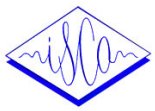

International Symposium on Chinese Spoken Language Processing (ISCSLP 2004)
Hong Kong
December 15-18, 2004

APPLY LENGTH DISTRIBUTION MODEL TO INTONATIONAL PHRASE PREDICTION

Jian-Feng Li, Guo-Ping Hu, Ming Fan, Li-Rong Dai

iFly Speech Lab, University of Science and Technology of China
Heifei, Anhui 230027
{lijianfeng, applecore}@ustc.edu

ABSTRACT

In this paper, a length distribution model for intonational phrase prediction is proposed. This model presents the probabilities that a certain length sentence is divided into some certain length intonational phrases. We will discuss how to estimate the probabilities in the model from training corpus, and how to apply it to intonational phrase prediction. We combine this model with a maximum entropy model which implements local context information. Experiment results show that length distribution is valuable information for intonational phrase prediction, and that it is able to make significant extra contribution over the maximum entropy model in terms of average score and unacceptable rate.

1. INTRODUCTION

In Chinese TTS systems, a widely used hierarchical prosody structure system consists of syllable, prosody word, intermediate phrase, intonational phrase and breath group[1] . For convenience sake, the 5 hierarchical layers are denoted by L0, L1, L2, L3 and L4. Among them, intonational phrase plays an important role on affecting the naturalness of synthesized speech. In this paper, we discuss about intonational phrase break prediction, which is to split a sentence into several intonational phrases, and also equals to decide whether a word boundary is an L3 break.

Recently, various kinds of statistic models were applied to this research field, including CART[1] (Classification And Regression Tree), Markov Model[3] , Maximum Entropy Model[5] , Memory Based Learning[4] and Artificial Neural Networks.

Although different statistic models were applied, similar information was exploited, including POS (Part-Of-Speech), Syllable number and the word itself in local context. The theory of machine learning tells us that there does not exist an algorithm who outperforms another at any case[6] . In practice, two learning models usually perform similarly if the same property set are used. Often, significant improvements can be achieved if new valuable properties are included.

Constrained by physical conditions, people is inclined to make an obvious pause (L3 break) after a certain number of syllables. Hence, we assume the length distribution of intonational phrase bears some statistic laws. We try to model the length distribution in this paper and investigate its contribution to intonational phrase break prediction.

The remainder of the paper is organized as following. Section 2 introduces the length distribution model. Section 3 simply describes the maximum entropy model. In section 4, the maximum entropy model and the length distribution model are combined into a mixed model, and in section 5, experiment settings and results are discussed.

2. LENGTH DISTRIBUTION MODEL

In this paper, the length distribution model of intonational phrase is defined as $P(l_1, l_2, \cdots l_n \mid L)$, where L is the length of the sentence, and l_i is the length of the i^{th} intonational phrase. Obviously, $L = \sum_i l_i$.

2.1. Probability estimation

Directly estimating the probabilities of length distribution from training corpus may cause data sparseness. For instance, given a sentence with a length of 17, there are altogether 2^{16} segmentation ways to split it into at least one, at most 17 intonational phrases. Due to a large number of potent segmentation ways for a sentence, great deal of training corpus is required to reliably estimate those probabilities in length distribution model.

In order to decrease the impact of data sparseness, three steps are adopted.

2.1.1. Model decomposition

According to probability theory, we have

$$P(l_1, l_2, \cdots, l_n \mid L) = P(n \mid L) \times P(l_1, l_2, \cdots, l_n \mid n, L) \quad (1)$$

The original length distribution model is decomposed into two sub-models: $P(n \mid L)$ and $P(l_1, l_2, \cdots, l_n \mid n, L)$,

which are denoted as sub-model-1 and sub-model-2 separately.

In sub-model-1, the probabilities are estimated directly from training corpus as follows:

$$P(n \mid L) = \frac{Count(n, L)}{Count(L)} \quad (2)$$

Comparing with the original length distribution model, probabilities with same number of intonational phrases are combined together into one probability in sub-model-1, which decrease the number of probabilities in model, and reduce the requirement for training corpus.

In sub-model-2, we relax the condition constraint of the probability $P(l_1, l_2, \cdots, l_n \mid n, L)$, not requiring the probabilities to be estimated only from the sentences with a length of L any more. In stead, they are permitted to estimate from all sentences, as long as the sentence contains some number of intonational phrases whose length summation equals to L. That is

$$P(l_1, l_2, \cdots, l_n \mid n, L) = \frac{Count(l_1, l_2, \cdots, l_n)}{\sum_{i_1, i_2, \cdots i_m; \sum_j l_{i_j} = L} Count(l_{i_1}, l_{i_2}, \cdots l_{i_m})} \quad (3)$$

2.1.2. L < 16

There are 2^{L-1} segmentation ways for a sentence with length of L. Larger L, more severe the data sparseness problem. So, we restrict L to be smaller than 16 in our model. In prediction process, for those sentences with length larger than 16, a specific algorithm is proposed to segment it into intonational phrases.

2.1.3. n < 4

Since L<16 is satisfied, it is nearly impossible that the number of intonational phrases n is larger than 4. So, we restrict n to be smaller than 4 in the model. For those $n \geq 4$, we set $P(n \mid L)$=0.

After the above steps, the problem of data sparseness can be efficiently reduced. Now, we are able to reliably estimate $P(n \mid L)$ and $P(l_1, l_2, \cdots, l_n \mid n, L)$ from a corpus of limited size.

3. MAXIMUM ENTROPY MODEL

Maximum entropy model is a probability model, which estimates probabilities based on the principle of making as few assumptions as possible, other than the constraints imposed. A constraint can be expressed by a binary feature function $fi(x,y)$, in which, x denotes the context, and y denotes the outcome. If some constraint is satisfied, $fi(x,y)$ is set to 1, otherwise 0.

A maximum entropy model can be represented as[7]:

$$p(y \mid x) = \frac{1}{Z(x)} \exp(\sum_i \lambda_i f_i(x, y)) \quad (4)$$

In which, λ_i is the weight of feature $fi(x,y)$, which can be estimated by IIS algorithms[7] . Z(x) is the normalization factor.

For more information about maximum entropy model, please refer to [7] .

4. MIXTED MODEL

Maximum entropy model only exploits the POS, syllable number and lexical information in the local context, excluding the length distribution information of intonational phrases. This may result in too short or too long intonational phrases. For instance, consider the following sentence.

“摆(0.00)上(0.00)了(0.58)越来越多(0.00)的(0.38)寻常(0.01)百姓家(0.00)的(0.18)餐桌”

The figures in the brackets indicate the probabilities that the boundary acts as an L3 break, which are generated by our maximum entropy model. Since the largest figure is 0.58, the prediction result of the maximum entropy model is “摆上了 # 越来越多的寻常百姓家的餐桌”, in which “#” indicates an L3 break. Obviously, the first intonational phrase is too short, while the second one is relatively too long.

In order to avoid this disadvantage, a mixed model is constructed that consists of the maximum entropy model and the length distribution model discussed in the previous two sections. The length distribution model is used as a post-process module which adjusts the result of maximum entropy model and presents the final prediction result.

4.1 For sentences with length L < 16

Given a sentence with a length of less than 16, the length distribution model is able to present a probability $P(l_1, l_2, \cdots l_n \mid L)$ for any segmentation way $(l_1, l_2, \cdots, l_n)$. Similarly, the maximum entropy model can also generate a likelihood probability $P(l_1, l_2, \cdots, l_n \mid ME)$. As it is known, one segmentation way $(l_1, l_2, \cdots, l_n)$ corresponds to a sequence of word boundary tags $(t_1, t_2, \cdots, t_N)$, $t_i \in \{L3, \overline{L3}\}$, and maximum entropy model is able to yield the probability $P(t_i \mid b_i)$ that each word boundary b_i is tagged as t_i. Hence, we have:

$$P(l_1, l_2, \cdots, l_n \mid ME) = \prod_{i=1}^{N} P(t_i \mid b_i) \quad (5)$$

where N is the number of word boundaries in the sentence.

In the mixed model, the probability for segmentation way $(l_1, l_2, \cdots, l_n)$ is computed as:

$$P(l_1, l_2, \cdots, l_n) = P(l_1, l_2, \cdots, l_n \mid ME) \times P(l_1, l_2, \cdots, l_n \mid L)^{\alpha} \quad (6)$$

In our experiment, α is set to 0.5 empirically.

Based on the above discussion, we choose the best segmentation way which has the maximum probability as the final prediction result.

4.2 For sentences with length L≥16

For sentences with length of larger than 16, the maximum entropy model can still yield the probability $P(l_1, l_2, \cdots, l_n \mid ME)$, but the length distribution model can not present $P(l_1, l_2, \cdots l_n \mid L)$ directly. We implement Algorithm 1 to solve the problem, which is showed in Figure 1.

Algorithm 1:

1) Set B to be the beginning of the sentence;
2) Set E to be the word boundary with maximum L3 break probability in the scope of 10 to 16 characters after B;
3) For the sentence part between B and E, search for the best segmentation which has the maximum probability;
4) If the best segmentation contains 1 phrase, reserve it;
5) If the best segmentation contains k (k>1) phrases, reserve the first k-1 phrases, and combine the last one with unprocessed sentence part.
6) Set B to be the head of unprocessed sentence part, and repeat step 2) to 6), until the whole sentence is processed.

Figure 1: Prediction algorithm for long sentences

We illustrate Algorithm 1 with the example "李沛瑶充分肯定我省在农村转移工作中取得的成绩". In the first iteration, the under-processed sentence part is "李沛瑶充分肯定我省在农村", which is split as "李沛瑶充分肯定 # 我省在农村" at step 3). Then, at step 5), the second phrase is merged with the un-processed part into "我省在农村转移工作中取得的成绩", which will be segmented as "我省在农村转移工作中 # 取得的成绩" in the second iteration. In the end, the final result is "李沛瑶充分肯定 # 我省在农村转移工作中 # 取得的成绩".

The computation complexity of step 3) grows exponentially with the length of under-processed sentence part. This is another reason which makes it necessary in Algorithm 1 to split long sentences into short sentence parts.

5. EXPERIMENTS

5.1. Experiment settings

5.1.1. Corpus

20,000 sentences were random selected from People's Daily, and used in the experiments. Word segmentation, POS tagging and person name recognition were carried out by a preprocessing program. The accuracy of word segmentation is 96% and the accuracy of POS tagging is 91%.

Tags need be labeled at each word boundary to indicate L3 breaks or non L3 breaks. Lack of the corresponding speech, the annotators labeled word boundaries by reading the sentences themselves. As it is known, different people might label the same sentence differently. Through testing, the labeling consistency among the four annotators was 75%, which is the upper limit for automatic prediction.

All of the sentences were divided into two parts, 1000 sentence for testing and the others for training.

5.1.2. Evaluation metric

We utilize the F-Score as the evaluation metric in the experiments, which is defined as follows:

$$precision = \frac{number\ of\ correctly\ identified\ breaks}{number\ of\ identified\ breaks}$$

$$recall = \frac{number\ of\ correctly\ identified\ breaks}{number\ of\ correct\ breaks\ in\ test\ set} \quad (7)$$

$$f-score = \frac{2 \times precision \times recall}{precision + recall}$$

As discussed above, there may be more than one correct label results for one sentence, so the F-Score is inclined to over-estimate the error rate. In order to reflect the true error rate, we have the automatic labeled sentences scored by human. A 5-grade scoring system is applied, and those sentences scored under 3 are considered as unacceptable ones. After human scoring, average score and unacceptable rate are computed. Average score is the arithmetic average of all sentences, and unacceptable rate is the percentage of unacceptable sentences in all ones.

5.2. Experiment results

5.2.1. In terms of F-Score

Following [5] , a maximum entropy model was trained from the training corpus containing 19,000 sentences, and tested on the test set. We got an F-Score of 66.2%.

The length distribution model is also trained from the training corpus. Combining the maximum entropy model and the length distribution model as section 4 discussed, and testing the mixed model on the test set, we obtained an F-Score of 66.9%.

Table 1 shows the F-Score of human labeling consistency, maximum entropy model and mixed model. The relative F-Score compared to human labeling consistency are also listed in it. Although the value of F-Score is not high, they approach to 90% relative to human labeling consistency. We list the relative improvement of mixed model over maximum entropy model in Table 2.

	F-Score	Relative to human labeling consistency
Human labeling consistency	75.0%	100%
Maximum Entropy model	66.2%	88.3%
Mixed model	66.9%	89.2%

Table 1: F-Score of different model

	Maximum entropy	Mixed Model	Relative Improvement
F-Score	66.2%	66.9%	1.1%
Average score	4.0	4.3	7.5%
Unacceptable rate	1.8%	0.9%	50.0%

Table 2: Performance comparison of maximum entropy model and mixed model

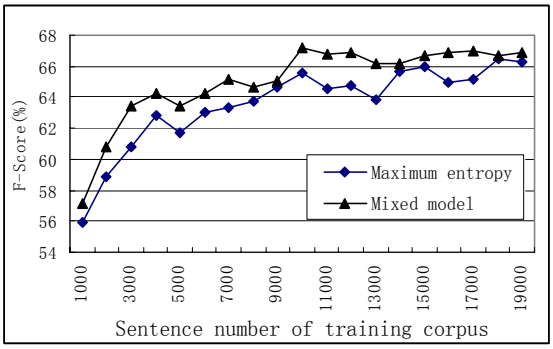

Figure 2: Performance curve of maximum entropy model and mixed model

5.2.2. In terms of average score and unacceptable rate

We also compared the performance of maximum entropy model and mixed model in terms of average score and unacceptable rate, which is listed in Table 2. Table 2 shows that, the mixed model makes an improvement of 1.1% in terms of F-Score, 7.5% in terms of average score, and 50.0% in terms of unacceptable rate. It is obvious that the improvement of F-Score is little while the improvement of average score and unacceptable rate is great. So, main contribution of length distribution model is to make the prediction result more acceptable by human, not more accurately matching the pre-defined standard labeling.

In order to investigate the performance variation of the two models along with the training corpus size, we trained them from different size of corpus. Two performance curves are drawn in Figure 2.

Figure 2 illustrates that the mixed model outperforms maximum entropy model consistently. Averagely, at each point, about 1.5 percent is achieved.

6. CONCLUSION

In this paper, we proposed a length distribution model. This model is to exploit the length distribution information which is assumed valuable for intonational phrase prediction. The length distribution model is used as a post-process module of a maximum entropy model, and experiment results proved that it contributes to improve average score and to reduce unacceptable rate.

The length distribution information is combined with local context information by equation (6), where α is a weight to balance the influence of the length distribution information and local context information. In this paper, α is not optimized through thorough experiments. Exploring the optimized α is the future work.

In this paper, length distribution information and local context information are exploited in different models, which make it unnatural to integrate them together. We will try to work out a uniform model to integrate all the information later.

References

[1] M. Chu, Y. Qian, “Locating Boundaries for Prosodic Constituents in Unrestricted Mandarin Texts”. *Computational Linguistics and Chinese Language Processing*, February 2001, Vol.6, No.1: 61-82.

[2] J. Hirschberg, P. Prieto. “Training intonational phrasing rules automatically for English and Spanish text-to-speech”. *Speech Communication*, 1996.

[3] NIE Xin, WANG Zuo-ying. “Automatic Phrase Break Prediction in Chinese Sentences”. *Journal of Chinese information Processing*, 2003, 17(4):39-44.

[4] G. J. Busser, W. Daelemans, Van den Bosch, A. “Predicting phrase breaks with memory-based learning”. *Proceedings 4th ISCA Tutorial and Research Workshop on Speech Synthesis*, Perthshire Scotland, August 29th - September 1st, 2001.

[5] Jian-feng Li, Guo-ping Hu, Wan-ping Zhang, Ren-hua Wang. “Chinese Prosody Phrase Break Prediction Based on Maximum Entropy Model”. *International Conference on Spoken Language Processing (ICSLP 2004)*. Oct 4-6, Korea.

[6] Richard O. Duda, Peter E. Hart, David G. Stork. “Pattern Classification”.

[7] Adam L. Berger, Stephen A. Della Pietra, Vincent J. Della Pietra. “A maximum entropy approach to natural language processing”. *Computational Linguistics* 1996, 23(4): 597-618.