Prosodic Boundary Prediction based on Maximum Entropy Model with Error-Driven Modification[1]

Xiaonan Zhang, Jun Xu, Lianhong Cai

Key Laboratory of Pervasive Computing (Tsinghua University), Ministry of Education, Beijing 100084, P.R.China
zxndelia724@gmail.com, xujun00@mails.tsinghua.edu.cn, clh-dcs@tsinghua.edu.cn

Abstract. Prosodic boundary prediction is the key to improving the intelligibility and naturalness of synthetic speech for a TTS system. This paper investigated the problem of automatic segmentation of prosodic word and prosodic phrase, which are two fundamental layers in the hierarchical prosodic structure of Mandarin Chinese. Maximum Entropy (ME) Model was used at the front end for both prosodic word and prosodic phrase prediction, but with different feature selection schemes. A multi-pass prediction approach was adopted. Besides, an error-driven rule-based modification module was introduced into the back end to amend the initial prediction. Experiments showed that this combined approach outperformed many other methods like C4.5 and TBL.

Keywords: Prosodic Boundary Prediction, Maximum Entropy Model, Error-Driven Rule-Based Modification

1 Introduction

When people talk, they rarely speak out a whole sentence without a break. Instead, an utterance is divided into smaller units with perceivable boundaries between them. These phonetic spurts or chunks of speech, which are commonly known as prosodic units, signal the internal structure of the message and serve important function in helping people to clearly express their ideas as well as their feelings.

These units are different in their categories, levels, and boundary strength. Smaller and lower-level units are contained in larger and higher-level units to form a prosodic hierarchy. In Mandarin Chinese, this hierarchical structure is often simplified to 3 layers [1] (from down to top): prosodic word, prosodic phrase and intonation phrase.

In current TTS, input text is firstly processed by a Text Analysis Model, whose output is a string of syntactic words, each with a POS tagging. Then its prosodic structure is expected to be constructed out of linguistic information to enhance the naturalness and understandability of the synthetic speech. However, as grammatical structure does not necessarily correspond to its prosodic counterpart, misjudgments

[1] This work is supported by China National Natural Science Foundation(60433030, 60418012)

always occur in assigning the proper prosodic boundary, which has become the major impediment for TTS systems to achieve human-like performance.

That's why more and more attention has been paid to addressing the problem of automatic prosodic boundary prediction. For Mandarin, as intonation phrases are usually distinguished by punctuation marks, most efforts focus on locating prosodic word and prosodic phrase boundaries based on syntactic information.

In the earlier time, rule-based methods were usually adopted. It mainly starts from Gee and Grosjean's work on performance structures [2], and has had various extensions over the years. For Chinese, similar investigation has also been carried out into the formation of prosodic constitutes, such as that reported by Jianfen Cao [1][3] and Hongjun Wang [4]. The central idea of all these work is to find some explicit rules that could recreate the prosodic structure of a sentence from syntax, by way of a large number of experiments and empirical observation. This method is easily explicable and understandable, but also poses strict demand for the system developer to summarize these rules. Moreover, it is hard to update and improve, and the set of rules is usually constrained to one branch of language, which hinders its general application.

With the availability of increasing prosodically annotated corpora and the rapid development of statistical learning, stochastic-based approach has been more and more widely used in prosodic boundary prediction. As in most cases, it is assumed that syntactic word is the smallest unit (i.e. leaf node) in a prosodic hierarchy tree, the task of building prosodic structure could be reduced to deciding the type for each syntactic word boundary, which is actually a classification problem. Thus many different statistical methods used for classification have been tried, such as Classification and Regression Tree (CART) used by Wang and Hirschberg [5], and Hidden Markov Model proposed by Paul and Alan [6]. Researchers in Chinese have also begun to adopt this approach during recent years. Besides those mentioned above, Zhao has described methods for automatically predicting prosodic phrase by combining decision tree and TBL [7]. And in Li's experiment, he attempted to predict prosody phrase break based on Maximum Entropy Model [8]. Generally, these methods relate each boundary site with some features (e.g. length and POS of adjacent words). By extracting and absorbing these features from a large collection of annotated sentences, a statistical model is trained and then applied to unlabeled texts. For each potential boundary site in the text, a probability is estimated for each possible outcome, and the one with largest likelihood is determined as the correct type.

In this paper, we proposed to predict the prosodic boundary using Maximum Entropy Model, which nowadays has gained more and more popularity with NLP. Unlike previous efforts, we applied it to both prosodic word and prosodic phrase boundary labeling, and a multi-pass approach was employed for the latter task. Moreover, an error-driven rule-based modification module was added at the back end to improve the performance furthermore.

The rest of this paper is organized as following. Section 2 first introduced the Maximum Entropy Model briefly. Then the feature selection method and the multi-pass procedure for prosodic boundary prediction are presented. Section 3 described the back-end modification module. Experiments and results are given in Section 4. Section 5 gives conclusions.

2 Prosodic Boudary Prediction based on Maximum Entropy Model

In our experiment, a basic assumption is that a prosodic boundary only occurs at syntactic word boundaries. For Mandarin Chinese, it is reasonable as statistics show that only 6% of prosodic words are part of a long syntactic word, and all the rest agree with this assumption.

Given a string of consecutive syntactic words, for each boundary between two of them (say w_i and w_{i+1}), there are 3 types: LW (w_i and $w_{i+1}$are within the same prosodic word), PW(within the same prosodic phase but different prosodic words) and PPh(within different prosodic phrases). Then our task comes down to deciding the right type for each syntactic word boundary, which could be accomplished with a Maximum Entropy Model.

2.1 Maximum Entropy Modeling

Consider a random process that produces an output value *y* based on some contextual information *x*, with *x* and *y* being a member of a finite set X and Y respectively. In our case, *y* is the type of a syntactic word boundary (i.e. LW, PW or PPh), and *x* could include any available information about that boundary.

Our task is to construct a stochastic model that accurately represents the behavior of the random process. In other words, it should give a reliable estimation of *p(ylx)*, which denotes the conditional probability that, given a context x, the process will output y.

For this purpose, we observe the behavior of the random process for some time, collecting N samples $(x_1, y_1), (x_2, y_2) . . . . . (x_N, y_N)$. To express these facts, a **feature function** or **feature** for short is defined as:

$$fi(x,y) = \begin{cases} 1 & \text{if } y = y_i \text{ and } x = x_i \\ 0 & \text{Otherwise} \end{cases}$$

The expected value of each feature f_i with respect to the statistics of training samples could then be calculated as:

$$\bar{p}(f_i) = \sum_{(x,y)} \bar{p}(x,y) f_i(x,y) \quad , \qquad . \tag{1}$$

where $\bar{p}(x,y)$ is the empirical probability distribution of the samples, defined by:

$$\bar{p}(x,y) \equiv \frac{1}{N} \times \textit{number of times that } (x,y) \textit{ occurs in the sample} \quad . \tag{2}$$

On the other hand, the expected value of f_i with respect to the unknown model *p(ylx)* is:

$$p(f_i) = \sum_{(x,y)} \bar{p}(x) p(y|x) f_i(x,y) \ , \tag{3}$$

where $\bar{p}(x)$ is the empirical distribution of x in the training sample. We require the model to accord with the observed statistics by constraining this value to be the same as the expected value of f in the training set. That is, for each f_i

$$p(f_i) = \bar{p}(f_i) \ . \tag{4}$$

Requirement (4) is called a **constraint equation** or simply a **constraint.**
Combining (1), (3) and (4) we have:

$$\sum_{(x,y)} \bar{p}(x)\, p(y \mid x)\, f_i(x,y) = \sum_{(x,y)} \bar{p}(x,y)\, f_i(x,y) \ . \tag{5}$$

Suppose we have n features, then all the probability distribution that satisfy the constraints exerted by these features constitute a set C:

$$C \equiv \left\{ p(y \mid x) \mid p(f_i) = \bar{p}(f_i) \quad for \ \ i \in \{1, 2, ..., n\} \right\} \ . \tag{6}$$

Among all the models p in C, the maximum entropy philosophy dictates that we select the one with maximum conditional entropy [9]

$$H(p) \equiv -\sum_{x,y} \bar{p}(x)\, p(y \mid x) \log p(y \mid x) \ , \tag{7}$$

and

$$p^* = \arg\max_{p \in C} H(p) \ . \tag{8}$$

It is a constrained optimization problem to find p^*. The target maximum entropy model has the following form[9]:

$$p^*(y \mid x) = \frac{1}{Z_\lambda(x)} \exp(\sum_i \lambda_i f_i(x,y)) \ . \tag{9}$$

where $Z_\lambda(x)$ is a normalizing constant and λ_i is a Lagrange multiplier which is commonly computed from the training set using GIS algorithm. Detailed steps are omitted here.

2.2 Feature Selection Strategy

The principle of Maximum Entropy Model is to agree with all that is known and assume nothing about what is unknown. Yet it poses another important question: how to find appropriate facts that are most relevant to the task in hand? Put another way, how to select a limited number of features that represents the ‘known’ fully and accurately?

In the first place, as prosodic phrase lies above prosodic word in the prosodic hierarchy, it should exhibit some ‘higher-level’ features than the latter. Taking this into account, we built two distinct models by incorporating into them different features for prosodic word and prosodic phrase prediction respectively.

Like Li [8], we used a semi-automatic approach for feature selection. First, feature ‘templates’ are manually designed, which in effect defined the space of candidate features; then the most “useful” features are selected automatically using a simple count cut-off method with a threshold of 3.

The feature ‘templates’ are so devised as to capture as much information about the random process as possible. For our specific application, most commonly used features include POS (Part of Speech Tagging), WLen (length in syllables) and Word (the word itself) of the words surrounding the boundary, which have also proved to be the most important determinants of prosodic boundary types [10]. On account of this, we added them into too both templates, with a window length of 2 for POS, i.e. we considered the POS of 2 words immediately before and after the boundary in question, and a window length of 1 for WLen and Word. A point to note, though, is that ‘word’ has different meaning under the two scenarios. For prosodic word, it indicates syntactic word and is readily available from the input text. For prosodic phrase, which is built upon prosodic words rather than syntactic words, the meaning accordingly changes to prosodic word. Here ‘POS’ property of a prosodic word is acquired by simply concatenating POS’s of the syntactic words it contains (e.g. POS of ‘我们/rr 的/ud’ is ‘rr ud’)

Besides these widely used features, another category of features—‘dynamic feature’ were also introduced into the templates. The first is ‘lastType’, which denotes the last prosodic boundary type. The motive for adding this information came from the observation that current boundary type is influenced by that of last one, which applies to prosodic word as well as prosodic phrase boundary. For example, a ‘lastType’ of PPh could well reduce the possibility that current boundary is still PPh.

The second was specially proposed for prosodic phrase segmentation. We noted that, to a large extent, insertion of prosodic phrase boundaries in natural spoken language is to balance the length of the constituents in the output. Hence it is not surprising that most PPh breaks occur in the middle part of a long sentence, and a prosodic phrase is usually 5~7 syllables long, but rarely shorter than 3 or longer than 9 syllables. For this reason, we took into consideration length measures by including ‘dBack’ and ‘dFront’ in our templates for prosodic phrase prediction, which means the distance (in syllables) from current boundary to the last and next nearest PPh location.

This category of features is by definition ‘dynamic’ in that they rely on the result of previous prediction, and remains unknown until judgment on preceding boundaries has been made. By contrast, the usual ‘static’ features are fixed and known all the way once the input is given.

The feature templates contained both atomic and composed ones. Atomic templates considered only one element mentioned above, while composed templates are combination of atomic ones. Table 1 lists all the atomic templates.

Table 1. Atomic Templates Used in Two Maximum Entropy Model

<table>
<tr><th></th><th colspan="2">PW&PPh Prediction 1^{st} Pass (Model 1)</th><th colspan="2">PPh Prediction 2^{nd}&3^{rd} Pass(Model 2)</th></tr>
<tr><td rowspan="6">Atomic Templates</td><td>Symbol</td><td>Meaning</td><td>Symbol</td><td>Meaning</td></tr>
<tr><td>POS-2
POS-1
POS+1
POS+2</td><td>POS of the $1^{st}/2^{nd}$ syntactic word before/after the boundary</td><td>POS-2
POS-1
POS+1
POS+2</td><td>POS of the $1^{st}/2^{nd}$ prosodic word before/after the boundary</td></tr>
<tr><td>Word-1
Word+1</td><td>1^{st} syntactic word itself before/after the boundary</td><td>Word-1
Word+1</td><td>1^{st} prosodic word itself before/after the boundary</td></tr>
<tr><td>WLen-1
WLen+1</td><td>length of 1^{st} syntactic word before/after the boundary</td><td>WLen-1
WLen+1</td><td>length of 1^{st} prosodic word before/after the boundary</td></tr>
<tr><td rowspan="2">lastType</td><td rowspan="2">boundary type after last syntactic word (LW/PW/PPh)</td><td>lastType</td><td>boundary type after last prosodic word (PW/PPh)</td></tr>
<tr><td>dFront/
dBack</td><td>distance from current position to last/next PPh boundary</td></tr>
</table>

2.3 Multi-Pass Prediction

In both our experience and experiments, we found that it's much easier to locate prosodic word boundaries accurately. It could be explained by the observation that distribution of this kind of boundaries largely depends on local syntactic constraints and exhibits more regular patterns that could be derived from low-level syntax analysis. On the other hand, prosodic phrasing is a compromise between the need to respect the syntax structure of the sentence and the prosodic constraints, which could hardly be decided in the normal one-pass classification solution.

That's why we came up with the idea of multi-pass prediction to determine prosodic phrase boundaries. The whole process is described in Figure 1. As mentioned earlier, 2 separate models were trained with different feature sets during the training stage. In testing, the 1^{st}-pass prediction used Model 1 at every syntactic word boundary to decide its type: LW, PW, or PPh. At this time, our major concern was to differentiate between PW and LW boundaries, and merely those most 'credible' PPh's were labeled as PPh's. That is, only when Model 1 decided that the probability of a boundary to be PPh is higher than a certain threshold (say threshold1), were we assured that it actually was PPh. Otherwise we still classified it as PW and left it to the next pass. It is worth to mention that though Model 1 was mainly targeted at PW prediction, it's sensible and necessary to label out some PPh's at the same time. For one thing, those PPh's with a high degree of confidence are mostly where we 'have to break' governed by syntax or grammatical constraints. For another, identifying these PPh positions also enabled us to acquire the 'dBack' feature in following predictions.

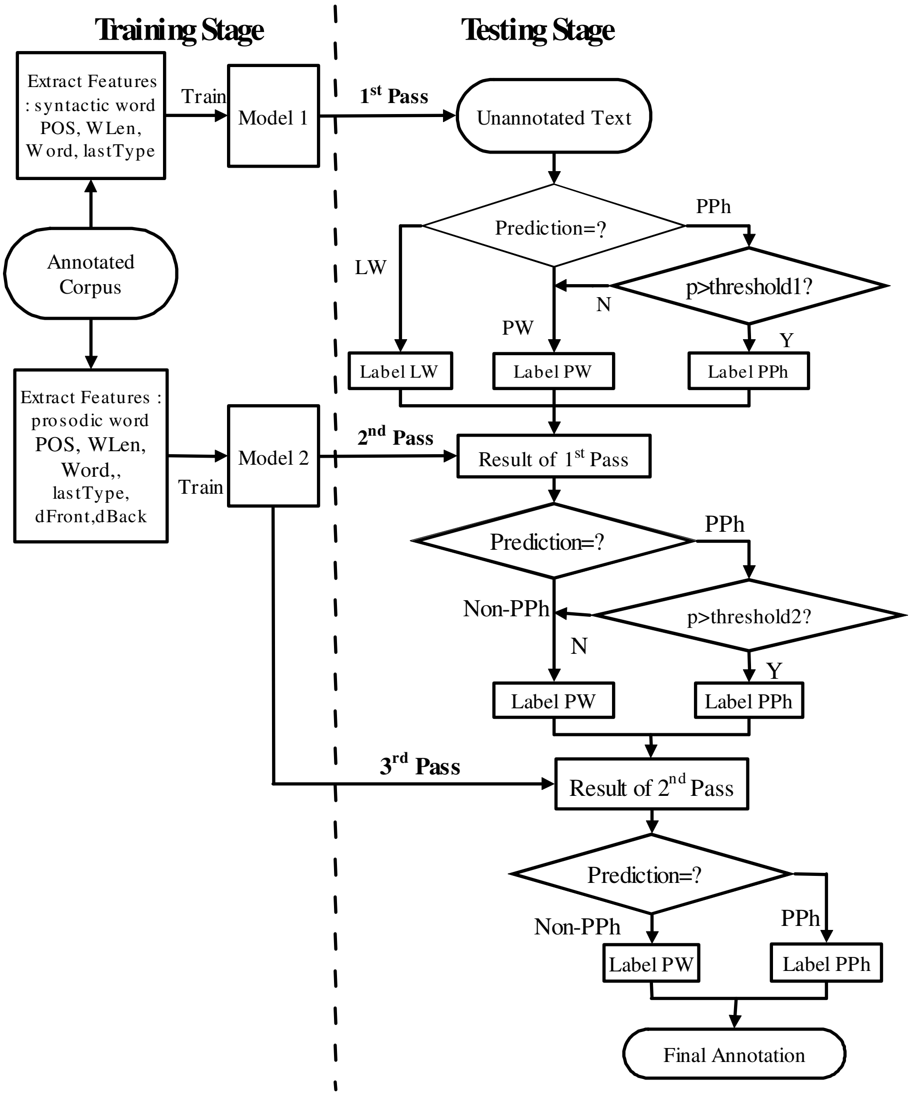

Fig. 1. Procedure of Multi-Pass Prediction

The 2nd and 3rd pass only worked on PW boundaries labeled in the 1st pass. They both used Model 2 to decide whether a PW indeed was PW, or should be classified as PPh. The only difference with these two pass is that during the 2nd pass, we still didn't take the result literally: only when the estimated probability of a boundary to be PPh was higher than threshold2, did we trust it to be an 'authentic' PPh. However, in the last pass, we accepted the model's judgment unconditionally.

We did this mainly because those PPh's decided in the latter stage of our prediction chiefly correspond to those breaks that we 'don't have to make but could make' out of prosodic constraints, and thus had better to be refined step by step to achieve the best balance in length.

Another question unaddressed is the set of threshold1 and threshold2, which turned out to have a considerable influence on the final outcome. After repeated experiments, it was found that a value of 0.65 and 0.7 for threshold1 and threshold2 respectively achieved the best performance.

3 Error-Driven Rule-Based Modification

In our preliminary experiment using only Maximum Model for prediction, we found that there were always some obvious mistakes that humans would never commit as they obviously contradicted to some 'fixed patterns' we were accustomed to. It then occurred to us that these mistakes might be corrected with manually-made rules.

3.1 Rules

Every rule is a pair with the form of 'predicate => action'. When and only when the pre-condition described by 'predicate' is satisfied, will a rule be activated, and then corresponding 'action' will be taken.

For example, the fact that 'A boundary succeeded by syntactic word "的"or "得"must be a LW boundary' could be written as the following rule:

$$WORD\text{-}1 = \text{的}\ or\ WORD\text{-}1 = \text{得} \Rightarrow Boudary \leftarrow LW$$

3.2 Basic Process

The rule-based modification module was added at the back-end of the system to amend the prediction from Maximum Entropy Model. To evaluate whether adding a rule does improve the performance, the metric 'F-Score' (detailed later in 4.1) was used. A brief working process is shown in Figure 2.

We compared the result of automatic annotation with manual annotation to detect errors made by the machine. By observing the statistics, a rule was derived to correct them. In most cases, errors first got rectified were those most amendable ones, i.e., errors which exhibited some evident patterns. Every time a rule was worked out, it was tried out on the whole testing corpus to see whether the resulting new F-Score was notably higher than that of last time. If it didn't, the rule was just ignored; otherwise it was adopted and applied. New errors might occur and this process repeated, until no rules could be manually found.

The underlying idea of this module is a bit like that of TBL (Transformation-Based Error-Driven Learning) [11]: It starts from an initial state, and by use of a series of transformation rules, it modifies the result bit by bit to achieve the best score according to the objective function used. It's only that in TBL, the transformations are learnt automatically (typically by greedy search algorithms); but in our solution, these rules are manually formulated to avoid the heavy computational cost.

For now a total of 15 rules were added.

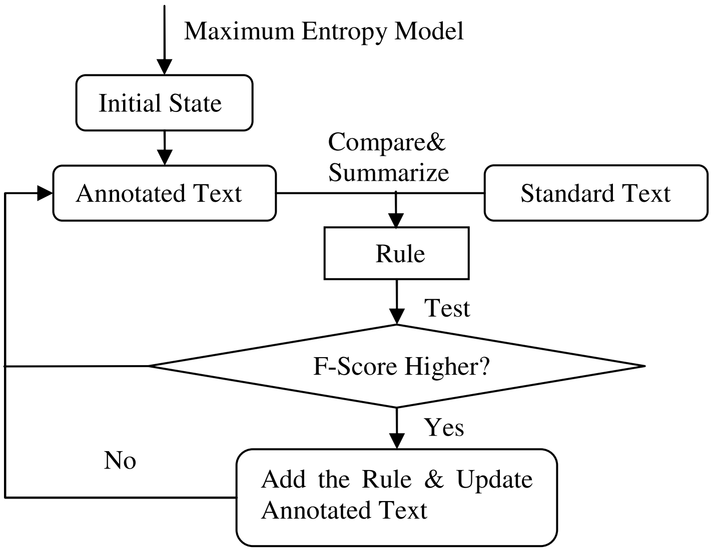

Fig. 2. Basic Process of Post-modification

4 Experiment and Result

4.1 Preparation

Our raw corpus comprises 10000 sentences randomly selected from People's Daily 2000. Each sentence had been segmented into a sequence of syntactic words with POS tags according to "Specification for Corpus Processing at Peking University: Word Segmentation, POS Tagging and Phonetic Notation" (shortened as "Specification 2003"). On average, one sentence contains 45.24 syllables and 25.15 syntactic words.

Then this corpus was prosodically annotated by two trained people, who were consistent on more than 90% of their annotation. Both prosodic words and prosodic phrase boundaries were marked out. The whole process was guided and supervised by an expert (Jianfen Cao of Chinese Academy of Social Sciences).

Among the 10000 sentences, 4500 were used for training and 2000 were used for testing in all of the following experiments. The two sets did not overlap each other.

4.2 Evaluation Criteria

Since subjective tests are time-consuming and costly to perform, we adopted an objective point of reference.

As mentioned earlier, there are altogether 3 prosodic boundary types between two syntactic words w_i and w_{i+1}: LW, PW and PPh. For simplicity in notation, we here refer to them as B_0, B_1 and B_2 respectively. To evaluate the performance of our system, the prosodic boundaries automatically assigned to the testing set were compared to human-annotation. In this way a confusion matrix was acquired, as shown in Table 2. In the table, C_{ij} denotes the number of boundaries manually labeled as B_i and predicted to be B_j.

Table 2. Confustion Maxtrix

Manually Labeled Type	Predicted Type		
	B_0	B_1	B_2
B_0	C_{00}	C_{01}	C_{02}
B_1	C_{10}	C_{11}	C_{12}
B_2	C_{20}	C_{21}	C_{22}

Our evaluation metric *Precision* and *Recall* were computed as:

For prosodic word boundary prediction:

$$\text{Precision}_1 = \frac{\sum_{j=1}^{2}\sum_{i=1}^{2} C_{ij}}{\sum_{j=0}^{2}\sum_{i=1}^{2} C_{ij}} \quad , \quad \text{Recall}_1 = \frac{\sum_{j=1}^{2}\sum_{i=1}^{2} C_{ij}}{\sum_{i=0}^{2}\sum_{j=1}^{2} C_{ij}} \; . \qquad (\mathbf{10})$$

For prosodic phrase boudnary prediction:

$$\text{Precision}_2 = \frac{C_{22}}{\sum_{j=0}^{2} C_{2j}} \quad , \quad \text{Recall}_2 = \frac{C_{22}}{\sum_{i=0}^{2} C_{i2}} \; . \qquad (\mathbf{11})$$

Another measurement F-Score takes both into consideration:

$$F\text{-Score}_i = \frac{2 \times \text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \quad (i = 1, 2) \; . \qquad (\mathbf{12})$$

4.3 Effect of Adding Rules

Figure 3 shows the test results for prosodic words segmentation when the number of rules gradually increased from 0 to 15. Since Maximum Entropy Model alone was able to achieve a relatively high accuracy with a large training corpus (4500 sentences in our case), the post-processing module doesn't seem to be playing a significant part.

Yet there is still notable rise in all three measures. When training material is of a small size and linguistic feature values are sparse, more remarkable improvement could be expected.

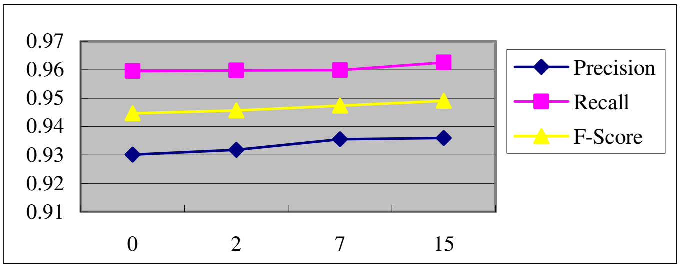

Fig. 3. Prosodic Word Boundary Prediction Result When Adding Rules

4.4 General Performance

Table 4 gives the testing result of our system, with threshold1=0.65 and threshold2= 0.7 for multi-pass prediction. Best results of some other approaches adopted in previous experiments are also listed for comparison.

Table 4. Best Test Result in Comparison with Other Methods

	Prosodic Word			**Prosodic Phrase**		
	Precision	Recall	F-Score	Precision	Recall	F-Score
C4.5 [7]	0.814	0.822	0.818	0.712	0.829	0.766
TBL [7]	0.782	0.848	0.814	0.853	0.613	0.713
ME Model[8]	N/A	N/A	N/A	N/A	N/A	0.652
Our Approach	0.936	0.963	**0.949**	0.798	0.784	**0.791**

Due to difference in the corpora and evaluation metric, these results may not be comparable in all respects. Yet from the statistics above, we could safely say that our approach is a successful attempt towards prosodic boudary prediction. .

5 Conclusions

In this paper, we addressed the problem of prosodic boudary prediction based on syntactic information. The Maximum Entropy Model was utilized with two separate instantiation for prosodic word and prosodic phrase segmentation. Our feature selection strategy was distinctive in that it not only drew on generally 'static' syntactic context, but also considered the interplay among successive boundary positions by incorporating 'dynamic' features into the model. It gave a satisfying performance especially for prosodic word prediction. For prosodic phrase prediction, even though lack of high-level syntactic and semantic information impeded its accurate prediction, a multi-pass procedure served to strike a better prosodic balance. Besides, the combination of machine learning power and human wisdom through error-driven ruled-based modification further enhanced its performance.

Future work should focus on extraction of 'deeper' contextual information such as sense group and semantic chunk to aid the perception of prosodic phrase boundaries. Moreover, the inherent uncertainty in prosody structure in natural language may require a more flexible approach to its prediction, possibly by using a minimum error-rate criterion (MERC) [12] in place of the traditional maximum correct-rate criterion currently adopted by us.

References

[1] Jianfen Cao: Prediction of Prosodic Organization Based on Grammatical Information. Journal of Chinese Information Processing, Vol. 17. (2003) 41–46

[2] Gee J.P., Grosjean F: Performance structures: A psycholinguistic and Linguistic Appraisal. Cognitive Psychology, Vol. 15. (1983) 411–458

[3] Jianfen Cao, Weibin Zhu: Syntactic and Lexical Constraint in Prosodic Segmentation and Grouping. In: Proceedings of Speech Prosody 2002. Aix-en-Provence, France (2002)

[4] Hongjun Wang: Prosodic words and prosodic phrases in Chinese. Chinese Language,Vol. 6. (2000) 525–536

[5] Wang M., Hirschberg J.: Predicting Intonational Boundaries Automatically from Text, In: the ATIS Domain Proceedings of the DARPA Speech and Natural Language Workshop, (199l) 378–383

[6] Paul Taylor, Alan. W. Black: Assigning phrase breaks from part-of speech sequences, Computer Speech and Language, Vol. 12(4). (1998) 99–117

[7] Z Sheng, T Jianhua, C Lianhong: Learning rules for Chinese prosodic phrase prediction. International Conference on Computational Linguistics, Proceeding of the first SIGHAN workshop on Chinese language processing, Vol. 18. (2002)

[8] Jian-Feng Li, Guo-Ping Hu, Renhua Wang: Chinese prosody phrase break prediction based on maximum entropy model, In: Interspeech 2004, Jeju Island, Korea. (2004) 729–732

[9] BERGER AL, STEPHEN A, DELLA PIETRA SA, et al: A maximum entropy approach to natural language processing [J]. Computational Linguistics, Vol. 22(1). (1996) 39–71

[10] Min Zheng, Lianhong Cai: Prosodic Constituents Segmentation and Syntax of Chinese: In: Proceeding of 5th Chinese Lexical Semantics Workshop, Singapore (2004)

[11] Brill, Eric: Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging. Computational Linguistics. Vol. 21(4). 543–565

[12] Min Chu: The Uncertainty in Prosody of Natural Speech and Its Application in Speech Synthesis, Journal of Chinese Information Processing, Vol. 18(4). (2004) 66–71