A Pair-based Language Model for the Robust Lexical Analysis in Chinese Text-to-Speech Synthesis

Wu Liu[1], *Dezhi Huang*[1], *Yuan Dong*[1,2], *Xinnian Mao*[1], *Haila Wang*[1]

[1]Speech and Natural Language Processing Unit, France Telecom R&D Beijing, China
[2]Beijing University of Posts and Telecommunications

{wu.liu,dezhi.huang,xinnian.mao,haila.wang}@orange-ftgroup.com, yuandong@bupt.edu.cn

10.21437/Interspeech.2007-529

Abstract

This paper presents a robust method of lexical analysis for Chinese text-to-speech (TTS) synthesis using a pair-based Language Model (LM). The traditional way of Chinese lexical analysis simply regards the word segmentation and part-of-speech (POS) tagging as two separated phases. Each of them utilizes its own algorithms and models. Actually, the POS information is useful for word segmentation, and vice versa. Therefore, a pair-based language model is proposed to integrate basic word segmentation, POS tagging and named entity (NE) identification into a unified framework. The objective evaluation indicates that the proposed method achieves the top-level performance, and confirms its effectiveness in Chinese lexical analysis.

Index Terms: Chinese text-to-speech, Pair-based LM, Chinese word segmentation, POS tagging, Named entity identification

1. Introduction

Text analysis for text-to-speech (TTS) synthesis aims to convert a text of orthographic characters into a linguistic representation for speech synthesis. In Chinese, text analysis [1] consists of five sub-tasks, namely word segmentation [2], part-of-speech (POS) tagging, text normalization, grapheme-to-phoneme conversion and prosodic phrasing. This paper focuses on the first two tasks. Robust lexical analysis, including Chinese word segmentation and POS tagging, is essential to develop a high-quality TTS system for unrestricted applications. The past decades have witnessed the fast and great development of Chinese TTS system. However, lexical analysis capability, particularly the capability in processing real text is still rudimentary in a sense. When applying these lexical analysis methods into unrestricted applications, the quality of synthetic speech degrades sharply. Therefore, it is an urgent task for Chinese text-to-speech synthesis.

Generally, Chinese lexical analysis consists of two phases: Chinese word segmentation and POS tagging. The traditional way of Chinese lexical analysis simply regards the word segmentation and POS tagging as two separated phases. Each of them utilizes its own algorithms and models. The separation of Chinese word segmentation and POS tagging can lower the complexity of the system design, but it also decrease the performance as well, because these two phases are fully integrated when a human processing a sentence. Fortunately, many researchers have already noticed it and pay more attention to the integration of word segmentation and POS tagging, such as tri-gram integrated model [3] and hierarchical hidden Markov model [4], and so on.

Actually, Chinese word segmentation can be divided into Chinese basic word segmentation and Chinese named entity identification (including person names, location names and organization names). The tri-gram integrated model [3] is actually an integrated lexical analysis method, but it doesn't involve named entity identification. The hierarchical hidden Markov model [4] can identify the basic words and named entity words in a hierarchical model, but it is basically a sequential model.

In this paper, we present a robust lexical analysis method for Chinese TTS. In converting a real orthographic Chinese text into a linguistic representation for speech synthesis, three problems usually arise, namely basic word segmentation, POS tagging and named entity identification. These problems interact and cannot be resolved separately. When word segmenting, the POS information is useful for discriminating the word boundary, and vice versa. If separating word segmentation and POS tagging, the fault of word segmentation maybe cause the fault of POS tagging. This paper proposes a pair-based language model to solve the three issues. The model not only integrates the three problems into a unified framework, but also uses the information of each other to discriminate easily. When the method is applied to Chinese lexical analysis, the F-score (detailed in Section 4) of POS tagging of the method is higher 1% than that of separating model.

The remainder of this paper is organized as follows: Section 2 introduces pair-based language model. Section 3 details our Chinese lexical analysis frameworks. Section 4 reports the evaluation results of our system on public corpora. Our conclusion remarks are detailed in Section 5.

2. Pair-based LM to Chinese Lexical Analysis

A class-based LM [5] is a stochastic model which predicts a class given the strings by estimating the conditional probability $p(C|S)$, which can integrate Chinese word segmentation and NE identification into a unified framework. A pair-based LM extends the class-based LM by defining similar classes as class-POS pairs. Actually, the POS information is introduced into the class-based LM. It will be demonstrated to be a more effective way of robust lexical analysis in our evaluation section. In this paper, the pair-based LM is applied to integrate Chinese word segmentation, POS tagging and NE identification in a unified framework.

2.1. Definition of class-POS Pairs

In this paper, each kind of class-POS is defined as a pair, where each kind of named entity (including person names, location names and organization names) is defined as a class. In practice, in order to represent different constructions for each kind of class-POS, we further divide each pair into sub-pairs. The detailed definitions of the pairs are shown in Table 1. In addition, each word with one POS in a lexicon is defined as a pair.

Table 1: *Definition of class-POS pairs.*

class-POS pairs		Description	Examples
(PN,nr)	(PN2,nr)	Surname+one-character given name with ”nr” pos	李 鹏/nr: Li Peng/nr
	(PN3,nr)	Surname+two-character given name with ”nr” pos	江 泽 民/nr: Jiang Zemin/nr
	(FPN,nr)	Foreign Person Name with ”nr” pos	布 什/nr: bush/nr
(LN,ns)	(CLN,ns)	Chinese Location name with ”ns” pos	北 京/ns: Beijing/ns
	(FLN,ns)	Foreign Location Name with ”ns” pos	巴 黎/ns: Paris/ns
(ON,nt)	(CON,nt)	Chinese Organization Name with ”nt” pos	北 大/nt: Peking University/nt
	(FON,nt)	Foreign Organization Name with ”nt” pos	欧 佩 克/nt: OPEC/nt
(BLW,T)		Basic lexicon word with pos	开 心/adj: happy/adj
Total		|P|+3	

The class-POS pairs used in our model are shown in Table 1, where |P| means the number of class-POS pairs saved in the lexicon. Supposed that the size of the lexicon is |V|, each word with three POS, so the |P| is equal to the 3∗|V|. The ”T” in (BLW, T) represents the POS of each word in the lexicon.

2.2. Pair-based LM for Chinese Lexical Analysis

Given a sequence of Chinese string $S_1^n = s_1, s_2, ..., s_n$, the goal of Chinese lexical analysis is to find the optimal sequence pairs $(C_1^m, T_1^m) = (c_1, t_1), (c_2, t_2), ..., (c_m, t_m)(m <= n)$ that maximizes the probability $P((C_1^m, T_1^m)|S_1^n)$, where C_1^m is the class sequence and T_1^m is the POS sequence. We use Bayes rule to rewrite $P((C_1^m, T_1^m)|S_1^n)$ as equation (1):

$$P((C_1^m, T_1^m)|S_1^n) = \frac{P(C_1^m, T_1^m) \times P(S_1^n|(C_1^m, T_1^m))}{P(S_1^n)} \quad (1)$$

The $P(S_1^n)$ doesn't affect the value, so

$$P((C_1^m, T_1^m)|S_1^n) \propto P(C_1^m, T_1^m) \times P(S_1^n|(C_1^m, T_1^m)) \quad (2)$$

The pair-based statistical model consists of two components: the contextual language model $P(C_1^m, T_1^m)$ and the local language model $P(S_1^n|(C_1^m, T_1^m))$.

2.2.1. Contextual Language Model

The contextual language model $P(C_1^m, T_1^m)$ estimates the generative probability of a pair. The classical HMM assumes that the conversion from one state (that means class here) to another is not effected by the current observation value (i.e. the current POS), and the generation of current observation value is independent on other observation value. That is:

$$\begin{aligned} &P(C_1^m, T_1^m) = P(C_1^m)P(T_1^m|C_1^m) \\ &\approx \prod_{i=1}^{m} P(c_i|c_1, c_2, ..., c_{i-1}) \prod_{i=1}^{m} P(t_i|c_1, c_2, ..., c_m) \end{aligned} \quad (3)$$

Furthermore, only N previous states are considered when the current state is generated. And only the current state is involved when the current POS is generated.

$$\begin{aligned} &P(C_1^m, T_1^m) = P(C_1^m)P(T_1^m|C_1^m) \\ &\approx \prod_{i=1}^{m} P(c_i|c_{i-N}, c_{i-N+1}, ..., c_{i-1}) \prod_{i=1}^{m} P(t_i|c_i) \end{aligned} \quad (4)$$

That is the so-called N-order model or the (N+1)-gram model. In practice, we often use bi-gram or tri-gram model to mitigate data sparseness.

From the deduction above, we can see that the classical HMM has made an assumption that does not accord with the characteristic of Chinese. Actually, there is tight association between classes and POS tags in Chinese text, so we use the method below:

$$\begin{aligned} &P(C_1^m, T_1^m) \\ &= \prod_{i=1}^{m} P((c_i, t_i)|(c_1, t_1), (c_2, t_2), ..., (c_{i-1}, t_{i-1})) \\ &\approx \prod_{i=1}^{m} P((c_i, t_i)|(c_{i-1}, t_{i-1})) \end{aligned} \quad (5)$$

By this transformation, we have broken down the HMM's assumption, and introduced lexical information into Chinese lexical analysis to strengthen its discriminative ability.

2.2.2. Local Language Model

The local language model estimates the generative probability of a Chinese character sequence given a class-POS pair. The $P(S_1^n|(C_1^m, T_1^m))$ can be approximated using the equation (6):

$$\begin{aligned} &P(S_1^n|(C_1^m, T_1^m)) \\ &= P((s_1, s_2, ..., s_n)|((c_1, t_1), (c_2, t_2), ..., (c_m, t_m))) \\ &\approx P(([s_1, ..., s_{c_1-end}], ..., [s_{c_m-start}, ..., s_n])| \\ &\quad ((c_1, t_1), (c_2, t_2), ..., (c_m, t_m))) \\ &\approx \prod_{j=1}^{m} P([s_{c_j-start}, ..., s_{c_j-end}]|(c_j, t_j)) \end{aligned} \quad (6)$$

For the pair of (PN, nr):

$$\begin{aligned} &P([s_{c_j-start}, ..., s_{c_j-end}]|(c_j, t_j)) \\ &\approx \frac{C([s_{c_j-start}, ..., s_{c_j-end}], c_j, t_j)}{C(c_j, t_j)} \end{aligned} \quad (7)$$

Because the class PN and the POS "nr" appear simultaneously, we can rewrite the equation (7) as:

$$\begin{aligned} & P([s_{c_j-start}, ..., s_{c_j-end}]|(c_j, t_j)) \\ & \approx \frac{C([s_{c_j-start}, ..., s_{c_j-end}], c_j, t_j)}{C(c_j, t_j)} \\ & \approx \frac{C([s_{c_j-start}, ..., s_{c_j-end}], c_j)}{C(c_j)} \\ & \approx P([s_{c_j-start}, ..., s_{c_j-end}]|c_j) \end{aligned} \tag{8}$$

The same to the pairs of (LN, ns) and (ON, nt).

For the other pairs of (c_j, t_j), the c_j is the word in a lexicon, so we can rewrite the equation (6) as:

$$\begin{aligned} & P([s_{c_j-start,c_j-end}]|(c_j, t_j)) \\ & \approx \frac{C([s_{c_j-start}, ..., s_{c_j-end}], [s_{c_j-start}, ..., s_{c_j-end}], t_j)}{C([s_{c_j-start}, ..., s_{c_j-end}], t_j)} \\ & = 1 \end{aligned} \tag{9}$$

Therefore, the local language model is the same as the entity language model detailed in [5]. According to the analysis above, we can divide pair-based LM into two specified parts: contextual language model (represented by equation (5)) and local language model (represented by equation (8)).

2.3. Data Smoothing

If we introduce lexical information into pair-based LM, it will make data sparseness problem more serious. It is found that many class-POS pairs in test set do not occur in training set. So we must adopt data smoothing method. From equation (5), we can obtain:

$$\begin{aligned} & P((c_i, t_i)|(c_{i-1}, t_{i-1})) \\ & \approx P(t_i|(t_{i-1}, c_{i-1}))P(c_i|(t_{i-1}, c_{i-1}, t_i)) \end{aligned} \tag{10}$$

Or

$$\begin{aligned} & P((c_i, t_i)|(c_{i-1}, t_{i-1})) \\ & \approx P(c_i|(c_{i-1}, t_{i-1}))P(t_i|(c_{i-1}, t_{i-1}, c_i)) \end{aligned} \tag{11}$$

Now, we don't know which one representation is the better. The two equations will be compared in the evaluation section.

3. Framework of Pair-based Lexical Analysis

As illustrated in Figure 1, pair-based Chinese lexical analysis consists of three steps: class with POS candidates, named entity with POS candidates and Viterbi decoding.

Given a sequence of Chinese string, the decoding process is as follows:

- Step1: Lexical class-POS pair candidate generation. All possible class-POS pairs are generated according to a Chinese lexicon.
- Step2: NE-POS pair candidate generation. NE-POS candidates are generated in two steps: (1) NE-POS candidates are generated; (2) each candidate is assigned a probability by using the corresponding local language model probability.
- Step3: Viterbi Search. The N-best searching method is used here for decoding.

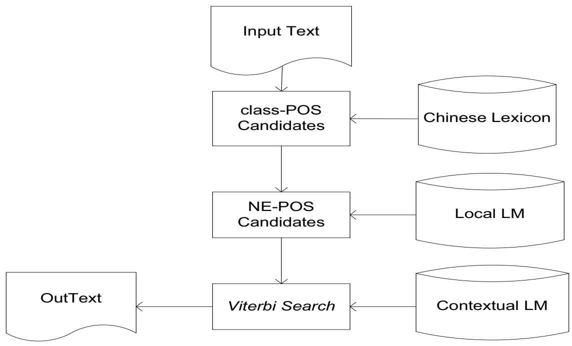

Figure 1: Pair-based Chinese lexical analysis.

4. Evaluation and Discussion

4.1. Evaluation Measures

We conducted evaluations in terms of the precision (P) rate, recall (R) rate and F-score:

$$P = \frac{Count(correctly\ identified\ WP)}{Count(identified\ WP)} \tag{12}$$

$$R = \frac{Count(correctly\ identified\ WP)}{Count(standard\ WP)} \tag{13}$$

$$F = \frac{2 \times P \times R}{P + R} \tag{14}$$

Where the "WP" can represent word or POS.

4.2. Experimental Results

In this experiment, we use the 2000 People's Daily corpus to train our proposed language model. Meanwhile we use the first month corpus of 1998 People's Daily to test. The used POS set is the first stage set defined by [6].

In the previous experiments, we have compared many probabilistic models for Chinese lexical analysis, and found that the system [7] participated in SIGHAN 2006 (rank 1st in the open track of MSRA test set) as segmentation strategy and forward tri-gram Markov Model as POS tagging model (SIG+FTMM) reaches the best performance. Our comparisons will be done between the SIG+FTMM and the pair-based language model.

The following tables present the results of our experiments. In the section 2.3, we use data smoothing method to obtain two equations, and we will obtain the better one through experiments. For comparison, we also rewrite the contextual model into the equation (15):

$$\begin{aligned} & P(C_1^m, T_1^m) = P(C_1^m)P(T_1^m) \\ & \approx \prod_{i=1}^{m} P(c_i|c_{i-2}, c_{i-1})P(t_i|t_{i-2}, t_{i-1})) \end{aligned} \tag{15}$$

Where we assume that the class and POS are completely independent.

In the Table 2 and Table 3:

- ICLM: Independent contextual language model, a very simple strategy in Chinese lexical analysis which assumes it is independent between the class and POS represented by equation (15).

Table 2: *The performance of Word Segmentation.*

WORD	R(%)	P(%)	F(%)
ICLM	97.2	97.5	97.3
DLM	97.0	96.7	96.8
DCLM	97.4	97.6	97.5
SIG+FTMM	97.5	97.7	97.6

Table 3: *The performance of POS tagging.*

POS	R(%)	P(%)	F(%)
ICLM	85.6	85.9	85.7
DLM	95.9	95.6	95.7
DCLM	96.5	96.7	96.6
SIG+FTMM	95.5	95.8	95.6

- DLM: Dependent contextual language model, which is represented by equation (11).
- DCLM: Dependent converse contextual language model, which is represented by equation (10).
- SIG+FTMM: A strategy of Chinese lexical analysis. The process is divided into two steps, firstly word segmentation, then POS tagging.

According to the results above (Table 2 and Table 3), the DCLM model does better than the above two methods in every field. Therefore, we can know the equation (10) is the better representation of contextual language model. We can also find that these word segmentation performance of DCLM and SIG+FTMM are basically the same, but the POS tagging performance of DCLM is better than that of SIG+FTMM. The DCLM combines Chinese word segmentation, POS tagging and named entity identification in a unified way, but the SIG+FTMM implement Chinese word segmentation and POS tagging separately. It further demonstrates that the word and POS information is closely dependent. In a word, these experiments confirm that introducing the pair-based language model into a Chinese lexical analysis system is effective.

Class-POS pair candidate need output all possible class-POS pairs, so there maybe hundreds of class-POS pairs in one place. It will cause the heavy computational workloads. In our lexical analysis method, we use the N-best Viterbi algorithm to search the optimal sequence pairs. From Figure 2 and Figure 3, we can find that the influence of the N-best algorithm to POS tagging is more than that of word segmentation. In Figure 2, we can see the performance of word segmentation raised fixed point from the 5-best, but in Figure 3, the fixed point is raised in 10-best. It is clear that the performance of POS tagging achieves great improvement from 1-best to 10-best.

5. Conclusions

In this paper, we present a hybrid method of incorporating Chinese word segmentation, POS tagging and named entity identification into pair-based language model. We only use one enhanced language model to implement Chinese lexical analysis, but the method can achieve better performance. Through our experiments, it can denote that the word and POS information are closely dependent in Chinese lexical analysis. The separation of word segmentation and POS tagging is unreasonable. The model can also be used in other language lexical analysis. Applying this model to other language, the corpus need be annotated and class-POS pairs need be redefined, which is dependent on language.

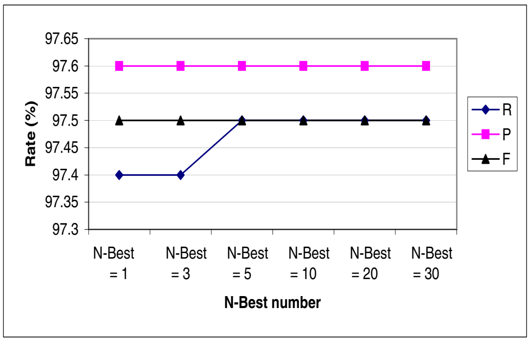

Figure 2: Word segmentation with N-best.

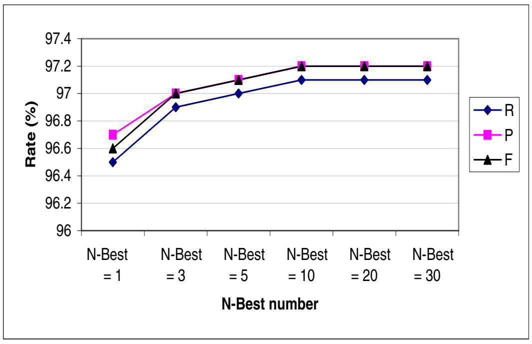

Figure 3: POS tagging with N-best.

6. References

[1] Guohong Fu, Min Zhang, GuoDong Zhou, Kang-Kuong Luke, "A Unified Framework for Text Analysis in Chinese TTS", the 5th International Symposium on Chinese Spoken Language Processing, vol. 1, pp. 200-210, 2006.

[2] Shih Chilin and Richard Sproat, "Issues in text-to-speech conversion for Mandarin", Computational Linguistics and Chinese Language Processing, vol. 1, no. 1, pp. 37-86, 1996.

[3] Shan Gao and Yan Zhang, "The Research on Integrated Chinese Word Segmentation and Labeling based on trigram statistical model", Natural Language Uderstanding & Machine Translation (JSCL-2001), 2001. (In Chinese)

[4] Huaping Zhang, Qun Liu, etc., "Chinese lexical analysis using hierarchical hidden Markov model", 2nd SIGHAN workshop affiliated with 41th ACL, pp. 63-70, 2003.

[5] Jian Sun, Jianfeng Gao, Lei Zhang, Ming Zhou, Changning Huang, "Chinese Named Entity Identification Using Class-based Language Model", Proceedings of the 19th International Conference on Computational Linguistics, pp. 967-973, 2002.

[6] Shiwen Yu, etc., "Specification for Corpus Processing at Peking University:Word Segmentation, POS Tagging and Phonetic Notation", Journal of Chinese Language and Computing, 13(2), pp. 121-158, 2003.

[7] Wu Liu, etc., "France Telecom R&D Beijing Word Segmenter for Sighan Bakeoff 2006", Proceedings of the Fifth SIGHAN workshop on Chinese Language Processing, pp. 122-125, 2006.