Language Model Adaptation Based on the Classification of a trigram's Language Style Feature

Qi LIANG, Thomas Fang ZHENG, Mingxing XU, and Wenhu WU
Center for Speech Technology, the State Key laboratory of Intelligence Technology and System,
Department of Computer Science and Technology, Tsinghua University, Beijing, 100084
Email: liangq, fzheng, xumx, wuwh@cst.cs.tsinghua.edu.cn

***Abstract*-In this paper, an adaptation method of the language style of a language model is proposed based on the differences between spoken and written language. Several interpolation methods based on trigram counts are used for adaptation. An interpolation method considering Katz smoothing computes weights according to the confidence score of a trigram. An adaptation method based on the classification of a trigram's style feature computes weights dynamically according to the trigram's language style tendency, and several weight generation functions are proposed. Experiments for spoken language on the Chinese corpora show that these methods, especially the method considering both a trigram's confidence and style tendency, can achieve a reduction in the Chinese character error rate for pinyin-to-character conversion.**

I. INTRODUCTION

Usually the corpus used for training a statistical language model is based on written language, and therefore this kind of language model can be called a written language model. However, daily communication such as face-to-face talking, phone calls, chats on the Internet, and mobile phone short messaging, is mostly carried out in spoken language. There are many differences between spoken language and written language. It can be foreseen that if a written language model is used for continuous speech recognition [1] and/or full sentence input of Chinese [2] where spoken language dominates, the performance of the system will degrade due to the mismatch between the training conditions and the application conditions of the language model [3].

According to the principle of maximum likelihood estimation (MLE) [4], if we have enough spoken language text to train a spoken language model, the mismatch problem can be better resolved. But actually collecting a sufficient amount of spoken language text is difficult. An alternative way to create a spoken language model might be to adapt a written language model trained from a large amount of text with a spoken language model trained from a relatively small amount of text.

A traditional language model adaptation technique is to combine a well-trained general language model with a poorly-trained domain-specific language model to form a new language model in the specific domain [5]. To distinguish this approach, such methods are referred to as domain adaptive methods in this paper. Usually there are two types of domain adaptive methods used: interpolation-based [6, 7] and maximum entropy-based [8]. The interpolation method is more commonly used. It is easy to implement and has a high computational performance, however, it is difficult to maintain the integrity of the language model and to achieve the best interpolation results. The maximum entropy method can optimize the interpolation results, but it often leads to a large amount of computation. In addition, language model adaptation methods can be either online or offline according to how the adaptive corpus is collected. Offline adaptation requires more concrete application conditions and needs a larger adaptive corpus than online adaptation, while online adaptation usually addresses the needs of a specific user. In this paper, we will focus on an offline adaptation method for a spoken language model based on the interpolation method.

This paper is organized as follows: The differences between spoken language and written language in Chinese will be described briefly in Section II. Next, Section III will present common language model adaptation algorithms, and Section IV will describe special adaptive algorithms facing to spoken language and written language. We will report experiment results and analysis in Section V. Finally, in Section VI, we will give a summarization of this paper.

II. DIFFERENCES BETWEEN SPOKEN LANGUAGE AND WRITTEN LANGUAGE

Spoken language is the most basic form of language, and is quite different in wording, phrasing and construction from written language. In regards to wording and phrasing, there are many more substantive words showing daily life and having concrete meanings, there are less abstract words, there are more words expressing various feelings, and there are more interjections. In regards to construction, it is more vivid, brief and changeable.

In fact, the differences between spoken language and written language can be shown statistically. Table I below lists the statistical results of probabilities of some Chinese words in a large-scale written language corpus and a small-scale spoken language corpus, both of which are in Chinese. From these samples we can see that the probabilities of interjections and words used specially for spoken language in the spoken language corpus are much greater while the

0-7803-9361-9/05/$20.0 © 2005 IEEE

TABLE I

COMPARISON OF PROBABILITIES OF SOME WORDS IN THE WRITTEN LANGUAGE CORPUS AND THE SPOKEN LANGUAGE CORPUS

		Written Language Corpus(W)	Spoken Language Corpus(S)	Ratio of Probability (W/S)
		Probability	Probability	
Interjection	啊	3.2×10^{-5}	6.8×10^{-3}	4.7×10^{-3}
	吧	4.2×10^{-5}	1.1×10^{-2}	3.8×10^{-3}
	吗	5.2×10^{-5}	1.2×10^{-2}	4.3×10^{-3}
	呀	1.8×10^{-5}	2.9×10^{-3}	6.2×10^{-3}
Word used specially for the spoken language	爸	3.7×10^{-6}	5.7×10^{-4}	6.5×10^{-3}
	妈	5.7×10^{-6}	1.2×10^{-3}	4.8×10^{-3}
Word used specially for the written language	父亲	4.6×10^{-5}	1.3×10^{-5}	3.5
	母亲	4.9×10^{-5}	1.7×10^{-5}	2.9

probabilities of words used specially in written language in the written language corpus are greater.

The differences between spoken language and written language are not the same as those between different domains. It is not suitable to directly use the domain adaptive method mentioned in Section I for language style adaptation for the reasons listed below:

(1) In the present, the domain adaptive method for language models is mostly for written language in which the style of construction is the same; however, in this paper, the language model adaptive method is for the adaptation from the written language model to the spoken language model where the styles of construction are different.
(2) Usually domain-specific words are not high-frequency ones; while a language model's style specific words are typically the ones occurring very often in its own style and very seldom in another style.

To differentiate the domain adaptive method mentioned above for language model adaptation, we call the adaptive method for adaptation from a written language model to a spoken language model (or vice versa) as the language style adaptive (LSA) method in this paper.

III. COMMONLY USED LANGUAGE MODEL ADAPTATION METHODS

Symbols to appear in the following equations are: $P(x)$ denotes the probability of event x and $C(x)$ the count; items with subscript c are related to the *common* language model trained from a large-scale written language corpus, those with subscript a are related to the *adaptive* language model trained from a small-scale spoken language model, and those corresponding items without any subscript are related to the resulted language model after adaptation; w denotes the current unit and h the historical units. The common language model, the adaptive language model, and the resulted language model in this paper are all trigram language models.

A. General interpolation method

Usually an interpolation method can be described using the equation in terms of probability as follows [6, 7]:

$$P(w\mid h)=\lambda P_c(w\mid h)+(1-\lambda)P_a(w\mid h) \tag{1}$$

In this paper we will use the count-based equation as follows:

$$C(h,w)=C_c(h,w)+\alpha C_a(h,w) \tag{2}$$

The two equations are in fact equivalent, which can be proved as follows:

$$\begin{aligned} P(w\mid h) &= \frac{C(h,w)}{C(h)} \\ &= \frac{C_c(h,w)+\alpha C_a(h,w)}{C_c(h)+\alpha C_a(h)} \\ &= \frac{\frac{C_c(h,w)}{C_c(h)}+\alpha\frac{C_a(h,w)}{C_a(h)}\cdot\frac{C_a(h)}{C_c(h)}}{1+\alpha\frac{C_a(h)}{C_c(h)}} \end{aligned} \tag{3}$$

From (1) and (2), if we define

$$\frac{C_a(h)}{C_c(h)}=A(h), \tag{4}$$

equation (3) can be rewritten as

$$P^{(C)}(w\mid h)= \frac{1}{1+\alpha\cdot A(h)}P_c(w\mid h)+\frac{\alpha\cdot A(h)}{1+\alpha\cdot A(h)}P_a(w\mid h) \tag{5}$$

where

$$\lambda=\frac{1}{1+\alpha\cdot A(h)} \tag{6}$$

We can see that (1) is equivalent to (5). What's more, in (5) λ can be elaborately adjusted according to different h.

It should be noticed that α is a predefined interpolation weight, and the larger the value assigned to α, the greater effect the adaptive model will take. $A(h)$, related to the two corpora, is the ratio of the count of h in the adaptive corpus to that in the common corpus, and reflects the different importance of h in the two corpora. The larger $A(h)$ is, the greater the effect the adaptive model will have. Equation (5) also indicates the physical meanings of α and $A(h)$.

B. Interpolation method considering Katz smoothing

Trigram language models are currently widely used. A problem they must face is the data sparseness issue, that is, parameters are too abundant to be trained directly from corpus. An effective approach to this problem is language model smoothing combining both discounting and regression. The basic idea of such language model smoothing is straightforward: it simply tries to take out a number of

occurrence counts from the seen units and redistribute them to the unseen ones. In this paper we will use the Katz smoothing technique [9, 10] which can be expressed with the following equation:

$$P_{Katz}(w_i \mid w_{i-2}, w_{i-1}) = \begin{cases} C(w_{i-2}, w_{i-1}, w_i) / C(w_{i-2}, w_{i-1}) \\ \quad if \quad C(w_{i-2}, w_{i-1}, w_i) > r_T \\ d_r C(w_{i-2}, w_{i-1}, w_i) / C(w_{i-2}, w_{i-1}) \\ \quad if \quad 0 < C(w_{i-2}, w_{i-1}, w_i) \le r_T \\ \alpha(w_{i-2}, w_{i-1}) P_{Katz}(w_i \mid w_{i-1}) \\ \quad if \quad C(w_{i-2}, w_{i-1}, w_i) = 0 \end{cases} \quad (7)$$

where r_T is the count threshold used for discount, and $\alpha(w_{i-2}, w_{i-1})$ and d_r are the smoothing parameters.

In the Katz smoothing method, trigrams whose counts are less than r_T will be discounted. Those trigrams with low counts are considered to have low statistical confidence, so they will be discounted to trigrams with count 0. Similarly, trigrams with high counts are considered to have high statistical confidence, and their probabilities will remain unchanged.

Because the size of the adaptive corpus is smaller than that of the common corpus, the value of α is set greater than 1 so that the adaptive model can achieve an ideal effect. The value is related to the differences of the scale and the trigram distribution between the two corpora. If a relatively large weight α is assigned to a trigram with count lower than r_T, this trigram's original low confidence value might be enhanced artificially (and excessively), which will result in those trigrams that should be discounted escaping from being discounted. Furthermore, any trigram with a low confidence should not be assigned a large probability value, because: (1) if its probability is larger than other counterpart trigrams when it should not be larger, those trigrams will not survive in decoding and thus decoding errors will be caused; (2) if its probability is smaller than other counterpart trigrams when it should not be smaller, the worst possible outcome is that it will not survive in decoding when it should, which will only affect itself, and not affect any other trigrams. Based on the above analysis, the probabilities of trigrams with low confidence should be too small rather than too large [10].

It can be seen that α in (2) and (5) is not very suitable for trigrams with low counts (less than r_T). Accordingly, Equation (2) is modified to (8).

On the one hand, because the value of r_T in Katz smoothing is rather small, if the value of β in (8) is too large, some trigrams with low confidence will escape from being discounted. On the other hand the value of β as a weight of adaptive model should not be smaller than 1. Therefore, the value of β is chosen as 1 in this paper.

$$C(h, w) = \begin{cases} C_c(h, w) + \alpha C_a(h, w) \\ \quad if \quad C_c(h, w) + \beta C_a(h, w) > r_T \\ C_c(h, w) + \beta C_a(h, w) \\ \quad if \quad C_c(h, w) + \beta C_a(h, w) \le r_T \end{cases} \quad (8)$$

IV. Adaptation Methods Based On The Classification Of A Trigram's Language Style Feature

The distribution of trigram units in a written language corpus is different from that in a spoken language corpus, which Table I indirectly points to. Actually, for most trigram units, the ratio of the count in a written language corpus to that in a spoken language corpus is comparative to the ratio of the scales of the two corpora, because the spoken language and the written language of a same language must follow the same rules of this language, and as a result the difference between them is limited.

However there are still some trigram units having rather distinct distributions in the two corpora, and it is these units that represent the features of spoken or written language. For example, the pinyin string "*wo shi wei ni hao*"("我是为你好"), is decoded into "我市为你好" using the language model trained from a written language corpus. As can be analyzed, "我市" appears more in a written language corpus than in a spoken one while such subjective language phenomena as "我是" appear more rarely in a written language corpus but more frequently in a spoken language corpus. Thus the use of a spoken language model in decoding will result in the correct sentence.

The size of the spoken language corpus in our experiments is small and hence the trigrams' counts in it are pervasively less than those in the written language corpus, and if every trigram is equally weighted those trigrams with spoken language features might not survive in full sentence decoding. In the above example, if "我是" and "我市" are equally weighted, the decoding result will remain as "我市为你好." But too large of weights may destroy the probabilities of those fully and accurately trained trigrams and lead to performance degradation of the overall model. Therefore the weights should be carefully selected. An advisable idea is to give larger weights to trigrams with spoken language features like "我是" and to give no weight or a bit smaller weights to trigrams with written language features like "我市." This is the basic idea of the language style adaptive method we propose in this paper. In other words, we propose to give different weights to trigrams according to their language style, spoken or written. In this method, the identification or classification of a trigram's language style feature is rather important.

In fact, the basic idea of language model adaptation is to compensate a common model with information contained in the adaptive model that is insufficient in the common model,

so as to get a new language model that could better match the application conditions. The common model itself contains a great deal of general language information and written language specific information but very little spoken language specific information, while the adaptive model contains some general language information and a comparatively great deal of spoken language specific information. As the common model is fully trained, the general language information contained in the adaptive model will take little effect and hence needs no weight (or a relatively small weight); the spoken language specific information is what the common model lacks (and therefore what must be compensated), and thus should be weighted more. With regards to written language specific information, if it will affect the performance of the new language model after adaptation, it could be weakened or given no weight.

A. Dynamic-weight adaptation method based on a trigram's language style feature

Based on the above analysis, the first step we should take is to classify a trigram's language style feature before adaptation.

In this paper, we proposed to take $\frac{C_c(h,w)}{C_a(h,w)}$ as the criterion to judge the tendency towards a certain language style (we'll call this language style tendency) of a trigram (*h*, *w*). Obviously, as shown in Equation (4), it is the reciprocal of *A*(*h*, *w*). The bigger $\frac{C_c(h,w)}{C_a(h,w)}$ is, the more likely the trigram is of written language; whereas, the smaller $\frac{C_c(h,w)}{C_a(h,w)}$ is, the more likely the trigram is of spoken language.

Based on the mentioned criterion, Equation (2) is modified into

$$C(h,w)=\begin{cases} C_c(h,w)+\alpha C_a(h,w) \\ \quad if\ \ \frac{C_c(h,w)}{C_a(h,w)}<\theta_s\ \ (spoken\ \ language) \\ C_c(h,w)+\beta C_a(h,w) \\ \quad if\ \ \theta_s\le\frac{C_c(h,w)}{C_a(h,w)}\le\theta_w \\ \gamma C_c(h,w)+C_a(h,w) \\ \quad if\ \frac{C_c(h,w)}{C_a(h,w)}>\theta_w\ \ (written\ \ language) \end{cases} \tag{9}$$

where α is the weight for a trigram's count in the adaptive model when the trigram is of the spoken language style; β is the weight for a trigram's count in the adaptive model when the trigram is a general one; γ is the weakening weight for a trigram's count in the common model when the trigram is of the written language style; θ_s is the threshold used to judge whether a trigram is of the spoken language style; and θ_w is the threshold used to judge whether a trigram is of the written language style . Generally, $\alpha>\beta\ge 1$, $\gamma\le 1$.

In this paper, a weight generating function is constructed as below:

$$\alpha = f\left(\frac{C_c(h,w)}{C_a(h,w)}\right) \tag{10}$$

The function *f* can be selected in several ways, for example, *f*(*x*) can be defined as a constant function, an increasing linear function, an increasing convex function, or an increasing concave function. Different forms of the weight generating function have different meanings: a constant function means that all the trigrams classified as spoken language are equally weighted; an increasing function means that the trigram will be more greatly weighted if its spoken language style tendency is stronger; and the convex or concave tendency denotes an increasing speed when the weight increases along with the spoken language style tendency. In this paper, these different forms of *f*(*x*) will be compared experimentally.

B. Dynamic-weight adaptation method based on a trigram's language style feature with the consideration of Katz smoothing

Based on the analysis in Sections III-*B* and IV-*A*, a method combining both the dynamic-weight adaptation method based on a trigram's language style feature and the Katz smoothing is proposed in this section. The basic idea here is given in the following equation, which is a combination of (8) and (9):

$$C(h,w)=\begin{cases} C_c(h,w)+\alpha C_a(h,w) \\ \quad if\ \ \frac{C_c(h,w)}{C_a(h,w)}<\theta_s\ and\ C_c(h,w)+C_a(h,w)>r_T \\ C_c(h,w)+\beta C_a(h,w) \\ \quad if\ \ \theta_s\le\frac{C_c(h,w)}{C_a(h,w)}\le\theta_w\ and\ C_c(h,w)+C_a(h,w)>r_T \\ \gamma C_c(h,w)+C_a(h,w) \\ \quad if\ \ \frac{C_c(h,w)}{C_a(h,w)}>\theta_w\ and\ C_c(h,w)+C_a(h,w)>r_T \\ C_c(h,w)+C_a(h,w) \\ \quad if\ \ C_c(h,w)+C_a(h,w)\le r_T \end{cases} \tag{11}$$

The meanings of symbols in (11) are the same as those in (8) and (9) except that the values of some parameters are a little different: α is defined by the weight generating function (10) and its overall value is less than that in (8) and larger than that in (9); β and γ are less affected and hence could be adjusted slightly or could remain unchanged.

V. Experiments And Analysis

The written language corpus used to train the common language model consists of about 274MB of text taken from Chinese newspapers such as *People's Daily* and *Economic Daily*; the spoken language corpus used to train the adaptive model consists of about 7.4MB of text which are actual short messages collected through mobile phones. The spoken language corpus used for testing consists of 500 short messages provided by Nokia Research Center in China and is not included in the training corpus. The size of the vocabulary in the trigram language model is 25,851.

Similar to the written language corpus, the spoken language corpora used for training and testing are of miscellaneous domains, therefore the experiments below are seen to be not general domain adaptation but rather language style adaptation from a written language model to a spoken language model. The experimental platform used here is a Chinese pinyin-to-character conversion system [5].

Several weight generating functions $f(x)$, mentioned in Section IV-*A*, were compared for the adaptation methods described in Sections IV-*A* and IV-*B* (referred to as Methods IV-*A* and IV-*B* hereinafter). The resulting character error rates (CER) of pinyin-to-character conversion are listed in Table II.

It can be seen from Table II that the performance when using a constant function as the weight generating function is the best in Method IV-*A* while the performance when using an increasing convex function as the weight generating function is the best in Method IV-*B*. The reason is that the Katz smoothing is not considered in Method IV-*A* and that the language style features' tendency of the trigrams with low confidence is not reliable itself. Thus, in this condition, a static weight outperforms a dynamic one. Such experimental results also show the necessity of Method IV-*B*.

In the following experiments, the weight generating function $f(x)$ with the best performance was selected for Method IV-*A* and Method IV-*B*, respectively. A pinyin-to-character conversion system based on the common language model trained by a large-scale written language corpus was used as Baseline 1; the pinyin-to-character conversion system based on the adaptive model trained by a small-scale spoken language corpus was used as Baseline 2. The experimental results of the language model adaptation using the methods described in Sections III-*A*, III-*B*, IV-*A*, and IV-*B* are listed in Table III in terms of the pinyin-to-character conversion CER.

TABLE II

CER Comparison of Different Weight Generating Functions Used in Methods IV-*A* and IV-*B*

	Constant function	Increasing linear function	Increasing convex function	Increasing concave function
Method IV.*A*	3.74%	3.88%	3.76%	3.82%
Method IV.*B*	3.43%	3.46%	3.32%	3.40%

TABLE III

The Performance Comparison of Several Adaptation Methods

	Baseline 1	Baseline 2	Method III-*A*	Method III-*B*	Method IV-*A*	Method IV-*B*
CER	6.66%	4.35%	3.90%	3.43%	3.74%	3.32%
CER decline from baseline 1			41.4%	48.5%	43.8%	50.2%
CER decline from baseline 2			10.3%	21.1%	14.0%	23.7%
CER decline from Method III-*A*				12.1%	4.1%	14.9%

From Table III, the following conclusions can be drawn:

(1) When the testing corpus is of spoken language, the performance of the language model trained using the 274MB written language corpus is worse than that of the language model trained using the 7.4MB spoken language corpus. The results show that when the training condition and the testing condition mismatch with each other, no matter how big the training corpus is the performance is still very poor; and on the contrary when the training condition and the testing condition match with each other, the performance will be rather good even if the training corpus is not so big. The results also show that differences between written language and spoken language do exist.

(2) Any of the four proposed adaptation methods outperform any of the two baselines, which confirm our idea of adaptation from written language to spoken language.

(3) Any of Methods III-*B*, IV-*A* and IV-*B* outperforms the general interpolation method described in Section III-*A*, which shows: a) the idea of the trigram weighting according to the confidence score of the trigram is reasonable; b) considering the language style feature's tendency of a trigram could improve adaptation performance, which also confirms that the difference of language styles is not the same as the difference of domains.

VI. Conclusions

In this paper, a method for language model adaptation from a written language model to a spoken language one is proposed based on the classification of a trigram's language style feature. In this method, each trigram is first classified into either having a written language style tendency or a spoken language style tendency, and then is given a different adaptation weight based on such a language style feature's tendency. In the commonly used adaptation methods, the interpolation method considering Katz smoothing actually computes the weights according to the confidence score of a trigram. The adaptation method based on the classification of a trigram's language style feature computes weights dynamically according to the trigram's language style

tendency. The pinyin-to-character conversion experiments with a spoken language testing corpus show that the dynamic-weight adaptation method based on a trigram's language style feature with consideration of Katz smoothing is the best, and can reduce the CER to a great extent. In this method, an increasing convex function is the most effective dynamic weight generating function.

Finally, we should mention that there are several parameters that are not very easy to determine. These parameters need to be changed when the training condition of the language model changes. How to find a better solution to this will be the focus of our subsequent research.

ACKNOWLEDGEMENT

The authors gratefully acknowledge Nokia Research Center in China for providing the testing corpus, and our colleague Michael Brasser for his helpful writing advice.

REFERENCES

[1] Fang Zheng, Zhanjiang Song, Pascale Fung, William Byrne. "Mandarin Pronunciation Modeling Based on CASS Corpus," *J. Computer Science & Technology*, 17(3): 249-263, May 2002.

[2] Gengqing Wu, Fang Zheng. "A Method to Build a Super Small but Practically Accurate Language Model for Handheld Devices," *J. Computer Science & Technology*, 18(6): 747~755, 2003.

[3] R. Rosenfeld, *et al*. "Error Analysis and Disfluency Modeling in the Switchboard Domain," In: Proceedings of the 4th *International Conference on Speech and Language Processing* (*ICSLP*). Philadelphia, PA, USA, 1996.

[4] L.R. Bahl, F. Jelinek, and R. L. Mercer. "A Maximum likelihood approach to continuous speech recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. PAMI-5, No. 2, pp. 179-190, 1983.

[5] Genqing Wu, Fang Zheng, Ling Jin, Wenhu Wu. "An online incremental language model adaptation method," *EuroSpeech*, 3:2139-2142, Aalborg, Denmark, Sept. 3-7, 2001.

[6] Rukmini M. Iyer, Mari Ostendorf. "Modeling long distance dependence in language: topic mixtures versus dynamic cache models," *IEEE Transactions on Speech and Audio Processing*, Volume 7 Issue 1. Page(s): 30-39, Jan 1999.

[7] Daniel Gildea, Thomas Hofmann. "Topic Based Language Models Using EM," In Proceedings of 6th *European Conference on Speech Communication and Technology* (*Eurospeech*'99).

[8] R. Rosenfeld. "A Maximum Entropy Approach to Adaptive Statistical Language Model," *Computer Speech & Language*, 10: 187~228, 1996.

[9] S. M. Katz. "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer," *IEEE Transaction on Acoustic, Speech and Signal Processing*, 35(3): 400~401, 1987.

[10] Genqing Wu. Fang Zheng, Wenhu Wu, Mingxing Xu, Ling Jin, "Improved Katz smoothing for language modeling in speech recognition," *International Conference on Spoken Language Processing* 2002, pp. 925-928, Colorado, USA, Sep. 16-20, 2002.