ISCA Archive
http://www.isca-speech.org/archive

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

# AN IMPROVED MAP METHOD FOR LANGUAGE MODEL ADPAPTATION

Langzhou Chen    Taiyi Huang

Institute of Automation, Chinese Academy of Science, Beijing, China 100080
email {clz,huang}@nlpr.ia.ac.cn

## ABSTRACT

This paper presents an improved MAP method for language model adaptation. The traditional MAP method mixes the task independent corpus and task dependent corpus using a fixed weight. In the method presented in this paper, we replaced the fixed weight with a function of history word. Another work in this paper is that a fuzzy controller was introduced in adaptation process, and three factors were used to be the input of the controller, they are: 1) the confidence of the estimation value, 2) the importance of the word, 3) the difference between the estimation value from the general corpus and from the adaptive corpus. The experiments showed that the improved method has the better performance than traditional model.

## 1 INTRODUCTION

The performance of large vocabulary speech recognition system depends on the prediction ability of the language models strongly[1].Statistical N-gram language model has been used to guide searching possible word string successfully. But the n-gram model's performance depends on the topic of training corpus strongly. the general language model always have no good performance in some special domain. Reference[2] proposed that cluster the corpus into different domain, then train the domain special language model using the domain dependent corpus respectively. But this method need that every domain have enough corpus to train the domain dependent model. When the domain dependent corpus is not enough, the problem of sparse data [3]will be very serious. In this case, language model adaptation, i.e. how to change a task independent language model to a good performance task dependent model using a small corpus in such special domain, should be a useful technology to solve this problem. The method using maximum a posteriori (MAP) criterion is a effective adaptation method. As we know, the MAP criterion can be express as:

$$P = \arg\max_{p} (f(x|p) * f(p)) \tag{1}$$

where $f(x/p)$ is the likelihood probability of task dependent corpus assumed that estimation value of language model $p$ is known. And $f(p)$ is the a-priori distribution of $p$. Let $L(p) = f(x|p) * f(p)$ and $d(L(p))/d(p) = 0$ ,we can know that:

$$p(w|v) = \frac{Count_I(vw) + \varepsilon Count_A(vw)}{\sum_w (Count_I(vw) + \varepsilon Count_A(vw))} \tag{2}$$

where $\varepsilon$ is a constant.

The method presented above has been proved to be very effective[4]. But it is not the best result. We think that we can get better result using two method. the first, we replaced the constant $\varepsilon$ in equation 2 with a function of history word $v$, the second, we have introduced a fuzzy controller to mix the task dependent and independent corpus. We will discuss it in detail in following section

## 2 WEIGHTING FUNCTION FOR ADAPTATION PROCESS

.

Traditional MAP method can improve the performance of language model very much. But it is not rational that make $\varepsilon$ to be a constant. We have introduced a weight function $\varepsilon(v)$ to replace the constant $\varepsilon$ in Eq(2). Then Eq(2) can be rewrote as :

$$p(w|v) = \frac{Count_I(vw) + \varepsilon(v) Count_A(vw)}{\sum_w Count_I(vw) + \sum_w \varepsilon(v) * Count_A(vw)} \tag{3}$$

We think that the probability estimated from adaptive corpus $P_A(w/v)$ has the adaptation information that more reliable than $Count_A(vw)$ has, so we defined $\varepsilon(v)$ as :

$$\varepsilon(v) \triangleq \gamma * Count_I(v) / Count_A(v) \tag{4}$$

where $\gamma$ is the coefficient, then we can get

10.21437/Eurospeech.1999-422

$$p(w/v) = \frac{Count_I(vw) + \gamma * Count_I(v) * P_A(w|v)}{\sum_w (Count_I(vw) + \gamma * Count_I(v) * P_A(w|v))} \quad (5)$$

After modified the Eq(2) as Eq(5), the performance of model has improved evidently.

## 3 FUZZY CONTROLLER FOR ADAPTATION PROCESS

Adaptation using Eq(5) is better than traditional method, but it still has some potential to improve the performance. In this paper, we try to tune the coefficient $\gamma$ to get more improvement.

When we tune the coefficient $\gamma$ ,we take into account the factors that:

1) The confidence of estimation value $P_A(w/v)$. As we know, the adaptation corpus is always small, and many parameters estimated from it have low confidence. We can assume that the higher confidence of $P_A(w/v)$, the more it can represent the real estimated value of $w$'s probability in the special domain, so the coefficient $\gamma$ will be big, otherwise, $\gamma$ will be small.
2) The importance of the word $w$.We only define the content word's importance. The importance of $w$ can be defined as :

$$U(w) = P(T_i|w) \quad (6)$$

where $T_i$ is the special domain we discussing. The more important word should have big $\gamma$ value.

3) The difference between $P_A(w/v)$ and $P_I(w|v)$ .If $P_I(w|v) << P_A(w/v)$, it means $P_A(w/v)$ have much useful information, else it is not very useful. The experiment shows that we give the bigger $P_A(w/v)$ a bigger $\gamma$ , the performance of the language model will improved a little.

According to these three factors, we present a fuzzy method to determine the value of $\gamma$ , as showed in figure 1, x(vw),y(vw),z(vw) is the inputs of fuzzy controller:

x(vw) represents the confidence of $P_A(w/v)$

$$x(vw) = Count_A(vw) \quad (7)$$

We divided x into 2 levels: HC(high confidence) and LC(low confidence), the membership function was showed in figure 2.

y(vw) represents the importance of $w$. We sort all the words in vocabulary from big to small according to the value of U(w) defined in Eq(6). Then we defined the y(w) as the series number of w. It means that the smaller y(w) is the more important w is. We divided y into 2 level too, i.e. IM(important), NI(no important). The membership function was showed in figure 3.

z(vw) represents the difference between $P_A(w/v)$ and $P_I(w|v)$, defined as

$$z(vw) = (P_I(w|v) - P_A(w|v)) / P_I(w|v) \quad (8)$$

z(w) was divided into 3 levels: N ($P_I(w|v) << P_A(w/v)$), Z (there is not much difference between $P_A(w)$ and $P_I(w)$), P ($P_I(w|v) >> P_A(w/v)$)

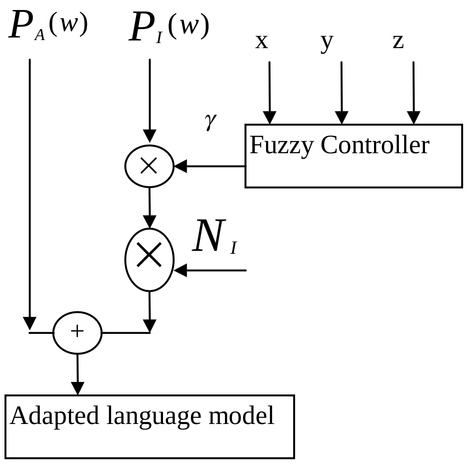


Figure 1

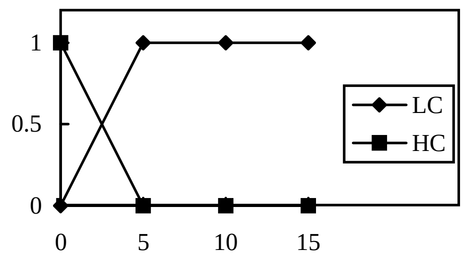


Figure 2

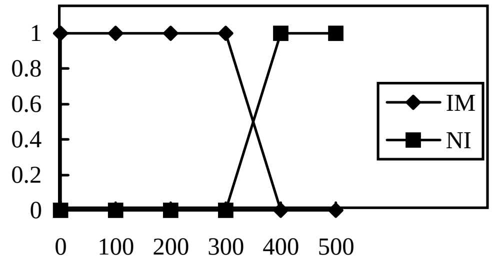


Figure 3

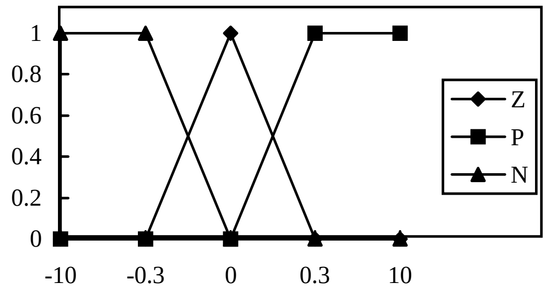


Figure 4

The output of the fuzzy controller is $\gamma$ . We divided it into 5 levels, i.e. ZE(zero), PS(positive small), PM(positive medium),PB(positive big), VB(very big). The controlling rules is showed in table 1.

| | z is N | z is Z | z is P |
|---|---|---|---|
| x is LC and y is NI | ZE | ZE | ZE |
| x is LC and y is IM | PS | PS | PS |
| x is HC and y is NI | PM | PB | VB |
| x is HC and y is IM | PM | PB | VB |

Table 1

Tuning the value of ZE, PS, PM, PB, VB seriously, the performance of adapted model can be improved evidently.

## 4. EXPERIMENT

We used the corpus from "People Daily" to train our independent language model and collected some corpus about the topic of special domain to train our domain special model. The size of corpora is showed in table 2
Using a corpus of sport domain about 30K words as our test corpus, the test result of bigram model is showed in table3.
It showed that using conventional MAP method , we can improved our model very much . But it is not the best result. Our improved method can reduce the perplexity farther. The perplexity reduced about 6.2% after we introduced $\varepsilon(w)$, reduced about 10.3% after we introduced fuzzy method.

| corpus | size (words) |
|---|---|
| domain independent | 61,461,753 |
| international news | 1,059,795 |
| national news | 1,503,522 |
| economy | 1,257,723 |
| sport | 265,718 |
| arts | 1,324,261 |
| science and technology | 118,969 |

Table 2

| Language model | perplexity |
|---|---|
| independent model | 281.34 |
| adapted model using Eq2 | 195.72 |
| adapted model using Eq5 | 183.67 |
| adapted model using Eq5+Fuzzy control | 175.35 |

Table 3

Table 4 is the perplexities of new method comparing with traditional MAP method in other 5 domains. Every domain has its special test corpus.

| | Adapted model using Eq2 | adapted model using Eq5+Fuzzy control |
|---|---|---|
| international news | 190.62 | 171.43 |
| national news | 189.53 | 169.88 |
| arts | 187.65 | 168.34 |
| economy | 189.01 | 169.12 |
| science and technology | 198.42 | 180.41 |

Table 4

Our next experiment is testing the performance of new model when the size of adaptation corpus changed. Figure5 show the change of test corpus' perplexity when the adaptation corpus of sport domain increased from zero to 260K words.
We can find that when the adaptation corpus increased, the perplexity of both the traditional MAP method and our new method decreased, but the new method fell faster than traditional method. It means that the new method need smaller corpus to reach the same performance of the traditional method.

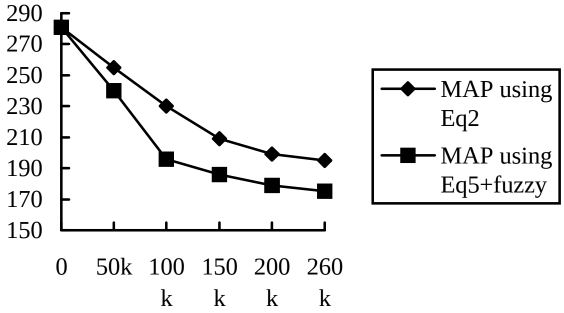


Figure 5

## 5. CONCLUSION

In this paper, we present a new language model adaptation method. Traditional MAP method mixes the topic independent corpus and topic dependent corpus using a fixed coefficient, our method introduced two method to improve it. The first, we think the probability estimated from adaptive corpus $P_A(w|v)$ has more adaptation information than $Count_A(vw)$ has, so we replaced the fixed coefficient with a function of history word , and this function is proportional to $P_A(w|v)$ .

The second, we designed a fuzzy controller to the adaptation process. Three factors are introduced as the input of fuzzy controller, they are 1. The confidence of the estimation value from adaptation corpus, 2.the importance of the word to be considered, 3. The difference between the estimation value from general corpus and adaptation corpus. After introduced these two improvement, the performance of the model is much better than traditional method.

We think that the fuzzy controller adopted here has provided a very good framework to the adaptation process. Many other information can be introduced easily into this framework.

## 6. REFERENCE

[1] F.Jelinek, Self-Organized Language Model for Speech Recognition. Readings in Speech Recognition . Morgan Kaufman Publisher,1990.

[2] Reinhard Kneser and Jochen Peters. Sematic Clustering for Adaptive Language Model. In Proceeding of ICASSP-97.

[3] SLAVA M.KATZ . Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer. IEEE Transon ASSP , VOL. ASSP35,No.3,March.

[4] Marcello Federico, Bayesian Estimation Methods for N-Gram Language Model Adaptation. In Proceeding of 1996 International Conference of Spoken Language Processing.：240－243.

[5] Abraham Kandel et al. Fuzzy Control System, Boca Raton :CRC Press,c1994