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

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

A Novel Language Model Based on Self-organized Learning

Taiyi Huang , Langzhou Chen
Institute of Automation, Chinese Academy of Science, Beijing 100080
{huang, clz}@nlpr.ia.ac.cn

10.21437/Eurospeech.1999-424

ABSTRACT

Statistical language model is very important to speech recognition. To a system of special topic, domain dependent language model is much better than general model. There are two problems in traditional method to train topic dependent model: 1. The corpus of special topic is not as enough as general corpus. 2. An individual article always relates to more than one topics, traditional method has not considered this phenomena. This paper try to solve these two problems. We have present a new method to organize the corpus--- the method based on fuzzy training subset. And the training of domain dependent models are based on these fuzzy subsets. At the same time, a self organized learning approach is introduced in training process to improve the models' predicting ability. The self organized learning can improve the performance of models evidently.

1. INTRUDUCTION

In speech recognition, statistical n-gram model has been successfully used to guild the search and score a path of word string[1]. But a general language model can not use the topic information of speech content efficiently. So the performance of general model will drop when it is used for a specific domain. Topic dependent language model is a effective way to get better performance in special domain. There are two ways to build the topic dependent language model. One is mixture models[2], in this structure, the language models of different topics are all interpolated together according to the mixing factors. It can be expressed as.

$$p(w_i|w_{i-1}) = \sum_k x_k p_k(w_i|w_{i-1}) \qquad \textbf{(1)}$$

where x_k is the mixing factor of topic *k* and $p_k(w_i|w_{i-1})$ is the language model of topic *k*. The other is single model structure[3]. In this structure, topic dependent model is interpolated with a general language model which can be expressed as:

$$p(w_i|w_{i-1}) = x * p_g(w_i|w_{i-1}) + (1-x) * p_k(w_i|w_{i-1})$$

(2)

where *x* is weighting factor and $p_g(w_i|w_{i-1})$ a general language model.

In order to introduce the topic dependent language models into speech recognition, there are two problems must be solved. Firstly, how to update the mixing factors. This problem can be solved by EM algorithm[4], which the mixing factors can be estimated as:

$$x_{n+1}(i) = \frac{1}{M}\sum_{m=0}^{M-1} \frac{x_n(i) * p_i(w_{n-m}|h_{n-m})}{\sum_{j=1}^{I} x_n(j) * p_j(w_{n-m}|h_{n-m})}$$

(3)

where $x_n(i)$ is the mixing factor of topic *i* in the *nth* iteration, *M* is the length of training data.

The second problem is the training of topic dependent language models. Traditional method to build the topic dependent language model is as follows: tagging the training corpus into different topic manually or automatically, then training the topic dependent language model using the tagged corpus. This method sometimes will bring to two problems: 1. Corpus of a specific domain is not as sufficient as general corpus, it will lead to data sparseness problem. 2. An individual article in training text sometimes involved with more than one topic, traditional training process did not consider this problem. In order to describe the phenomena of an article related to several topics, we presented a fuzzy method to get the corpus of a special topic. If we defined a general corpus $U = \{u_1, u_2, \cdots, u_n\}$ as our universe of discourse, unlike the traditional method to divide the corpus into different topics, we defined the corpus of special topic as a fuzzy subset of the general corpus. There are not precise boundaries between topics, and every topic is defined by a membership function of its fuzzy subset.

We have presented the topic dependent models' training algorithm based on the fuzzy training subset. In order to improve the prediction ability of models, a self organized learning process has been introduced into the training. The experiments showed that the models trained based on fuzzy training subset and self organized learning are better than the traditional method both in mixture model structure and single model structure.

2. BUILDING FUZZY TRAINING SUBSET

Fuzzy training subset is the base of the new models. It can be expressed as

$$U = \{u_1, u_2, \cdots, u_n\}$$

$$Topic_j = \sum_{i=1}^{n} \frac{A_j(u_i)}{u_i} \qquad (4)$$

where $Topic_j$ is the fuzzy training subset of topic j , $A_j(u_i)$ is membership function. The main work of building fuzzy training subset is to determine the membership function to every fuzzy subset relevant to different topics. We think that the topic feature of articles is contained in the keywords. The keywords are the words which are most representative in every topic. The frequency of the appearance of the keyword has the burst character, i.e. it occur frequently in relevant topic while seldom occur in other articles. The information of single keyword is not confident, but if many keywords co-occur in same article, they can provide the reliable evidence for topic detection.

2.1 KEYWORD SELECTION

Since the keywords occur frequently in some articles but seldom occur in other place, we can detect the keywords according to this burst character.

To every word w_i, the probability of it appearing in article u_j is

$$p(w_i \in u_j) = \frac{Count(w_i, u_j)}{Count(w_i)} \qquad (5)$$

Cluster the probability values $p(w_i \in u_j),\ j = 1,2,\cdots n$ into two classes and maximize the value

$$d(w_i) = \frac{|m_2 - m_1|}{\sigma_1 + \sigma_2} \qquad (6)$$

where m_1, m_2 are the means of two classes and σ_1, σ_2 are the variances of two classes. We can find that if $d(w_i)$ is very large, it means that w_i occurs frequently in some articles while seldom occurs in others and w_i is a keyword; if $d(w_i)$ is small, the distribution of $p(w_i \in u_j)$ is even in different article, w_i is not a keyword. Through this way we have selected 3,000 keywords from a lexicon of 40,000 words.

2.2 MEMBERSHIP FUNCTION

Because the keywords contain the topic information of articles, we propose the keyword distribution vectors as the feature to calculate the membership function of special topic. Every article in corpus $u_i, i = 1\cdots n$ can be represented as a keyword distribution vector $vector(u_i)$. The dimension of the vector is the number of keywords, in our system, the dimension is 3000. Every element of the vector is the times that relevant keyword happen in articles. At the same time, every topic was represented as a kernel vector. The kernel vectors represent the center of every topic in keyword space and they were determined as follows. Divide the general corpus into different topic manually or automatically like the traditional method, then, we can get the mean vector of every topic as Eq7:

$$\mathrm{Ker}(Topic_j) = \frac{1}{n_j} * \sum_{aticle \in Topic_j} vector(article) \quad (7)$$

where n_j is number of articles in topic j which determined by traditional method, and $vector(*)$ is the keyword distribution vector of articles. These mean vector is the kernel vector of every topic. Because the topic information of the article is contained in its keyword distribution vector, we can think that the degree of the correlation between an article and a topic is determined by the similarity of article's keyword distribution vector and topic's kernel vector. When two vectors overlap completely, it means the content of article coincide with the topic, if two vectors are perpendicular, the article has no relationship with the topic. So an article's membership value to a topic is determined by the cosine angle between article's keyword distribution vector and topic's kernel vector. Therefore the membership function of topic *j* can be expressed as follows:

$$A_j(u_i) = f\left(\frac{vector(u_i) \bullet \mathrm{Ker}(Topic_j)}{|vector(u_i)| * |\mathrm{Ker}(Topic_j)|} \right) \quad (7)$$

From Eq. 7, the support of a topic can be expressed as:

$$\mathrm{Supp}Topic_j = \left\{ u \left| \frac{vector(u) \bullet \mathrm{Ker}(Topic_j)}{|vector(u)| * |\mathrm{Ker}(Topic_j)|} > 0 \right. \right\}$$

(8)

In our system, the membership function is designed as a step function:

supposed that the step function have m values $0 < \lambda_1 < \lambda_2, < \cdots < \lambda_m < 1$ and $\Phi_j(u_i) = \frac{vector(u_i) \bullet \mathrm{Ker}(Topic_j)}{|vector(u_i)| * |\mathrm{Ker}(Topic_j)|}$ represented the relationship between article and topic, then the membership function can be expressed as:

$$A_j(u) = \bigcup_{i=1,\cdots,m} \lambda_i \Phi_j(u) \qquad (9)$$

where $\lambda_i \Phi_j(u) = \lambda_i \wedge \Phi_j(u)$.

3. ESTIMATION OF MODEL'S PARAMETER

The estimation of model's parameter from the fuzzy training subset can be considered as a defuzzy process. The maximum likelihood estimation of bigram based on general training set can be expressed as:

$$p(y|x) = \frac{count(xy)}{count(x)} \qquad (10)$$

where *x*,*y* is the word in lexicon.

In the case of fuzzy training set of $Topic_j$, the parameters are estimated as follows. Supposed that the membership function has m values $\{\lambda_1, \lambda_2, \cdots \lambda_m\}$ (it was determined by Eq. 9), we can divide the support of $Topic_j$ as follows:

$$\mathrm{Supp}Topic_j = \bigcup_{i=1,\cdots,m}((A_j)_{\lambda i} - (A_j)_{\lambda i+1}) = \bigcup_{i=1,\cdots,m} \Psi_i$$

(11)

where $(A_j)_{\lambda i}$ is the λ_i cut set of $Topic_j$. According to Eq. 11, the articles in every subset Ψ_i have the same membership value. Using the articles in subset Ψ_i to training the model, we can get:

$$p(y|x) = \frac{count_i(xy)}{count_i(x)} \qquad (12)$$

where $count_i(*)$ is the times that event '*' being seen in subset Ψ_i. Because all the article in subset Ψ_i have the same membership value λ_i, we can think that the model estimated by the training data in Ψ_i similar to the model of topic *j* in a degree of λ_i. So we calculate the language model of topic *j* as a weighted mean showed in Eq13:

$$p_j(y|x)=\sum_{i=1}^{m}\frac{\tilde{\lambda}_i * count_i(xy)}{\tilde{\lambda}_i * count_i(x)} \qquad (13)$$

where $\tilde{\lambda}_i$ is a normalized membership value.

Comparing with the traditional method, the new method has several advantage:

Firstly, traditional method only use the article that tagged as a special topic to train the topic dependent language model, while in the new method, the training data are extended to the support of the fuzzy training subset, the problem of data sparseness is mitigated.

Secondly, comparing with the traditional method, new method reflected the phenomena that one article related with more than one topics. At the same time the new method also described the degree of an article related with a topic precisely.

4. SELF ORGANIZED LEARNING

The topic dependent language model training based on fuzzy training subset, has the advantage mentioned above, but at the same time, it bring to a problem is that the distance between models of different topic has been reduced. In traditional method, the training data of different topic have not overlap, so the distance of different model is large. We hope the language model based on fuzzy training subset not only has the advantage mentioned in last section but also keeps the large distance between models of different topic.

In this paper, we have introduced a self organized learning process to increase the distance between different models. In the case of bigram, we can build a pattern classifier as figure 1.

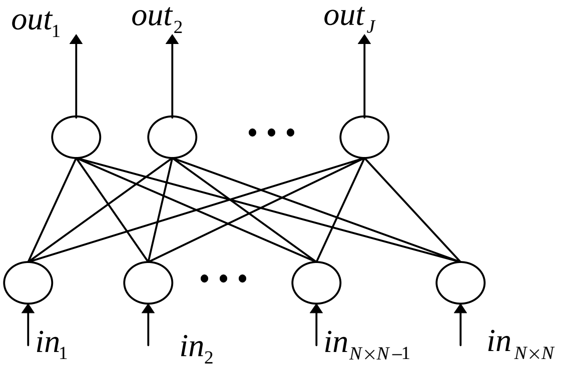

Figure 1 self organized learning

Every Language model of a specific topic is correspond to a kind of pattern. Every sentence in training data is converted into a input vector of network. The dimension of the vector is $N \times N$, where *N* is the size of vocabulary. Every element of the input vector is the times of relevant word pair was seen in the sentence. The initial weights of the network is the logarithm parameters of the models which were trained based on fuzzy training data, according to Eq. 13. The output of the network can be expressed as:

$$out_j=\sum_{i=1}^{N\times N} in_i * w_{ji} = \sum_{x,y} Count(yx) * \ln p(x/y)$$

(14)

When the training data inputted to the network, we will sort the outputs according to their values. To every input sentence, there is a model that has biggest output, the concerned weights of this model should be increased. At the same time, a lateral inhibition process is introduced, i.e. the concerned weights of several model that have smallest output should be decreased.

We using Eq16 to evaluate the distance between models.

$$\text{Distance}(Q,P)=D(Q(x/y),P(x/y))+ D(P(x/y),Q(x/y)) \qquad (16)$$

where

$$D(Q(x/y),P(x/y))=\sum_{x,y}Q(x,y)\log\frac{Q(x/y)}{P(x/y)}$$ is Kullback-Liebler distance. Experiments showed that after self organized learning the distance between models increase evidently.

5. EXPERIMENT AND RESULT

We used the corpus about 8,450,000 words, 20,000 articles from “People Daily” as our training corpus. The first experiment is based on the single model. At first, we have tagged the corpus manually and trained a language model using the data that tagged as “sport”, then we trained another model based on the fuzzy training subset of topic “sport”, the experiment result has been showed in table 1. We can find that the model based on fuzzy training subset is much better than traditional one. The major reason is that the problem of data sparseness is very serious in traditional method, while the model based on fuzzy training subset can get much more training data. When the training data is very small, traditional method using topic adaptation to get the topic dependent model. MAP (maximum a posteriori) method[5] is a representative topic adaptation method. We have built a topic dependent model using MAP method, the result is also showed in table1. We can find that the model based on fuzzy training subset is better than that based on MAP too. it is because that the new method describe accurately the relationship between articles and the topic.

Table 1

Training method	perplexity
maximum likelihood training	165.8
MAP adaptation	126.7
based on fuzzy training subset and self organized learning	121.1

the second experiment is testing the performance of mixture models. At first, all the articles were divided into 5 topic manually, and traditional mixture language models were trained according to this manually tagging. Then self-organized language model based on fuzzy subset of special topic is built using the same training data. Our text corpus is selected from “People Daily” too, the size of test corpus is 200,000 words. The content of test corpus include variety of topics and there is not overlap between training and test corpus. The experiment result is showed in table2. We can find that the models based on fuzzy training subset is better than the traditional models.

Table 2

Training method	perplexity
traditional method	122.8
based on fuzzy training subset and self organized learning	112.3

We also compared the model before and after the self organized learning,. We have calculated the distance among five topic dependent models, they are: politics, sport, art, economy, science and technology. Table3 and table4 are the distance between models before and after self organized learning respectively. They were calculated according to Eq16. We can find that self organized learning increased the distance between models evidently.

Table 3

	Model 1	Model 2	Model 3	Model 4	Model 5
Model 1	0	12.2	7.9	10.3	17.5
Model 2		0	14.1	12.2	10.5
Model 3			0	8.2	10.3
Model 4				0	8.6
Model 5					0

Table 4

	Model 1	Model 2	Model 3	Model 4	Model 5
Model 1	0	13.7	9.7	16.9	18.6
Model 2		0	17.8	15.5	13.1
Model 3			0	10.5	13.2
Model 4				0	11.3
Model 5					0

6. CONCLUSION

This paper presented a new method to build the topic dependent language model, the language model based on fuzzy training subset and self organized learning. I the new method, we expressed the topic specific corpus as a membership function to describe the relationship between the articles and topics. It can describe the phenomena that an article related with several topics and build topic dependent language models more accurately.

At the same time, a self organized learning was introduced into the training process. From this process the distance between the models of different topic was increased and the performance of topic dependent language model improved evidently.

REFERENCE

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

[2] Yoshihiko Gotoh, Steve Renals, Document Space Models Using Latent Semantic Analysis, European Conference on Speech Communication and Technology, Greece :European Speech Communication Association,1997.1443-1446

[3] Sung-Chien Lin, Chi-Lung Tsai, Chinese Language Model Adaptation Based on Document Classification and Multiple Domain-Specific Language Models, Eurospeech1997.1463-1466

[4] R.Kneser, V.Steinbiss. On the Dynamic Adaptation of Stochastic Language modeling, 1993 International Conference of Acoustics, Speech and Signal Processing. Minneapolis, MN, Vol.2:586-589, April 1993

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