The Effect of Part-of-speech on Mandarin Speech Recognition

Caixia Gong, Xiangang Li and Xihong Wu
Speech and Hearing Research Center,
Key Laboratory of Machine Perception (Ministry of Education),
Peking University, 100871, Beijing, China
{gongcx, lixg, wxh}@cis.pku.edu.cn

***Abstract*—This paper** concentrates on **the effect of part-of-speech on Mandarin speech recognition by incorporating it into language model and pronunciation dictionary. This work is motivated by the two benefits of part-of-speech, one is to reduce the lexical ambiguity in language model to some extent and the other is to provide some information about the pronunciation of heteronyms. The experiments conducted on two corpora, tagged manually or automatically, show that a 3% relative character error rate (CER) reduction is achieved. Moreover, we find that performance improvement is mainly due to the relationship between part-of-speech and pronunciation of heteronyms.**

I. Introduction

Although a great deal of progress has been made during the last two decades, the performance of automatic speech recognition (ASR) systems still lags far behind human level performance. In almost all of the state-of-art ASR systems, the language information is modeled by n-gram, and acoustic parameters are modeled by Hidden Markov Models. In the human speech recognition, multi-level cues, including lexical, syntactic and semantic cues, are integrated together.

In order to bridge the gap of performance between human and ASR, many researchers devote to introducing kinds of rich language information into current speech recognition systems. In [1], the structured language model is proposed, which uses lexical and syntactic information to improve the performance of ASR. Besides, to introducing the semantic information, [2] adopts the maximum entropy technique, which has inspired many relevant efforts on language modeling. Moreover, nonlocal, syntactic and n-gram dependencies are combined for language modeling in [3].

However, the effects of introducing these multi-level linguistic cues into ASR systems have not been studied and discussed in detail. The state-of-the-art ASR systems are usually decomposed into three components: an acoustic model, a pronunciation dictionary and a language model. Each component has its own missions, where the acoustic model focuses on how to reduce the confuse of acoustic observation among each modeling unit, the pronunciation dictionary must deal with the pronunciation confusability of words, and the language model handle the uncertainty of words sequence or the complex language phenomena. The performance improvement comes from the confusion reduction of these three aspects. More specifically, the introduction of multi-level linguistic information may directly results in the change of pronunciation dictionary and language model.

Many questions remain to figure out where the performance improvement comes from: the effect of multi-level linguistic information on the pronunciation dictionary or language model, which effect plays a leading role, and how to get further more improvement. To answer these questions, this paper investigates the effects of part-of-speech for Mandarin speech recognition carefully.

The idea of using part-of-speech information into speech recognition has been proposed by many researches. These approaches can be divided into two categories. The first category introduces part-of-speech to model unseen events, such as Class-based LM [4] and its various variants [5-6]. By contrast, the second uses part-of-speech information to distinguish a word in different contexts [7-8]. These methods represent word as bundles of features which can include, for example, morphological, syntactic and semantic information.

The first category of methods takes only consideration on the effect of part-of-speech on language model while ignores the other components of ASR. In fact, parts-of-speech not only give large amount of information about a word and its neighbors, but also provide some information about its pronunciation. Methods of the second category make better use of the effect of part-of-speech on word's pronunciation. In this paper, similar to [8], the pair of the word and part-of-speech is considered as the basic unit in language modeling and pronunciation dictionary. In addition, to analyze the effects of part-of-speech on the language model and pronunciation respectively, we classify the multi-category words based on the amount of information about the pronunciations of these word that part-of-speech can provide, i.e. we can know the pronunciation totally or partially when knowing its part-of-speech. And then a series of experiments are conducted. The experimental results give some suggestions for introducing multi-level linguistic cues into ASR systems.

The rest of the paper is organized as follows. In Section II, we review the method of using part-of-speech information in an ASR system. Section III gives experiment and results. Discussions and conclusions are presented in Section IV and Section V, respectively.

II. REVISITING THE METHOD OF INCORPORATING PART-OF-SPEECH INTO ASR SYSTEMS

In [8], based on the fact that there are many morph entries that have different part-of-speech tags and also a lot of Kanji entries that have multiple pronunciations in Japanese, the author distinguished lexical entries by their notations, part-of-speech tags and phonetic transcription to improve language modeling.

In this paper, similar to [8], we append the part-of-speech to every word. In this case, the basic modeling unit is the pair of word and part-of-speech instead of word in traditional n-gram model. So the multi-category words in Chinese are considered as different tokens while the single-category words remain unchanged. Taking the word “黄” for example, it is changed into “黄/noun”(meaning “a surname”), “黄/verb”(meaning “fizzle out”) and “黄/adjective”(meaning “yellow”) after refinement. It is observed that lexical ambiguity is decreased in this way.

The pronunciation dictionary is modified correspondingly. For instance, the word “教授”(meaning “professor” or “teaching”) is pronounced differently (the verb is pronounced “jiao1shou4” and the noun “jiao4shou4”). Table I gives the comparison between the traditional and revised pronunciation dictionaries. As showed in Table I, two different pronunciations with equal probability are provided for the two pronunciations of word “教授” in the traditional dictionary. In this case, we have no idea about its pronunciation even when knowing the part-of-speech S. By contrast, the pronunciation is totally determined by the part-of-speech in our dictionary.

TABLE I
COMPARISON BETWEEN THE TRADITIONAL AND REVISED DICTIONARIES

Traditional dictionary	Revised dictionary
… 黄 huang2 教授 jiao1shou4 教授 jiao4shou4 …	… 黄/verb huang2 黄/noun huang2 黄/adjective huang2 教授/verb jiao1shou4 教授/noun jiao4shou4 …

III. EXPERIMENTS AND RESULTS

This section first describes the corpora used to train language models in this paper, and experiment settings. And then verify the effectiveness by introducing part-of-speech into ASR. To study how part-of-speech affects the speech performance, we analyze the multi-category words. Based on the classification of these words, further experiments are conducted. The experiments results are presented and analyzed at last.

A. Corpora

Two corpora tagged with parts-of-speech are selected to train language models. The first is the People's Daily corpus (PDC) [9] which is tagged manually. However, it is so expensive to annotate raw corpus manually so that another automatically-tagged corpus, called Tagged Chinese Gigaword corpus (TCGC) [10] is adopted to verify the effectiveness of the method.

The PDC was released in 2003 by the Institute of Computational Linguistics, Peking University. It contains 12 month's data from People's Daily (2000) and consists of about 13 million tokens. According to our statistics, only 16.61% of Chinese word types in the PDC are multi-category words, but they occupy a high proportion, over 50%. This means that multi-category words make up most of the corpus.

The TCGC was released in 2009 by Linguistic Data Consortium (LDC). It contains about 832 million Chinese words, including 501 million words from Taiwan’s Central News Agency (CAN), 312 million words from Mainland China's Xinhua News Agency (XIN) and 19 million words from Zaobao Newspaper (ZBN). Because all of the language models in this paper are evaluated on a Mandarin speech recognition task, we discard the corpus from CAN which is published in traditional characters. The part-of-speech tag set of the TCGC is different from the one used in the PDC. Convenient for the following experiments, the two tag sets are uniformed by mapping the latter into the former based on [11].

B. Preliminary experiment overall multi-category words

In our experiments, the SRILM toolkit [12] is used to train 3-gram language models on the two corpora described above and back-off Kneser-Ney smoothing [13] is employed. The part-of-speech tag is surplus to the needs of the traditional N-gram language models and thus it is removed when building them.

The acoustic model is trained on 1300-hour Mandarin Broadcast News speech by ML estimation. Feature extraction is carried out at a frame rate of 10ms using a 25ms Hamming window. A pre-emphasis factor of 0.97 is employed. 12 Mel Frequency Cepstral Coefficients (MFCCs) and log-energy with corresponding first and second order time derivatives is calculated. The acoustic model consists of about 8000 tied states and each state is modeled with 32 Gaussian mixtures. 1060 utterances consist of the test set, which comes from the same source as the acoustic training set, and there is no overlap between them. Table II gives the results of speech recognition when all multi-category words distinguished by part-of-speech.

TABLE II
SPEECH RECOGNITION RESULTS OF THE BASELINE AND THE PROPOSED MODEL ON PDC AND TCGC.

LMs	CER (%)	Rel.Impr. (%)
PDC_baseline	15.24	-
PDC_proposed	15.13	0.72
TCGC_baseline	11.72	-
TCGC_proposed	11.37	2.99

In Table II, results of the experiments on the two corpora show the recognition performance is improved by incorporating part-of-speech information, which verifies the validity of the method. A relative CER reduction of around

3% is achieved on the TCGC. The reason is that the PDC facing more severe data sparse problem since its size is smaller than the TCGC.

As mentioned before, part-of-speech can bring two aspects of benefits: reducing some confusion in language model and determining the pronunciation of heteronyms more or less. Although the experimental results in this section have showed the improvement of recognition performance, we do not know which benefit causes the improvement. Next, we show the analysis of multi-category words and related experiments.

C. *Analysis of multi-category words*

Based on the amount of information about the pronunciation of a word that part-of-speech can give, multi-category words are divided into three sets. The first set (A) contains the multi-category words that have only one pronunciation and part-of-speech gives no any information about their pronunciations. The second set (B) consists of heteronyms whose pronunciations are completely decided by part-of-speech, which means that one knows how to pronounce them when knowing their part-of-speech. The last set (C) includes those words whose pronunciations are partly decided by part-of-speech, e.g. there are two alternative pronunciations for the Chinese word "把" when being a noun, only one pronunciation being a verb or a quantifier. Table III shows the statistical information of these three sets of multi-category words on the PDC.

TABLE III
STATISTICS OF MULTI-CATEGORY WORDS FROM PDC

	#words	Percentage of vocabulary (%)	Percentage of corpus (%)
A	11,701	16.07	39.85
B	175	00.24	10.99
C	272	00.37	5.99
Total	12,148	16.68	56.83

From Table III, we can see that 16.68% of vocabulary are multi-category words and they take over more than half of the tokens of the corpus. The number of the words of set B and C is much smaller than A, but they occur more frequently. In another word, those heteronyms (Set A and B) are always high-frequency words in Mandarin Chinese. This suggests that the recognition performance will be improved significantly because the ambiguity caused by these words is decreased.

Based on the classification of multi-category words, a number of experiments are implemented on the PDC. Fig.1 illustrates the CERs of recognition on the PDC when refining different set of words, respectively. For example, in Fig.1 "A" means refining the words of set A solely while "B&C" means refining the words of set B and C as the same time.

From Fig.1 we can see that, if only refining the words of set A, the performance is worse than the baseline. By contrast, the CER is reduced when refining those words of set B or C. It is worth noting that, much more performance improvement is obtained by only refining the words of set B than refining all of the multi-category words (set A, B and C) in the corpus. That is to say, the performance degrades when refining the words of set A. This phenomenon is caused by the more severe sparseness data problem. Table IV shows the average frequency of the words of each set on the PDC before and after refined.

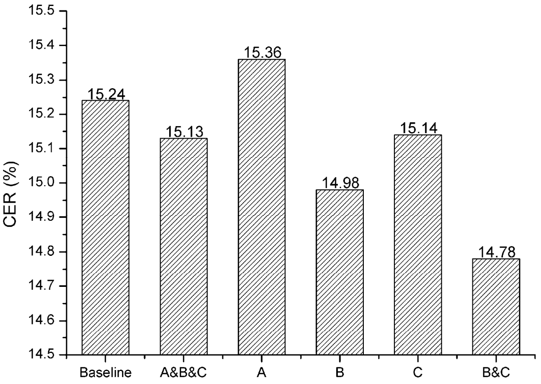

Fig. 1 CERs of recognition when refining different sets of multi-categories on the PDC.

TABLE IV
AVERAGE FREQUENCY OF MULTI-CATEGORY WORDS OF THE PDC BEFORE AND AFTER REFINING THEM

	Before refinement	After refinement
A	433	204
B	7,983	4,802
C	2,799	1,021

In Table IV, the average frequency of set A is 433, much smaller than those of set B (7,983) and C (2,799). After refined, the average frequency of set B and C is still significantly greater than set A. Data sparseness is more severe, the parameters estimation of language model is more unreliable. As for the words of set B or C, the performance should be poorer than the baseline since they meet the same problem. One explanation for this phenomenon is that the second benefit of part-of-speech mitigates the harm coming from data sparseness. What's more, the sparseness problem for set B and C is smaller than set A.

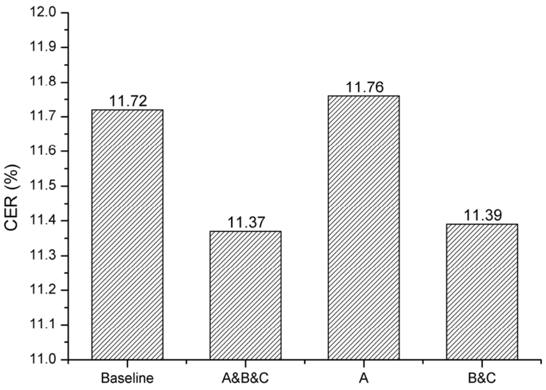

Fig. 2 CERs of recognition when refining different sets of multi-categories on the TCGC

The results of experiments when refining different sets of words on the TCGC are summarized in Fig.2.

As can be seen in Fig.2, the experimental results on the TCGC are similar to those on the PDC. It confirms that the method is also useful when using the corpus tagged automatically to train language models. When refining the words of set A solely, the performance is also worse than baseline but closer to it, which means that the problem of data sparseness is alleviated.

Comparing the experiments on the PDC and TCGC, an intuitive conclusion can be obtained: if the training data for language modeling is too sparse, only refining the words of set B and C brings the most performance improvement.

IV. Discussions

In Fig.1 and Fig.2, we can see that part-of-speech is useful for the performance improvement of speech recognition. The most improvement is gained by refining the words of set B and C, which is because part-of-speech reduces the pronunciation confusion of heteronyms significantly.

However, when refining the words of set A solely, the results of experiment on the two corpora are both worse than the baseline. The interpretation for this phenomenon is that the words of this set are rare, when refining these words, the data sparseness problem becomes more severe. As the corpus increases, the data sparseness problem weakens, that is why the result on the TCGC is closer to the baseline. The benefit of part-of-speech for the words of set A will be reflected when using a larger corpus than the TCGC. Furthermore, part-of-speech could not resolve the ambiguity for these words at the acoustic level at all.

Compared to the results of C in Fig.1, we find that refining the words of set B beings more significant improvement, which is because the pronunciations of the words of set B are more certain than the latter. However, for the words of set C, the acoustic ambiguity is partly resolved by introducing part-of-speech as well.

V. Conclusions

In this paper, we study the effect of part-of-speech on speech recognition at two levels, lexical and acoustic, by integrating part-of-speech into language model and pronunciation dictionary.

The results of experiments on the two corpora, with different sizes and tagged differently, show that the performance of ASR is significantly improved. The relative CER reductions are both up to 3%. It is worth noting that one of the corpora is automatically tagged and the experimental result improves as much as the corpus annotated by manual, which means the generality of the method is effective and can be easily extended to large-scale corpus.

What's more, we find that the part-of-speech helps to improve the performance of recognition mainly by resolving the acoustic ambiguity of heteronyms. When refining the words of set A solely, the experimental results on both corpora are worse than baseline, while refining the words of set B or C improves the recognition performance. It implies that the reduction of confusion of pronunciation improves the recognition performance significantly. Incorporating richer knowledge cues, such as syntactic structure and semantics, to reduce the confusion of the language model and pronunciation dictionary to obtain further more performance improvement.

Acknowledgment

The work was supported in part by the National Basic Research Program of China (2013CB329304) and National Natural Science Foundation of China (No. 61121002, No.91120001).

References

[1] Chelba C, Jelinek F. Structured language modeling [J]. Computer Speech & Language, 2000, 14(4): 283-332.

[2] Rosenfeld R. A maximum entropy approach to adaptive statistical language modelling [J]. Computer speech and language, 1996, 10(3): 187.

[3] Wu J, Khudanpur S. Combining nonlocal, syntactic and n-gram dependencies in language modeling [C]//Proceedings of Eurospeech. 1999, 99: 2179-2182.

[4] Brown, P. F., DellaPietra, V. J., deSouza, P. V., Lai, J. C. and Mercer, R. L. "Class-based n-gram models of natural language", Computational Linguistics., 18(4):467-479, 1992.

[5] Yamamoto H, Isogai S, Sagisaka Y. Multi-class composite N-gram language model [J]. Speech Communication, 2003, 41(2): 369-379.

[6] Zitouni I. Backoff hierarchical class N-gram language models: effectiveness to model unseen events in speech recognition [J]. Computer Speech & Language, 2007, 21(1): 88-104.

[7] Bilmes, J. A. and Katrin, K., "Factored language models and generalized parallel backoff", Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003--short papers-Volume 2. Association for Computational Linguistics, 2003.

[8] Kawahara T, Kobayashi T, Takeda K, et al. Japanese Dictation Toolkit: Plug-and-Play framework for speech recognition R&D [J]. 1999.

[9] http://www.icl.pku.edu.cn

[10] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=L\\DC2009T14

[11] Huang, C. R., Lee, L. H., Qu. W., et al., "Quality assurance of automatic annotation of very large corpora: a study based on heterogeneous tagging system", Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco, 2008.

[12] Stolcke, A., "SRILM-an extensible language modeling toolkit", In Hansen, J. H. L. and Pellom, B. editors, Proc. ICSLP, 2(901-904), Denver, Sep. 2002.

[13] Kneser, R. and Ney, H., "Improved backing-off for m-gram language modeling", Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, 181-184, 1995.