Using multiple linguistic features for Mandarin phrase break prediction in maximum-entropy classification framework

*Yu Zheng  **Byeongchang Kim  *Gary Geunbae Lee

*Department of CSE, POSTECH, Pohang, South Korea
**School of Computer & Information Communications Engineering, Catholic University of Daegu, Daegu, South Korea
{zhengyu, bckim, gbler}@postech.ac.kr

Abstract

We model Mandarin phrase break prediction as a classification problem with three level prosodic structures and apply conditional maximum entropy classification to this problem. We acquire multiple levels of linguistic knowledge from an annotated corpus to become well-integrated features for maximum entropy framework. Five kinds of features were used to represent various linguistic constraints including POS tag features, lexical features, phonetic features, length features, and distance features. Experiment results show that our method performs better than the previous methods and the conditional maximum entropy (ME) model is very effective for data sparseness problem in Mandarin phrase break prediction.

1. Introduction

Assigning the appropriate phrase breaks in text-to-speech systems is important for naturalness and intelligibility. Linguistic researchers have shown that the spoken language is structured as a hierarchy of prosodic units, including phonological phrase, intonation, and utterance [1]. However, the text language is often structured by syntactic units, such as words and phrases, which are not equivalent to prosodic ones. But we suppose the syntactic information would provide important cues for prosodic phrase prediction.

Many techniques have been introduced to predict phrase break, such as using Recurrent Neural Network (RNN) [2], Hidden Markov Model (HMM) [3], POS bi-gram and CART [4] and rule learning with C4.5 or TBL [5] [6].

Min Chu and Yao Qian [4] proposed CART-based approach in four-class prosodic structure and their method shows high accuracy of 83%.

Zhao and Tao with others [5] [6] proposed automatic rule-learning approach with two typical rule-learning algorithms (C4.5 and TBL). They report a higher accuracy of 87.9% where they used POS features, lexical features and length features. They added chunking features and achieved even better accuracy of 90.0%, but they only used two-class prosodic structure to evaluate the accuracy.

We treat entire phrase break prediction as a classification problem and apply conditional maximum entropy (ME) model. Various linguistic information is represented in the form of features, and five kinds of features were used in our system including POS tag features, lexical features, phonetic features, length features, and distance features.

One serious problem of Mandarin phrase break prediction is that we usually do not have a large sized phrase break annotated corpus. So, we have to acquire multiple levels of linguistic information from only a small sized annotated corpus. Here, the data sparseness is a critical problem in Mandarin phrase break prediction, and we show that ME framework is very effective to capture useful features in the sparse training data environment.

The remainder of this paper is organized as follows: Section 2 briefly introduces the ME framework with some justification. In section 3, we present our prosodic phrase framework and the five kinds of features used in the system. The effectiveness of our proposed method is verified by the experimental results given in section 4. Finally, conclusions are provided in section 5.

2. Conditional Maximum Entropy Framework

In general, a probabilistic model expresses the relation between input information and output information by a probability distribution, which is estimated from a training corpus. The more various types of input information is used for estimation, the more accuracy can be obtained to guess the output exactly. However, too many kinds of input information cause a data sparseness problem.

In ME modeling, relations between input information and output information are expressed as feature functions. The model whose entropy is maximized under the constraints defined by feature functions must be estimated. It means that the model would be close to the uniform distribution for input information which rarely appears in the training corpus. Mandarin phrase break prediction must relate many kinds of linguistic features from small Mandarin phrase break annotated corpus, and ME modeling can be effective for various feature integration.

A classifier obtained by means of an ME technique consists of a set of parameters that are estimated using an optimization procedure. Each parameter is associated with one feature observed in the training data. The main purpose is to obtain the probability distribution that maximizes the entropy, that is, maximum ignorance is assumed and nothing apart from the training data is considered. In addition to the effectiveness for sparse data problem, some advantages using the ME framework are that even knowledge-poor features can be estimated accurately by asking only elementary questions to the surrounding contexts. So, we decide to use ME modeling for our Mandarin phrase break prediction task.

The ME model allows experimenters to encode various dependencies freely in the form of features [7]. Let us assume a set of contexts X and a set of classes C. The system chooses the class C with the highest conditional probability in the
We thank voiceware, Inc. to identify the three levels of Mandarin prosodic structure. The parameter intoonation phrase [8]. We used four types of boundaries to POS/C [9], unknown word prediction system [10], and

\[ P(c | x) = \frac{1}{Z(x)} \prod_{j=1}^{k} \lambda_{j}^{f_{j}(c, x)} \] (1)

\[ Z(x) = \sum_{c} \prod_{j=1}^{k} \lambda_{j}^{f_{j}(c, x)} \] (2)

3. Prosodic Phrasing

3.1. Phrase break prediction architecture

Linguistic research has suggested that Mandarin is structured in a prosodic hierarchy, in which there are mainly three levels of prosodic units: prosodic word, prosodic phrase, and intonation phrase [8]. We used four types of boundaries to identify the three levels of Mandarin prosodic structure. The following shows a prosodic boundary annotation example:

```
妹妹/ B1 她和/ B0 出/ B2 的/ B0 地点/ B2
才/ B1 将/ B0 他/ B1 送到/ B0 了/ B1 医院/ B3.
```

In this example, B0 represents no boundary, B1 for boundary of prosodic word, B2 for boundary of prosodic phrase, and B3 for boundary of intonation phrase.

Input sentence

Segmentation

POS tagging

Unknown word Prediction

Pinyin generation

Feature Exaction

ME-based phrase break prediction

Model parameters

output sentence

Figure 1: Overall architecture of the proposed method

Figure 1 shows the overall architecture of our ME-based phrase break prediction. In our research, we used previously developed word segmentation and POS tagging system called POSTAG/C [9], unknown word prediction system [10], and Chinese pinyin generation system [11]. Using the previous linguistic analysis systems, we can extract five kinds of linguistic and phonetic features, such as POS tag features, lexical word features, phonetic features, length features, and distance features. Then we use the conditional maximum entropy model parameters estimated by L-BFGS method [12] to predict the phrase break.

3.2. Multiple linguistic features

The features used in our model include:

1. POS tag features

   POS tag features are the most commonly used features in phrase break prediction. We include current, previous two and next two POS tags: P-2, P-1, P0, P1 and P2, through extensive experiments. Our POS tag features are binary features.

2. Lexical word features

   Word features include current, previous one and next one lexical words: W-1, W0 and W1. They are also binary features.

3. Phonetic features

   Phonetic features include right character’s pinyin of current word and left character’s pinyin of next word. Pinyin features are also binary features.

4. Length features

   Length features correspond to the word length in the number of characters of the current, previous two and next two words: WLEN-2, WLEN-1, WLEN0, WLEN1 and WLEN2. Lengths of words are floating-point value features. In this paper, we normalized the features using the function as follows:

\[ \text{Normalized word length} = \frac{\text{current word length}}{\text{maximum word length}} \]

5. Distance features

   Distance features are the distance in characters from the current point to the beginning (dis_start) and the end (dis_end) of a sentence. Distance features are also floating-point value features. In this paper, we normalized the features using the function as follows:

\[ \text{Normalized distance} = \frac{\text{distance}}{\text{sentence length}} \]

4. Experimental Results

4.1. Corpus

The experiments are performed on a database provided by Voiceware Inc\(^1\). The database has 2197 sentences, 52546 Chinese characters, which constitute 25974 Chinese words. The database is POS tagged, pinyin annotated and break-labeled with four class prosodic structures. We divide the database into 10 parts and conducted 10-fold cross validation.

\(^1\) http://www.voiceware.co.kr . We thank voiceware, Inc. to provide their Chinese TTS database for our research.
4.2. Performance measures

The performance is assessed with reference to N, the total number of junctures (spaces in the texts including any type of phrase breaks), and to B, the total number of real phrase breaks (only \(B_1\), \(B_2\) and \(B_3\)) in the test set. The errors can be divided into insertions, deletions and substitutions. An insertion (I) is a break inserted in the test sentence, where there is not a break in the reference sentence. A deletion (D) occurs when a break is marked in the reference sentence but not in the test sentence. A substitution (S) is the number where we correctly recognize a break existing, but assign a wrong level, such as tag \(B_1\) as \(B_3\), or \(B_1\) as \(B_2\). We used the following performance measures using these definitions [13].

\[
\text{Break Correct}(B,C) = \frac{B - D - S}{B} \times 100\% \\
\text{Juncture Correct}(J,C)(\text{Accuracy}) = \frac{N - D - S - I}{N} \times 100\%
\]

For more extensive comparisons, we used another performance measure, called adjusted score, which refers to the prediction accuracy in proportion to the total number of phrase breaks [14].

\[
\text{Adjusted Score}(A,S) = \frac{JC - NB}{1 - NB}
\]

where \(NB = \frac{N - B}{N}\) means the proportion of no breaks to the number of inter-word spaces and \(JC\) means the \text{Juncture Correct}.100.

4.3. Experiment analysis

We performed two experiments to show the phrase break prediction results of our ME-based method. First, using several types of feature combinations to show the best feature selection, we can select feature combinations for our ME framework. In the second experiment, we used POS bi-gram statistical model, HMM-based model and manual heuristic rules to predict phrase breaks on the same corpus for various levels of comparison. We evaluate with four boundary classes: \(B_0\), \(B_1\), \(B_2\) and \(B_3\), and also two boundary classes: \(B_{0}\) and \(B_{123}\). In this case, we merged three classes \(B_1\), \(B_2\) and \(B_3\) into \(B_{123}\).

In Table 1, POS(-2,-1,0,1,2) represents the current, previous two, and next two POS tag features. Word(-1,0,1) represents the current, previous one, and next one lexical word features. PY(-1,L,R,1) represents the current word’s left and right pinyin, left word’s right pinyin, and right word’s left pinyin. Wlen(-1,0,1) shows that we use word length features for current, previous and next words. Dis_start represents the distance in characters from the current point to the beginning of the sentence, while Dis_end represents the distance to the end of the sentence. Acc4 represents the accuracy (Juncture Correct) for four boundary classes, while Acc2 is the Juncture Correct for two boundary classes.

As the experiment shows, we can draw various conclusions on the effect of feature selection for phrase break prediction. Table1 shows the accumulated performance of best feature selection in each feature class.

1. the POS tag is a baseline feature, and the window size of 5 is the best value in this class.
2. Adding lexical word features are helpful and the window size of 3 is the best value in this class.
3. Pinyin features are also slightly useful for the phrase break prediction, and the best case is the right pinyin of the current word and the left pinyin of the right word.
4. Length is a useful feature and the window size of 5 is slightly better than the window size of 3 in our experiment.
5. Distance feature actually decreases the performance, but through the feature value’s normalization, distance feature becomes helpful to phrase break prediction.
6. Comparing Acc4 and Acc2, we discover that prosodic word boundaries (\(B_1\)) can be more accurately predicted than the prosodic phrase ones (\(B_2\)), and the distance features are more helpful to the prosodic phrase prediction than to the prosodic word prediction.

![Table 1: Results of the best feature selection](image)

Table 2 shows the comparison of our ME-based method with the previous two methods: HMM and POS bigram. In this comparison, we also used about 85 manual prediction rules for Chinese prosodic segmentation [15]. The following box shows some sample rules and we want to demonstrate the ME framework can systematically incorporate the heuristic manual rules in the feature selection level.
As shown in Table 2, both the accuracy of POS bi-gram and the accuracy of HMM-based method are lower than that of the ME-based method. Moreover, ME-based method alone outperforms the heuristic rule-based error correction of POS bi-gram and HMM results. However, error correction rules are less useful for ME-based method because the rules are already effectively integrated into the ME features. This is another benefit of the ME-based method because generating handcraft prediction rules is a laborious work.

**Table 2: Comparison with other methods**

<table>
<thead>
<tr>
<th></th>
<th>B_C %</th>
<th>Acc4 (J_C) %</th>
<th>A_S (J_C) %</th>
<th>Acc2 (J_C) %</th>
</tr>
</thead>
<tbody>
<tr>
<td>POS + bi-gram</td>
<td>81.82</td>
<td>78.79</td>
<td>0.703</td>
<td>81.82</td>
</tr>
<tr>
<td>POS + rules</td>
<td>85.80</td>
<td>81.02</td>
<td>0.732</td>
<td>85.86</td>
</tr>
<tr>
<td>HMM</td>
<td>75.62</td>
<td>75.05</td>
<td>0.623</td>
<td>75.62</td>
</tr>
<tr>
<td>HMM + rules</td>
<td>79.80</td>
<td>79.06</td>
<td>0.676</td>
<td>79.81</td>
</tr>
<tr>
<td>ME</td>
<td>85.80</td>
<td>86.48</td>
<td>0.810</td>
<td>90.33</td>
</tr>
<tr>
<td>ME + rules</td>
<td>85.92</td>
<td>86.55</td>
<td>0.812</td>
<td>90.54</td>
</tr>
</tbody>
</table>

5. Conclusion and Future works

We proposed a conditional maximum entropy (ME) model for Mandarin phrase break prediction tasks. We analyzed several combinations of linguistic features in order to identify which features are the best candidates for ME-based phrase break prediction. The results obtained from our proposed system show that the selected best feature sets guarantee the success of the disambiguation method. Because the ME model allows experimenters to encode various dependencies freely in the form of features and often Mandarin-phrase-break related features are dependent on each other, the ME model’s feature selection is more flexible than that of other machine learning models.

As shown in the experimental results, the ME-based model is more effective for data sparseness than other statistical methods such as the POS bi-gram method and the HMM-based method.

For future works, we will explore more complicated features that benefit Mandarin phrase break prediction, such as chunk-level features and parsing-level features to improve our performance.

Acknowledgements

This research was supported by grant NO. (R01-2003-000-10181-0) from the basic research program of the KOSEF (Korea Science and Engineering Foundation).

References