (12.0%), suggesting that radio anchors tend to mark statement boundaries with significant pitch falls. This feature was also important for BC, but to a lesser extent (4.1%). In both corpora, the most used energy feature was normalized maximal RMS value from the word following the boundary in question (6.3% in BN, 5.8% in BC). We also compared the Czech feature usage statistics with English. Since there are no published usage statistics for English BC, we could make the comparison only for BN using the statistics from [16]. The most used feature was the same for both languages – pause duration at the current boundary. A comparison of other frequent features revealed that features capturing final lengthening were more important for English, while features capturing final pitch fall were more important for Czech. This finding is in agreement with the fact that in comparison to English, Czech offers less opportunity for final lengthening because length also serves a lexical function in Czech.

The results of the models relying on both information sources are visualized in Figure 3. For the BN corpus, the best results were achieved by the boosting model, however, the gaps

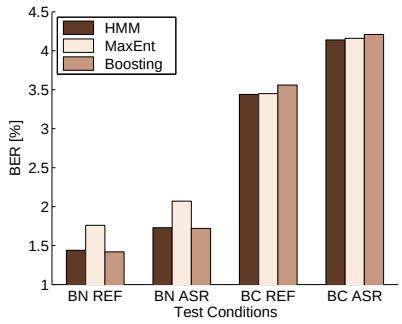

Figure 3: *SU segmentation error rates [BER%] for individual models when both textual and prosodic information is used.*

	BN		BC	
Model	**REF**	**ASR**	**REF**	**ASR**
SU Prior	8.11	8.01	6.81	6.89
HMM	1.44	1.73	3.44	4.14
MaxEnt	1.76	2.07	3.45	4.16
Boosting	1.42	1.72	3.56	4.21
Combination	**1.27**	**1.61**	**3.22**	**3.98**

Table 1: *SU segmentation error rates [BER %] for HMM, MaxEnt, Boosting, and their combination using both textual and prosodic information. 'SU Prior' corresponds to the percentage of SUs in the test sets.*

between boosting and HMM were statistically insignificant for both REF and ASR. On the other hand, MaxEnt performed significantly worse than BoosTexter or HMM, for both test conditions (significant at $p < .01$). The inferiority of MaxEnt on BN may be due to its use of prosodic information – MaxEnt uses discretized decision tree posteriors instead of using the prosodic features directly, which may miss some prosodic cues in BN where the boundaries are often prosodically marked by the professional anchors and reporters. In the BC corpus, the best results were achieved by the HMM model, but there is no significant difference among the three approaches.

Table 1 summarizes the results for the three models using both textual and prosodic information. In addition, the last row shows the results of a model combining HMM, MaxEnt, and boosting via posterior probability interpolation. The interpolation weights 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-approach models are significant at $p < .01$ for BN REF and BC REF, and at $p < .05$ for BN ASR and BC ASR.

5. Conclusion

This paper evaluated automatic sentence segmentation of spoken Czech based on textual and prosodic information, examining three different modeling approaches – HMM, MaxEnt, and adaptive boosting. All the models were evaluated them on two Czech corpora (broadcast news and broadcast conversations) using both manual and speech recognition transcripts. Among the three approaches, HMM showed most consistently good results, typically producing best or close to best results. Furthermore, the results suggest that the main advantage of the boosting model is in an effective combination of textual and prosodic cues – this approach was never the best when only one of the knowledge sources was employed. The MaxEnt model showed inferior performance on the BN corpus, probably because of a less precise approach to capturing rich prosodic information, which is important for SU segmentation of planned speech. Regarding prosodic information, we found that it benefits SU segmentation more for BN than for BC. Feature analysis revealed difference in prosodic feature usage patterns between BN and BC, as well as between English and Czech. Overall, superior results for all our test sets were achieved by a model combining HMM, MaxEnt, and boosting via posterior probability interpolation. All the improvements over the best single-approach models were statistically significant.

6. References

[1] V. Warnke, R. Kompe, H. Niemann, and E. Nöth, "Integrated dialog act segmentation and classification using prosodic features and language models," in *Proc. Eurospeech*, Rhodes, Greece, 1997.

[2] 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.

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

[4] D. Hillard, M. Ostendorf, A. Stolcke, Y. Liu, and E. Shriberg, "Improving automatic sentence boundary detection with confusion networks," in *Proc. HLT/NAACL*, Boston, USA, 2004.

[5] U. Guz, B. Favre, D. Hakkani-Tur, and G. Tur, "Generative and discriminative methods using morphological information for sentence segmentation of Turkish," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, pp. 895–903, 2009.

[6] J. Kolář, Y. Liu, and E. Shriberg, "Speaker adaptation of language and prosodic models for automatic dialog act segmentation of speech," *Speech Communication*, vol. 52, pp. 236–245, 2010.

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

[8] 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*, Taipei, Taiwan, 2009.

[9] J. Kolář and Y. Liu, "Automatic sentence boundary detection in conversational speech: A cross-lingual evaluation on English and Czech," in *Proc. ICASSP*, Dallas, TX, USA, 2010.

[10] P. Brown, V. D. Pietra, P. de Souza, J. Lai, and R. Mercer, "Class-based n-gram models of natural language," *Computational Linguistics*, vol. 18, no. 4, pp. 467–479, 1992.

[11] 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.

[12] 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.

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

[14] J. Kolář and J. Švec, "Structural metadata annotation of speech corpora: Comparing broadcast news and broadcast conversations," in *Proc. LREC*, Marrakech, Morocco, 2008.

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

[16] Y. Liu, "Structural event detection for rich transcription of speech," Ph.D. dissertation, Purdue University, 2004.