Prosodic Words Prediction from Lexicon Words with CRF and TBL Joint Method[1]

Heng Kang, Wenju Liu

National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences
{hkang, lwj}@nlpr.ia.ac.cn

Abstract. Predicting prosodic words boundaries will directly influence the naturalness of synthetic speech, because prosodic word is at the lowest level of prosody hierarchy. In this paper, a Chinese prosodic phrasing method based on CRF and TBL model is proposed. First a CRF model is trained to predict the prosodic words boundaries from lexicon words. After that we apply a TBL based error driven learning approach to refine the results. The experiments shows that this joint method performs much better than HMM.

Keywords: prosodic words, lexicon words, CRF, TBL

1 Introduction

In Chinese spoken language, the prosodic structure of an utterance can be viewed as three levels: prosodic word, prosodic phrase and intonation phrase[1]. The prosodic word (P-word) is defined as a group of syllables that are uttered closely and continuously in speech[2]. No boundary should be perceived within a prosodic word. That is to say, prosodic word is the basic prosodic structure in spoken Chinese. And in speech synthesis, experiments show the TTS system using the prosodic words as the basic prosodic units can get high naturalness than using lexicon words (L-word)[3]. Because the prosodic words influence the rhythm of synthetic utterances very much, it is of much importance to predict prosodic words from the input text.

Although lexicon words segmentation technology has become mature in Chinese language processing, it is still a difficult work on prediction prosodic words from unrestricted text. In all of previous work, Hidden Markov Model (HMM) based methods are adopted in many TTS systems because of the elegant methodology. Although most of them have taken full use of the feature information, the segmentation results are not good enough. The reason lies in the structure of HMM[4]. HMM is a probabilistic model of the way in which the data and labels are generated. This model has some drawbacks. Firstly the structure of HMM is often a poor model of the true process to produce data. Because of its first-order Markov

[1] This work is supported in part by the China National Nature Science Foundation (No. 60172055, No. 60121302), the Beijing Nature Science Foundation (No.4042025) and the National Fundamental Research Program (973) of China (2004CB318105).

property, any relationship between two labels must communicates via the intervening status, which cannot in general capture the relationships. Secondly is HMM generates each datum only from the corresponding status, which makes it difficult to utilize an input sliding window.

In this paper we proposed a prosodic words prediction method based on Conditional Random Field(CRF)[5] and Transformation Based Learning(TBL)[6,7] joint model. Firstly a CRF model is trained to predict the prosodic words boundaries from lexicon words. After that we apply a TBL based error driven learning approach to refine the results. The experiments shows that this joint method performs much better than HMM. In our project we apply this model on both prosodic words grouping and splitting.

This paper is organized as follows. Section 2 gives an introduction to the prosodic words. Section 3 describes our CRF model based method to predict prosodic words. Section 4 gives the description of the TBL method to refine the CRF predicting results. Experiments results will be presented in section 5. Finally we conclude our paper.

2 Prosodic Words

2.1 Prosodic Words and Lexicon Words

In Chinese, the hierarchy of prosody is not identical to that of syntax. The prosodic word is defined as a group of syllables that are uttered closely and continuously. While lexicon words is according to a lexicon. For example, in a Chinese sentence "他的帽子太大了(His hat is too large)" can be segmented to lexicon words "他/的/帽子/太/大/了". But in spoken speech the utterance should be segmented to prosodic words "他的/帽子/太大了". It can be seen that there is not a direct mapping between lexicon words and prosodic words. Statistical results show that only about half of lexicon words are prosodic words as well.

The prosodic words cannot be directly stored in a lexicon, because they will change greatly in different context. Therefore we cannot simply lookup in a lexicon to segment a sentence into prosodic words.

Studies show that most P-words consist of two characters, and very few P-words consist of 3 characters or above. This is due to the bi-character rhythm demand to build the prosodic foots in the phonology. Because of this reason, some one-character lexicon words should be grouped into one P-word, and long L-word is tended to be split into several P-words.

Our research is based on a corpus with 13000 sentences. For each sentence the prosodic words boundaries are manually labeled. The statistical results show that there are about 110,000 lexicon words total, and 81,000 prosodic words. Figure 1 illustrates the length distribution of P-words and L-words in this corpus. From this figure we can find that most P-words have 2 Chinese characters.

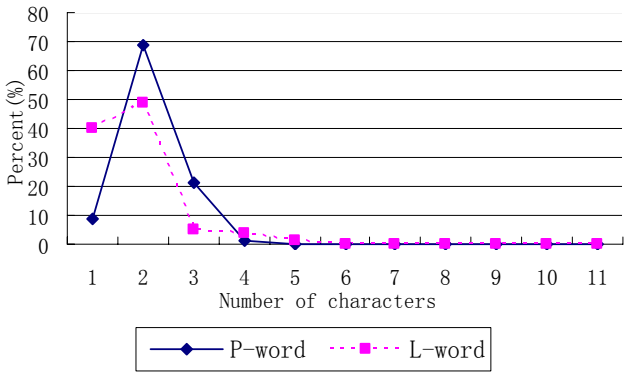

Fig. 1. Length distribution of P-words and L-words in the corpus

2.2 Prosodic Words Labeling

In order to predict prosodic words from lexicon words automatically, the training corpus should be labeled manually by a set of guidelines[2]. The labels include lexicon words boundaries, the POS(part of speech) of each lexicon words, and the prosodic words boundaries. To get high consistent labeling, only one annotator is asked to do this labeling work in our experiments.

We take the Chinese sentence "在这幸福的日子里我们歌唱祖国。" for example:
(1) 在这幸福的日子里我们歌唱祖国。
(2) 在/p 这/r 幸福/a 的/u 日子/n 里/f 我们/r 歌唱/v 祖国/n 。/w
(3) 在这| 幸福的| 日子里| 我们| 歌唱| 祖国。
(4) 在/p 这/r | 幸福/a 的/u | 日子/n 里/f | 我们/r | 歌唱/v | 祖国/n 。/w
where (1) is the original text. After POS tagging (2) is gotten, in which '/' means the boundaries and the symbols followed by '/' is the POS tag. (3) is labeled with the prosodic words, in which '|' means the prosodic words boundaries. For better usage of the boundaries and POS tag, (2) and (3) are combined to (4).

All the 13000 sentences in our corpus are labeled like this.

3 CRF Based Method to Predict Prosodic Words

HMM is an elegant and easy methodology that is adopted in many TTS system. It is a probabilistic model in which data and status are generated. However it suffer from some drawbacks. Firstly the structure of HMM is often a poor model of the true process to produce data. Because of its first-order Markov property, any relationship between two labels must communicates via the intervening status, which cannot in general capture the relationships. Secondly is HMM generates each datum only from the corresponding status, which makes it difficult to utilize an input sliding window.

Maximum Entropy Markov Models (MEMM)[8] attempt to maximize the conditional likelihood of data via a maximum entropy method. Although this model supports long-distance interactions, unfortunately it suffer from a label bias problem.

3.1 Introduction to CRF model

Conditional Random Fields are introduced to overcome these problems. CRFs are undirected graphical models that encode a conditional probability distribution with a given set of features. In the special case in which the designated output nodes of the graphical model are linked by edges in a linear chain, CRFs make a first-order Markov independence assumption among output nodes. Fig.2 shows the graphical structure of a chain-structured CRFs.

For sequential data $X = x_1 ... x_T$ and their corresponding labels (status) $Y = y_1 ... y_T$, a linear chain structure CRF defines the conditional probability as

$$P_\Lambda(Y \mid X) = \frac{1}{Z_X} \exp\left(\sum_{t=1}^{T} \sum_{k} \lambda_k f_k(y_{t-1}, y_t, x, t)\right) \tag{1}$$

where Z_x is the per-input normalization that makes the probability of all state sequences sum to one; $f_k(y_{t-1}, y_t, x, t)$ is a feature function which is often binary-valued, but can be real-valued, and λ_k is a learnt weight associated with feature f_k. The feature functions can measure any aspect of a state transition y and the entire observation sequence x centered at the current time step t. Large positive values for λ_k indicate a preference for such an event; large negative values make the event unlikely.

The model parameters f_k can be estimated by maximum likelihood—maximizing the conditional probability of a set of label sequences, each given their corresponding input sequences. The log-likelihood of the training set is

$$\begin{aligned} L_\Lambda &= \sum_i \log P_\Lambda(y_i \mid x_i) \\ &= \sum_i \left(\sum_{t=1}^{T} \sum_{k} \lambda_k f_k(y_{t-1}, y_t, x, t) - \log Z_{xi} \right) \end{aligned} \tag{2}$$

Traditional maximum entropy learning algorithms, such as GIS and IIS[9] can be used to train CRFs.

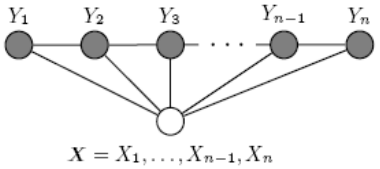

Fig. 2. Graphical structure of a chain-structured CRFs for sequences

For the given observation sequential data, the most probable label sequence can be determined by

$$y^* = \arg\max_{y} P_{\Lambda}(Y \mid X) \tag{3}$$

which can be efficiently determined using the Viterbi algorithm[10]. An N-best list of labeling sequences can also be obtained using modified Viterbi algorithm and A* search[11].

3.2 CRF for Prosodic Words Prediction

For automatically processing the labels by computers, the manually labeled data should be formatted as follows:

The sentence "在/p 这/r | 幸福/a 的/u | 日子/n 里/f | 我们/r | 歌唱/v | 祖国/n 。/w " is formatted to "在/p/B 这/r/E 幸福/a/B 的/u/E 日子/n/B 里/f/E 我们/r/S 歌唱/v/S 祖国/n/S 。/w/W", in which '/B' represents this lexicon is at the beginning of a prosodic word, '/E' means this lexicon word is at the end of a prosodic word, and '/I' means it is at the intermediate part.

After the labeling we can find that prosodic words prediction is a typical tagging problem which can be described as: given the observation sequence $X = x_1 \ldots x_T$, determine the corresponding labels $Y = y_1 y_2 \ldots y_N$. This can be solved directly by CRF.

To utilize the flexibility of CRF and considering the prosodic words prediction problem, we use the features in table 1.

In addition, the word length of the current prosodic words is used as the feature. For the lexicon word "幸福/a" in the last example, the feature for length of the prosodic words is f(PW-3B) = 1, which means this lexicon is at the beginning of a 3-characters-long prosodic words. Because most prosodic words have 2 or 3 Chinese characters, this type of feature is very import.

Table 1. Features used in CRF modeling

Features	Explanations
LW_{-2}	second previous lexicon word
LW_{-1}	previous lexicon word
LW_0	current lexicon word
LW_1	next lexicon word
LW_2	second next lexicon word
LW_0LW_1	Current lexicon word and next lexicon word
$LW_{-1}LW_0$	previous lexicon word and current lexicon word
$LW_{-2}LW_{-1}$	second previous lexicon word and previous lexicon word
$LW_{-1}LW_0LW_1$	previous lexicon word, current lexicon word and next lexicon word

4 Refining the Result with TBL

Although we use a very large corpus to train the CRF model, the sparseness problem still occurs because of the statistical method. For this reason we try to make use of TBL[6,7] to refine the prediction results.

TBL is an algorithm that can automatically get rules from a set of templates. Comparison with statistical methods like CRF, TBL is not so sensitive to the data sparseness.

Fig. 3 illustrates how to select rules with TBL. Before learning stage, the program compares the initial labels and the mannul labels results sentence by sentence. If they are totally the same the sentence will be skipped, otherwise, a candidate rule is generated from a template.

In the learning stage, the evaluation process will apply each rule in the candidate rules set. A score is given according to the number of errors that the rule can amend. The rule with the highest score is recorded, and the amended results by it will be saved as the initial status of next loop. This evaluation-application loop runs until no more errors can be corrected.

For our application, the configuration is as follows:

(1) initial labels: results from CRF prediction

(2) rule templates: templates should be designed to consider the prosodic words problem. In our application, the templates are very like the features used in CRF model:

If (LW_{-1}:POS=P1 & LW_{-1}: POS=P2 & LW_{-1}:LENGTH=L) then PTAG1 =>PTAG2

which means, if the POS of previous lexicon word is P1 and POS of current lexicon word is P2 and length of previous lexicon word is L, then the tag PTAG1 is corrected to PTAG2. PTAG is in the set {B, I, E, S, W}.

(3) evaluation: a score is given according to the number of errors that the rule can amend. A rule with the highest score is selected. A threshold should be set that only when the score is more than the threshold the rule could be recorded.

In our experiment,

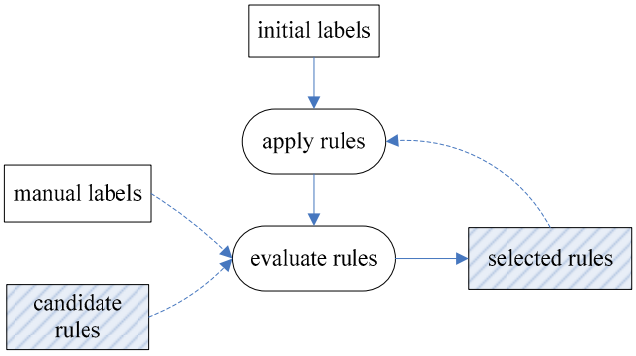

Fig. 3. TBL learning process

5 Experiments

For the Chinese prosody research, we collected and designed a large phonetically and prosodic enriched text corpus from different domains. In this large corpus there are about 13000 text sentences, which are labeled carefully by a well-trained researcher. We select 10000 sentences to train our model and the rest of them are for test. We also adopt HMM model based method for comparison.

In the training set, there are 120121 lexicon words and 90312 prosodic words. The longest lexicon word has 11 Chinese characters, and the longest prosodic word has 4 Chinese characters.

In our experiments, after training, we test on both training set and the test set. There are 2 evaluation criteria: precision and recall rate, which are defined as follows.

$$F_{pre} = \frac{N_1}{N_2} \times 100\% \tag{4}$$

$$F_{rec} = \frac{N_1}{N_3} \times 100\% \tag{5}$$

Where N_1 is the number of prosodic word boundaries predicted correctly, N_2 is the total number of prosodic word boundaries predicted, and N_3 is the total number of real prosodic word boundaries in the test set.

Table 2. Statistical results of the experiments

models \ results		10000 sentences close set		3000 sentences open set	
		precision	recall rate	precision	recall rate
No Model		59.07%	**96.71%**	59.65%	**96.54%**
CRF	CRF	90.52%	95.72%	90.12%	92.29%
	CRF + TBL	**95.87%**	95.90%	**93.22%**	94.44%
HMM	HMM	83.90%	94.77%	84.33%	94.52%

Table 2 illustrates our experimental results, in which “no model” method means the lexicon words boundaries is labeled as prosodic words boundaries directly. Apparently this method will get highest recall-rate and lowest precision.

From this table we can draw a conclusion that CRF model based method get high precision than HMM model based method, both in close set (training set) and open set (test set). And after applying TBL refinement, the precision and recall-rate increase more.

And we can also find that the precision and recall-rate don’t decrease much in the open set test. This indicates that the our method is robust and generalized.

6 Conclusion

In this paper, a Chinese prosodic phrasing method based on CRF and TBL model is proposed. First a CRF model is trained to predict the prosodic words boundaries from lexicon words. After that we apply a TBL based approach to refine the results. The experimental results show that this joint method performs much better than HMM.

Reference

1. Jianfen Cao, Acoustic phonetic features in the rhythm of Mandarin, The 4th national conference on modern phonetics, 1997
2. Min Chu, Yao Qian, Locating boundaries for prosodic constituents in unrestricted Mandarin texts, Computational linguistics and Chinese language processing, vol.6, no.1, 2002, pp.1-22
3. Yao Qian, Min Chu, Segmenting unrestricted Chinese text into prosodic words instead of lexicon words, Proceedings of the 2001 International conference on acoustic, speech and signal processing, 2001, Salt Lake City
4. Thomas G. Dietterich, Machine learning for sequential data: a review,
5. John Lafferty, Andrew McCallum, Fernando Pereiram, Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, Proceedings of the Eighteenth International Conference on Machine Learning, 2001
6. E. Brill, A rule-based approach to prepositional phrase attachment disambiguation, Proc. 15th international conference on computational linguistics, 1994, pp1198-1204
7. E. Brill, Automatic grammar induction and parsing free text: a transformation-based approach. Proc. Of the ARPA human language technology workshop, Princeton, N.J. 1993
8. McCallum, A., Freitag, D. & Pereira, F. Maximum Entropy Markov Models for Information Extraction and Segmentation, Proc. ICML 2000
9. S. della Pietra, V. della Pietra, and J. Lafferty, Inducing Features Of Random Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(4), 1995
10. L. Rabiner, A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, *Readings in Speech Recognition*, pages 267–296, 1990
11. R. Schwartz and Y. Chow, The N-best Algorithm: An Efficient and Exact Procedure for Finding the N most Likely Sentence Hypotheses, *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 1990*