CHINESE LANGUAGE MODEL ADAPTATION BASED ON DOCUMENT CLASSIFICATION AND MULTIPLE DOMAIN-SPECIFIC LANGUAGE MODELS

Sung-Chien Lin[1], Chi-Lung Tsai[1], Lee-Feng Chien[2], Ker-Jiann Chen[2], Lin-Shan Lee[1,2]
[1]Dept. of Computer Science and Information Engineering, National Taiwan University
[2]Institute of information Science, Academia Sinica
Taipei, Taiwan, Republic of China
lsc@speech.ee.ntu.edu.tw

ABSTRACT

Adaptation of language models to the specific subject domains is definitely important for real speech recognition applications. In this paper, a Chinese language model adaptation approach is presented mainly based on document classification and multiple domain-specific language models. The proposed document classification method using the perplexity value and word bigram coverage value as primary measures are able to model word associations and syntactic behavior in classifying documents into the clusters and thus creates more effective domain-specific language models. The adaptation of language model in speech recognition can be therefore effectively achieved by the proper selection of the most appropriated domain-specific language model. Preliminary tests have been made in application to Mandarin speech recognition and shown its exciting performance of the proposed approach in creating real applications.

I. INTRODUCTION

Statistical N-gram language models provide very useful linguistic constraints in speech recognition to guide the search for the most possible word string of dictated speech. Unfortunately, although such a language model is able to predict short-distance dependence of the language quite well, it is not really efficient in modeling long-distance dependence and adaptable with the change of different subject domains. Degradation of language model performance due to different subject domains has always been a serious problem [1]. Thus, adaptation of language models to the specific subject domains is definitely important for real applications [2,3,4]. In this paper, a Chinese language model adaptation approach is presented mainly based on document classification and multiple domain-specific language models.

In order to obtain more reliable model parameters, conventional N-gram language models needs large amounts of training corpus to construct. Since it may be resulted from uneasy collection of training corpus, traditionally the documents in the corpus are totally used to train a general language model without carefully considering their characteristics among different subjects. As a consequence, the language model will be averaged and smeared out, even though the amounts of training corpus can be continuously increased. With the growth of electronic documents published and distributed over the Internet, the problem of corpus collection is less difficult than it was before. Since training corpus can be seen as a large heterogeneous collection of homogeneous documents, a possible solution to alleviate the above difficulty is to classify the training documents into several subject domains based on the homogeneity of the documents. For each subject domain, we can then use the domain-specific documents to train a domain-specific language model. Such domain-specific language models are able to describe the special linguistic characteristics of the documents in the corresponding subject domains with limited training corpus. The adaptation of language model in speech recognition can be achieved effectively by the proper selection of the most appropriated domain-specific language model. In this paper an approach based on this idea is proposed.

The remaining parts of this paper are organized as follows. In Section II, we will describe the general concept of the proposed approach. Then in Section III the method of document classification, which is the core technology of this approach, will be introduced. Furthermore, the performance of the language model adaptation approach by testing the Mandarin dictation task will be presented and concluding remarks given in Section IV.

II. THE PROPOSED APPROACH FOR CHINESE LANGUAGE MODEL ADAPTATION

Basically, in the training phase the proposed approach classifies the documents in the corpus into several subject domains and constructs multiple domain-specific language models for all subject domains. And, in the recognition phase, it can adaptively select an appropriate domain-specific language model with the input of dictated speech and change of subject domain.

When a new training document is collected, it only needs to classify the document to a specific subject domain or uses it to define a new subject domain determined by the classification parameters. The classification of the training documents is the core technology of the proposed approach. It was designed based on concepts of perplexity and word bigram coverage. For more clear

description, it will be described in more details in the next section. Whenever all of the training documents classified, multiple domain-specific language models are then trained using these clustered documents, and each will be interpolated with a general domain model trained by using all documents in the collection. Such an interpolation makes each subject domain require only limited training texts. For example, a bigram probability of word w_i given word w_j for a certain domain-specific language model can be defined as,

$$\Pr_S'(w_i|w_j) \overset{def}{=} \lambda \Pr_S(w_i|w_j) + (1-\lambda)\Pr_G(w_i|w_j) \quad \text{.....(1)}$$

where $\Pr_S(w_i|w_j)$ and $\Pr_G(w_i|w_j)$ are the bigram probabilities which are estimated with the relative frequencies appearing in the corpus of subject S and in the general domain corpus, respectively. The weighting factor λ is dependent on the reliability of the obtained domain-specific probability values. It can be determined using the Bayesian estimation or MAP estimation [4].

During the recognition phase, limited number of beginning sentences are used to adapt the language model, in the other words, to select the most appropriate domain-specific language model for dictated speech. The process of speech recognition and language model adaptation is described in brief. The beginning sentences of the dictated speech are first linguistic decoded with respect to all the available domain-specific language models and for each domain-specific language model we can obtain a decoding word string with a decoding score. The decoding scores are then used in the selection of the final output for the sentence, because the scores act as the indices for the capability of the corresponding language models to provide linguistic constraints for speech recognition of the sentence. In such a way, the decoding word string with the largest decoding score is selected as the result. On the other hand, the decoding scores are also used in the selection of the domain-specific language model to be used for the following input speech according to the same reason. These decoding scores are accumulated with respect to all the domain-specific language models. When the accumulating score of a certain domain-specific language model is below a dynamic adjusting threshold, it provided a clue that the language model is not appropriate for linguistic decoding of the following speech. The language model can be filtered out temporarily and the following sentences are linguistic decoded with respect to all the remaining domain-specific language model.

III. DOCUMENT CLASSIFICATION FOR DOMAIN-SPECIFIC LANGUAGE MODEL

Document classification provides efficient clustering of documents with similar information or coefficients into clusters. It is a frequently-used skill in the area of information retrieval to provide efficient file access by limiting the searches to those document clusters which appear to be most similar to the corresponding queries [5]. Conventional approach to document classification is based on weight matrix processing, which needs to create a keyword-class (two dimension) real matrix that shows the relative importance of the keywords in each class. This kind of document classification is mainly based on distribution of keyword frequency and weak in applying word associations and syntactic information to the clustering. In our approach for classifying all of the training documents into different subject domains in order to have domain-specific language models, we use new measures, i.e., the perplexity value [6] and the word bigram coverage.

The perplexity value of a document with respect to a certain domain-specific language model (or document class) is defined as follows:

$$PP \overset{def}{=} 2^{-\frac{1}{N}\sum_{i=1}^{N}\log \Pr(w_i|w_{i-1})} \quad \text{.....(2)}$$

where N is the length of the examined document. Basically, the perplexity value are usually used to represent the prediction power with the certain language model to the document, which means the possible number of words followed a word in the document on average based on the prediction of the language model. The complexity of word frequencies and word associations conceived in this certain domain can be modeled. Therefore, the perplexity value is a very useful measure which is able to consider syntactic behavior in some degree for the classification of documents.

On the other hand, the word bigram coverage of a language model to a document is defined as,

$$C_{WB} \overset{def}{=} \frac{\sum_{i=1}^{N} d_i}{N} \times 100\% \quad \text{.....(3)}$$

where N is also the length of the examined document. d_i is used being the presence of the word pair (w_i,w_{i+1}) in the document appearing in the training document for the language model. If the word pair (w_i,w_{i+1}) has appeared once more in the training documents for the language model, the value of d_i is assigned to 1, otherwise to 0. In this way, for a document which is more similar to the documents in a certain subject domain than those in the

other domains, the estimated perplexity value of the corresponding domain-specific language model is possible to be lower than those of the other domain-specific language model to the document, but the word bigram coverage is possible to be higher, and vice versa.

Preliminary experiments were performed based on the above two measures. Five subject domains were used: philosophy (PHI), literature (LIT), baseball news (BBL), Windows software (WIN) and science (SCI). The training documents were obtained from newspapers, magazines and the Internet. The size of the training corpus for each subject domain are listed in Table 1, with BBL the largest and SCI the smallest. Domain-specific language models were then trained first using these domain-specific documents and then by interpolation with a general-domain model trained by the general-domain training documents also listed in Table 1.

Another set of testing documents were chosen outside of the above training documents, and for each document the domain-selection score was obtained based the perplexity value and word bigram coverage of each domain-specific language model with respect to the document. The domain classification was then performed using this score. After the classification, the average perplexity and word bigram coverage values for each set of testing documents classified into a specific subject domain evaluated with respect to each domain-specific language model are listed in Table 2 and 3 respectively.

From these two tables, it is clear that the correct domain specific language model always gives the lowest perplexity values and the highest word bigram coverage values. It is also interesting to note that the subject domain BBL was best classified (with significantly smaller perplexity and slightly higher word bigram coverage), while the classification of the subject domain SCI is the least apparent (with only slightly smaller perplexity and slightly higher word bigram coverage). It was believed that this was because not only the size of training texts for BBL was the largest but that for SCI was the smallest, but the domain-specific vocabulary and word association patterns are much more complicated for SCI than for BBL. Consequently, using multiple domain-specific language model makes it possible to capture the special linguistic characteristic among different subject domains.

IV. EXPERIMENTAL RESULTS IN MANDARIN SPEECH RECOGNITION AND CONCLUDING REMARKS

In order to realize the performance of the proposed language model adaptation approach. The speech recognition tests were then performed in speaker dependent mode also with the above set of testing tests, using the acoustic recognition module and complete recognition system for very large vocabulary Mandarin speech recognition [7]. For each dictated speech, both methods of the general domain language model and the language model adaptation using the proposed approach are tested. For language model adaptation, the domain selection is performed every 5 sentences. Fig. 1 shows the averaged prediction rate of subject domains for different numbers of input sentences. It is clear that the prediction rate of subject domain for the dictated speech is high in the speech recognition test. The finally decoded character accuracy rates are listed in Table 4. It can be found that with the selected domain-specific language models, the accuracy rates can be improved significantly for all cases. Again the most significant improvements were obtained in the subject domain BBL.

To sum up the above introduction, it is believed that adaptation of language models to the specific subject domains is definitely important for real speech recognition applications. In this paper, a Chinese language model adaptation approach has been presented based on the document classification and multiple domain-specific language models. The proposed document classification method using the perplexity value and word bigram coverage value as measures as shown in preliminary experiments are able to model word associations and syntactic behavior in classifying documents into clusters. Multiple domain-specific language models are then created. Meanwhile, the adaptation of language model in speech recognition can be achieved effectively by the proper selection of the most appropriated domain-specific language model. Preliminary tests have been made in application to Mandarin speech recognition and shown its exciting performance of the proposed approach in creating real applications.

REFERENCES

[1]Y-J. Yang, S-C. Lin, L-F. Chien, K-J. Chen, and L-S. Lee, “An Intelligent and Efficient Word-Class-Based Chinese Language Model for Mandarin Speech Recognition with Very Large Vocabulary, ” in Proc. ICSLP’94, pp. 1371-1374, Japan, 1994.

[2]S. Matsunaga, T. Yamada and K. Shikano, “Task Adaptation in Stochastic Language Models for Continuous Speech Recognition, ” in Proc. ICASSP’92, pp. 165-168, CA, USA, 1992.

[3]R. Kneser and V. Steinbiss, “On the Dynamic Adaptation of Stochastic Language Models,” in Proc. ICASSP’93, pp. 586-589.

[4]M. Federico, “Bayesian Estimation Methods for N-Gram Language Model Adaptation,” in Proc. ICSLP’96.

[5]R. R. Larson, “Experiments in Automatic Library of Congress Classification,” Journal of American Society of Information Science, Vol. 43, No. 2, pp. 130-149,

1992.

[6]F. Jelinek, "The Development of an Experimental Discrete Dictation Recognizer," Proceedings of IEEE, Vol. 73, No. 11, pp. 1616-1624, Nov., 1985.

[7]H-M. Wang, L-S. Lee, et. al., "Complete Recognition of Continuous Mandarin Speech for Chinese Language with Very Large Vocabulary," ICASSP'95, Vol. I, pp. 61-64, U.S.A., 1995.

subject domains	general-domain	PHI	LIT	BBL	WIN	SCI
no. of words	12,094,234	67,127	61,676	170,214	16,647	3,445
no. of characters	18,384,664	100,054	87,987	226,864	23,463	5,612

Table 1 The size of the training texts for different subject domains

		testing texts with different subject domains				
		PHI	LIT	BBL	WIN	SCI
domain-specific language models	PHI	**352**	765	1067	765	521
	LIT	554	**301**	1174	781	587
	BBL	601	838	**37**	604	534
	WIN	876	1239	1349	**402**	473
	SCI	1249	2221	2376	1812	**362**

Table 2 perplexity values for testing texts with different subject domains evaluated with respect to different domain-specific language model

		testing texts with different subject domains				
		PHI	LIT	BBL	WIN	SCI
domain-specific language models	PHI	**37.28%**	31.39%	23.45%	28.62%	12.42%
	LIT	25.95%	**52.24%**	20.77%	27.89%	8.97%
	BBL	24.68%	29.28%	**99.70%**	35.55%	11.96%
	WIN	13.73%	17.48%	15.93%	**40.28%**	12.94%
	SCI	5.02%	3.30%	3.30%	4.71%	**17.88%**

Table 3 word bigram coverage values for testing texts with different subject domains evaluated with respect to different domain-specific language model

		testing documents with different subject domains				
		PHI	LIT	BBL	WIN	SCI
language models used in linguistic decoding	general domain	82.52%	82.04%	81.43%	82.28%	87.04%
	domain adaptation	84.43%	85.68%	88.70%	87.13%	91.13%

Table 4 character accuracy for the speech recognition tests using different language models

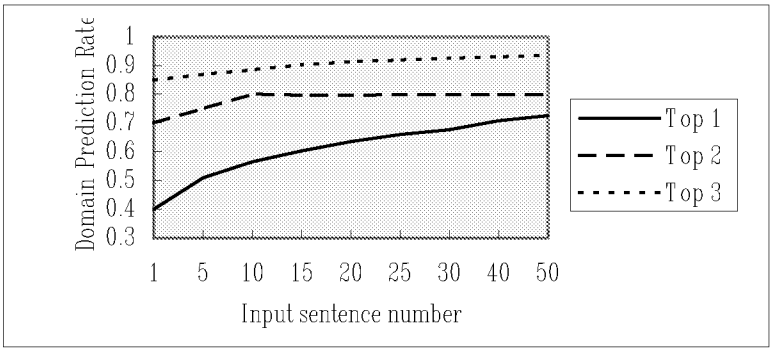

Fig. 1 subject domain prediction rate of different numbers of input sentences in the speech recognition tests