Cluster-based Language Model for Sentence Retrieval in Chinese Question Answering

Youzheng Wu **Jun Zhao** **Bo Xu**
National Laboratory of Pattern Recognition
Institute of Automation Chinese Academy of Sciences
No.95 Zhongguancun East Road, 100080, Beijing, China
(yzwu, jzhao,boxu)@nlpr.ia.ac.cn

Abstract

Sentence retrieval plays a very important role in question answering system. In this paper, we present a novel cluster-based language model for sentence retrieval in Chinese question answering which is motivated in part by sentence clustering and language model. Sentence clustering is used to group sentences into clusters. Language model is used to properly represent sentences, which is combined with sentences model, cluster/topic model and collection model. For sentence clustering, we propose two approaches that are One-Sentence-Multi-Topics and One-Sentence-One-Topic respectively. From the experimental results on 807 Chinese testing questions, we can conclude that the proposed cluster-based language model outperforms over the standard language model for sentence retrieval in Chinese question answering.

1 Introduction

To facilitate the answer extraction of question answering, the task of retrieval module is to find the most relevant passages or sentences to the question. So, the retrieval module plays a very important role in question answering system, which influences both the performance and the speed of question answering. In this paper, we mainly focus on the research of improving the performance of sentence retrieval in Chinese question answering.

Many retrieval approaches have been proposed for sentence retrieval in English question answering. For example, Ittycheriach [Ittycheriah, et al. 2002] and H. Yang [Hui Yang, et al. 2002] proposed vector space model. Andres [Andres, et al. 2004] and Vanessa [Vanessa, et al. 2004] proposed language model and translation model respectively. Compared to vector space model, language model is theoretically attractive and a potentially very effective probabilistic framework for researching information retrieval problems [Jian-Yun Nie. 2005].

However, language model for sentence retrieval is not mature yet, which has a lot of difficult problems that cannot be solved at present. For example, how to incorporate the structural information, how to resolve data sparseness problem. In this paper, we mainly focus on the research of the smoothing approach of language model because sparseness problem is more serious for sentence retrieval than for document retrieval.

At present, the most popular smoothing approaches for language model are Jelinek-Mercer method, Bayesian smoothing using Dirichlet priors, absolute discounting and so on [C. Zhai, et al. 2001]. The main disadvantages of all these smoothing approaches are that each document model (which is estimated from each document) is interpolated with the same collection model (which is estimated from the whole collection) through a unified parameter. Therefore, it does not make any one particular document more probable than any other, on the condition that neither the documents originally contains the query term. In other word, if a document is relevant, but does not contain the query term, it is still no more probable, even though it may be topically related.

As we know, most smoothing approaches of sentence retrieval in question answering are learned from document retrieval without many adaptations. In fact, question answering has some

Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, pages 56–63,
Sydney, July 2006. ©2006 Association for Computational Linguistics

characteristics that are different from traditional document retrieval, which could be used to improve the performance of sentence retrieval. These characteristics lie in:

1. The input of question answering is natural language question which is more unambiguous than query in traditional document retrieval.

For traditional document retrieval, it's difficult to identify which kind of information the users want to know. For example, if the user submit the query {发明/invent, 电话/telephone}, search engine does not know what information is needed, who invented telephone, when telephone was invented, or other information. On the other hand, for question answering system, if the user submit the question {谁发明了电话？/who invented the telephone?}, it's easy to know that the user want to know the person who invented the telephone, but not other information.

2. Candidate answers extracted according to the semantic category of the question's answer could be used for sentence clustering of question answering.

Although the first retrieved sentences are related to the question, they usually deal with one or more topics. That is, relevant sentences for a question may be distributed over several topics. Therefore, treating the question's words in retrieved sentences with different topics equally is unreasonable. One of the solutions is to organize the related sentences into several clusters, where a sentence can belong to about one or more clusters, each cluster is regarded as a topic. This is sentence clustering. Obviously, cluster and topic have the same meaning and can be replaced each other. ***In the other word, a particular entity type was expected for each question, and every special entity of that type found in a retrieved sentence was regarded as a cluster/topic.***

In this paper, we propose two novel approaches for sentence clustering. The main idea of the approaches is to conduct sentence clustering according to the candidate answers which are also considered as the names of the clusters.

For example, given the question {谁发明了电话？/who invented telephone?}, the top ten retrieved sentences and the corresponding candidate answers are shown as Table 1. Thus, we can conduct sentence clustering according to the candidate answers, that are, {贝尔/Bell, 西门子/Siemens, 爱迪生/Edison,库珀/Cooper, 斯蒂芬/Stephen}.

ID	Top 10 Sentences	Candidate Answer
S1	1876年3月10日贝尔发明电话/Bell invented telephone on Oct. 3th, 1876.	贝尔/Bell
S2	西门子发明了电机，贝尔发明电话，爱迪生发明电灯。/ Bell, Siemens and Edison invented telephone, electromotor and electric light respectively.	西门子/ Siemens 贝尔/Bell 爱迪生/ Edison
S3	最近，“移动电话之父”库珀再次成为公众焦点。/Recently, the public paid a great deal of attention to Cooper who is Father of Mobile Phone.	库珀/Cooper
S4	1876年，发明家贝尔发明了电话。/In 1876, Bell invented telephone.	贝尔/Bell
S5	接着，1876年，美国科学家贝尔发明了电话；1879年美国科学家爱迪生发明了电灯。/Subsequently, American scientist Bell invented the phone in 1876; Edison invented the electric light in 1879.	贝尔/Bell 爱迪生/Edison
S6	1876年3月7日，贝尔成为电话发明的专利人。/On March 7th, 1876, Bell became the patentee of telephone.	贝尔/Bell
S7	贝尔不仅发明了电话，还成功地建立了自己的公司推广电话。/Bell not only invented telephone, but also established his own company for spreading his invention.	贝尔/Bell
S8	在首只移动电话投入使用30年以后，其发明人库珀仍梦想着未来电话技术实现之日到来。/Thirty years after the invention of first mobile phone, Cooper still anticipated the date of the realization of future phone's technology.	库珀/Cooper

S9	库珀表示，消费者采纳移动电话的速度之快令他意外，但移动电话的普及率还没有达到无所不在，这让他有些失望。/Cooper said, he was surprised at the speed that the consumers switched to mobile phones; but the popularization of mobile phone isn’t omnipresent, which made him a little bit disappointed.	库珀/Cooper
S10	英国发明家斯蒂芬将移动电话的所有电子元件设计在一张纸一样厚薄的芯片上。/England inventor Stephen designed the paper-clicked CMOS chip which included all electronic components.	斯蒂芬/Stephen

Table 1 The Top 10 Retrieved Sentences and its Candidate Answers

Based on the above analysis, this paper presents cluster-based language model for sentence retrieval of Chinese question answering. It differs from most of the previous approaches mainly as follows. 1. Sentence Clustering is conducted according to the candidate answers extracted from the top 1000 sentences. 2. The information of the cluster of the sentence, which is also called as topic, is incorporated into language model through aspect model. For sentence clustering, we propose two novel approaches that are One-Sentence-Multi-Topics and One-Sentence-One-Topic respectively. The experimental results show that the performances of cluster-based language model for sentence retrieval are improved significantly.

The framework of cluster-based language model for sentence retrieval is shown as Figure 1.

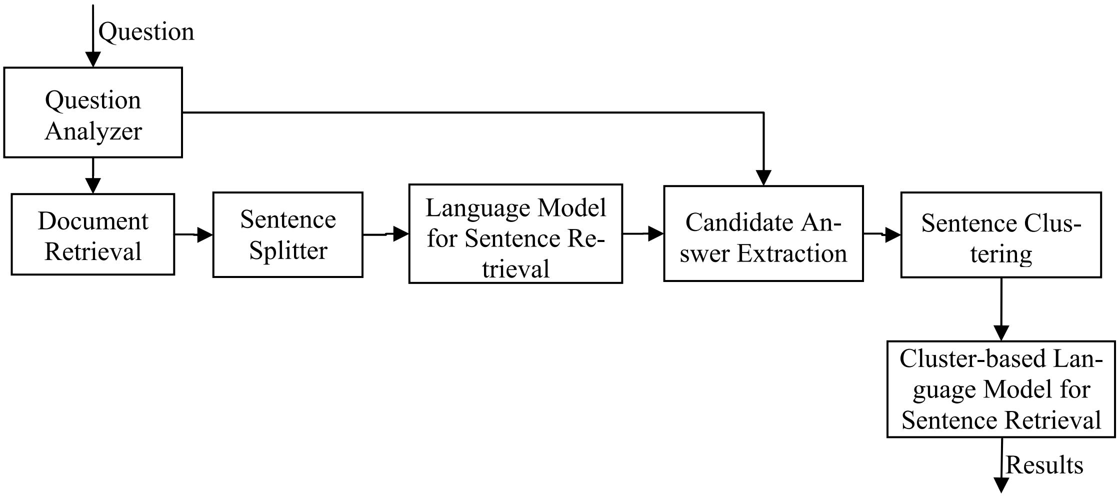
Question
Question Analyzer
Document Retrieval
Sentence Splitter
Language Model for Sentence Retrieval
Candidate Answer Extraction
Sentence Clustering
Cluster-based Language Model for Sentence Retrieval
Results

Figure 1 The Framework of Cluster-based Language Model for Sentence Retrieval

2 Language Model for Information Retrieval

Language model for information retrieval is presented by Ponte & Croft in 1998[J. Ponte, et al. 1998] which has more advantages than vector space model. After that, many improved models are proposed like J.F. Gao [J.F Gao, et al. 2004], C. Zhai [C. Zhai, et al. 2001], and so on. In 1999, Berger & Lafferty [A. Berger, et al. 1999] presented statistical translation model for information retrieval.

The basic approach of language model for information retrieval is to model the process of generating query Q. The approach has two steps. 1. Constructing document model for each document in the collection; 2. Ranking the documents according to the probabilities $p(Q|D)$. A classical unigram language model for IR could be expressed in equation (1).

$$p(Q \mid D) = \prod_{w_i \in Q} p(w_i \mid D) \qquad (1)$$

where, w_i is a query term, $p(w_i|D)$ is document model which represents terms distribution over document. Obviously, estimating the probability $p(w_i|D)$ is the key of document model. To solve the sparseness problem, Jelinek-Mercer is commonly used which could be expressed by equation (2).

$$p(w \mid D) = a \times p_{ML}(w \mid D) + (1-a) \times p_{ML}(w \mid C) \qquad (2)$$

where, $p_{ML}(w|D)$ and $p_{ML}(w|C)$ are document model and collection model respectively estimated via maximum likelihood.

As described above, the disadvantages of standard language model is that it does not make any one particular document any more probable than any other, on the condition that neither the documents originally contain the query term. In the other word, if a document is relevant, but does not contain the query term, it is still no more probable, even though it may be topically related. Thus, the smoothing approaches based on standard language model are improper. In this paper, we propose a novel cluster-based language model to overcome it.

3 Cluster-based Language Model for Sentence Retrieval

Note that document model *p(w|D)* in document retrieval is replace by *p(w|S)* called sentence model in sentence retrieval.

The assumption of cluster-based language model for retrieval is that topic-related sentences tend to be relevant to the same query. So, incorporating the topic of sentences into language model can improve the performance of sentence retrieval based on standard language model.

The proposed cluster-based language model is a mixture model of three components, that are sentence model $p_{ML}(w|S)$, cluster/topic model $p\_topic_{ML}(w|T)$ and collection model $p_{ML}(w|C)$. We can formulate our model as equation (3).

$$p(w|S) = a \times p_{ML}(w|S) + (1-a) \times \left(\beta \times p\_topic_{ML}(w|T) + (1-\beta) \times p_{ML}(w|C)\right) \tag{3}$$

In fact, the cluster-based language model can also be viewed as a two-stage smoothing approach. The cluster model is first smoothed using the collection model, and the sentence model is then smoothed with the smoothed cluster model.

In this paper, the cluster model is in the form of term distribution over cluster/topic, associated with the distribution of clusters/topics over sentence, which can be expressed by equation (4).

$$p\_topic(w|T) = \sum_{t \in T} p(w|t)p(t|S) \tag{4}$$

where, *T* is the set of clusters/topics. *p_topic(w|T)* is cluster model. *p(t|S)* is topic sentence distribution which means the distribution of topic over sentence. And *p(w|t)* is term topic distribution which means the term distribution over topics.

Before estimating the sentence model *p(w|S)*, topic-related sentences should be organized into clusters/topics to estimate *p(t|S)* and *p(w|t)* probabilities. For sentence clustering, this paper presents two novel approaches that are One-Sentence-Multi-Topics and One-Sentence-One-Topic respectively.

3.1 One-Sentence-Multi-Topics

The main idea of One-Sentence-Multi-Topics can be summarized as follows.

1. If a sentence includes M different candidate answers, then the sentence consists of M different topics.

For example, the sentence S5 in Table 1 includes two topics which are "贝尔发明电话/Bell invented telephone" and "爱迪生发明电灯/Edison invented electric light" respectively.

2. Different sentences have the same topic if two candidate answers are same.

For example, the sentence S4 and S5 in Table 1 have the same topic "贝尔发明电话/Bell invented telephone" because both of sentences have the same candidate answer "贝尔/Bell".

Based on the above ideas, the result of sentence clustering based on One-Sentence-Multi-Topics is shown in Table 2.

Name of Clusters	Sentences
贝尔/Bell	S1 S2 S4 S5 S6 S7 S8
西门子/Siemens	S2
爱迪生/Edison	S2 S5
库珀/Cooper	S3 S8 S9
斯蒂芬/Stephen	S10

Table 2 The Result of One-Sentence-Multi-Topics Sentence Clustering

So, we could estimate term topic distribution using equation (5).

$$p(w|t) = \frac{n(w,t)}{\sum_{w'} n(w',t)} \tag{5}$$

Topic sentence distribution can be estimated using equation (6) and (7).

$$p(t|S) = \frac{1/kl_{st}}{\sum_{t} 1/kl_{st}} \tag{6}$$

$$kl_{st} = KL(s||t) = \sum_{w} p_{ML}(w|s) \times log \frac{p_{ML}(w|s)}{p_{ML}(w|t)} \tag{7}$$

where, kl_{st} means the Kullback-Leibler divergence between the sentence with the cluster/topic. k denotes the number of cluster/topic. The main idea of equation (6) is that the closer the Kullback-Leibler divergence, the larger the topic sentence probability *p(t|S)*.

3.2 One-Sentence-One-Topic

The main idea of One-Sentence-One-Topic also could be summarized as follows.

1. A sentence only has one kernel candidate answer which represents the kernel topic no matter how many candidate answers is included.

For example, the kernel topic of sentence S5 in Table 1 is "贝尔发明电话/Bell invented telephone" though it includes three different candidate answers.

2. Different sentences have the same topic if two kernel candidate answers are same.

For example, the sentence S4 and S5 in Table 1 have the same topic "贝尔发明电话/Bell invented telephone".

3. The kernel candidate answer has shortest average distance to all query terms.

Based on the above ideas, the result of sentence clustering based on One-Sentence-One-Topic is shown in Table 3.

Name of Clusters	Sentences
贝尔/Bell	S1 S2 S4 S5 S6 S7
库珀/Cooper	S3 S8 S9
斯蒂芬/Stephen	S10

Table 3 The Result of One-Sentence-One-Topic Sentence Clustering

Equation (8) and (9) can be used to estimate the kernel candidate answer and the distances of candidate answers respectively. Term topic distribution in One-Sentence-One-Topic can be estimated via equation (5). And topic sentence distribution is equal to 1 because a sentence only belongs to one cluster/topic.

$$a_i^* = \underset{a_i}{argmin}\left\{SemDis_{a_i}\right\} \tag{8}$$

$$SemDis_{a_i} = \frac{\sum_j SemDis\left(a_i, q_j\right)}{N} \tag{9}$$

$$SemDis\left(a_i, q_j\right) = \left|Position_{a_i} - Position_{q_j}\right| \tag{10}$$

where, a_i* is the kernel candidate answer. a_i is the i-th candidate answer, $SemDis_{a_i}$ is the average distance of i-th candidate answer. q_j is the j-th query term, N is the number of all query terms. $Position_{q_j}$ and $Position_{a_i}$ mean the position of query term q_j and candidate answer a_i.

4 Experiments and Analysis

Research on Chinese question answering, is still at its early stage. And there is no public evaluation platform for Chinese question answering. So in this paper, we use the evaluation environment presented by [Youzheng Wu, et al. 2004] which is similar to TREC question answering track [Ellen. M. Voorhees. 2004]. The documents collection is downloaded from Internet which size is 1.8GB. The testing questions are collected via four different approaches which has 7050 Chinese questions currently.

In this section, we randomly select 807 testing questions which are fact-based short-answer questions. Moreover, the answers of all testing questions are named entities identified by [Youzheng Wu, et al. 2005]. Figure 2 gives the details. Note that, LOC, ORG, PER, NUM and TIM denote the questions which answer types are location, organization, person, number and time respectively, SUM means all question types.

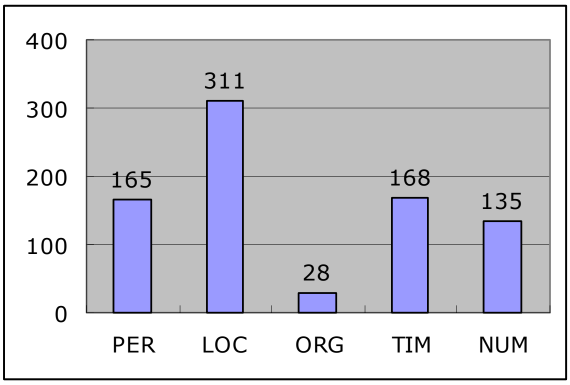

Figure 2 The Distribution of Various Question Types over Testing Questions

Chinese question answering system is to return a ranked list of five answer sentences per question and will be strictly evaluated (unsupported answers counted as wrong) using mean reciprocal rank (MRR).

4.1 Baseline: Standard Language Model for Sentence Retrieval

Based on the standard language model for information retrieval, we can get the baseline performance, as is shown in Table 4, where α is the weight of document model.

α	0.6	0.7	0.8	0.9
LOC	49.95	51.50	52.63	54.54
ORG	53.69	51.01	50.12	51.01
PER	63.10	64.42	65.94	65.69
NUM	48.43	49.86	51.78	53.26
TIM	56.97	58.38	58.77	61.49
SUM	53.98	55.28	56.40	57.93

Table 4 The Baseline MRR5 Performance

In the following chapter, we conduct experiments to answer two questions.

1. Whether cluster-based language model for sentence retrieval could improve the performance of standard language model for sentence retrieval?

2. What are the performances of sentence clustering for various question types?

4.2 Cluster-based Language Model for Sentence Retrieval

In this part, we will conduct experiments to validate the performances of cluster-based language models which are based on One-Sentence-Multi-Topics and One-Sentence-One-Topic sentence clustering respectively. In the following experiments, *β = 0.9.*

4.2.1 Cluster-based Language Model Based on One-Sentence-Multi-Topics

The experimental results of cluster-based language model based on One-Sentence-Multi-Topics sentence clustering are shown in Table 5. The relative improvements are listed in the bracket.

α	0.6	0.7	0.8	0.9
LOC	55.57 (+11.2)	55.61 (+7.98)	56.59 (+7.52)	57.70 (+5.79)
ORG	59.05 (+9.98)	59.46 (+16.6)	59.46 (+18.6)	59.76 (+17.2)
PER	67.73 (+7.34)	68.03 (+5.60)	67.71 (+2.68)	67.45 (+2.68)
NUM	52.79 (+9.00)	53.90 (+8.10)	54.45 (+5.16)	55.51 (+4.22)
TIM	60.17 (+5.62)	60.63 (+3.85)	62.33 (+6.06)	61.68 (+0.31)
SUM	58.14 (+7.71)	58.63 (+6.06)	59.30 (+5.14)	59.54 (+2.78)

Table 5 MRR5 Performance of Cluster-based Language Model Based on One-Sentence-Multi-Topics

From the experimental results, we can find that by integrating the clusters/topics of the sentence into language model, we can achieve much improvement at each stage of α. For example, the largest and smallest improvements for all types of questions are about 7.7% and 2.8% respectively. This experiment shows that the proposed cluster-based language model based on One-Sentence-Multi-Topics is effective for sentence retrieval in Chinese question answering.

4.2.2 Cluster-based Language Model Based on One-Sentence-One-Topic

The performance of cluster-based language model based on One-Sentence-One-Topic sentence clustering is shown in Table 6. The relative improvements are listed in the bracket.

α	0.6	0.7	0.8	0.9
LOC	53.02 (+6.15)	54.27 (+5.38)	56.14 (+6.67)	56.28 (+3.19)
ORG	58.75 (+9.42)	58.75 (+17.2)	59.46 (+18.6)	59.46 (+16.6)
PER	66.57 (+5.50)	67.07 (+4.11)	67.44 (+2.27)	67.29 (+2.44)
NUM	49.95 (+3.14)	50.87 (+2.02)	52.15 (+0.71)	53.51 (+0.47)
TIM	59.75 (+4.88)	60.65 (+3.89)	62.71 (+6.70)	62.20 (+1.15)
SUM	56.48 (+4.63)	57.65 (+4.29)	58.82 (+4.29)	59.22 (+2.23)

Table 6 MRR5 Performance of Cluster-based Language Model Based on One-Sentence-One-Topic

In Comparison with Table 5, we can find that the improvement of cluster-based language model based on One-Sentence-One-Topic is slightly lower than that of cluster-based language model based on One-Sentence-Multi-Topics. The reasons lie in that Clusters based on One-Sentence-One-Topic approach are very coarse and much information is lost. But the improvements over baseline system are obvious.

Table 7 shows that MRR1 and MRR20 scores of cluster-based language models for all question types. The relative improvements over the baseline are listed in the bracket. This experiment is to validate whether the conclusion based on different measurements is consistent or not.

	One-Sentence-Multi-Topics		One-Sentence-One-Topic	
α	MRR1	MRR20	MRR1	MRR20
0.6	50.00 (+14.97)	59.60 (+7.66)	48.33 (+10.37)	57.70 (+4.23)
0.7	50.99 (+13.36)	60.03 (+6.12)	49.44 (+9.92)	58.62 (+3.62)
0.8	51.05 (+8.99)	60.68 (+5.06)	51.05 (+8.99)	60.01 (+3.90)
0.9	51.92 (+5.81)	61.05 (+2.97)	51.30 (+4.54)	60.25 (+1.62)

Table 7 MRR1 and MRR20 Performances of Two Cluster-based Language Models

Table 7 also shows that the performances of two cluster-based language models are higher than that of the baseline system under different measurements. For MRR1 scores, the largest improvements of cluster-based language models based on One-Sentence-Multi-Topics and One-Sentence-One-Topic are about 15% and 10% respectively. For MRR20, the largest improvements are about 7% and 4% respectively.

Conclusion 1: The experiments show that the proposed cluster-based language model can improve the performance of sentence retrieval in Chinese question answering under the various measurements. Moreover, the performance of clustering-based language model based on One-Sentence-Multi-Topics is better than that based on One-Sentence-One-Topic.

4.3 The Analysis of Sentence Clustering for Various Question Types

The parameter β in equation (3) denotes the balancing factor of the cluster model and the collection model. The larger β, the larger contribution of the cluster model. The small β, the larger contribution of the collection model. If the performance of sentence retrieval decreased with the increasing of β, it means that there are many noises in sentence clustering. Otherwise, sentence clustering is satisfactory for cluster-based language model. So the task of this experiment is to find the performances of sentence clustering for various question types, which is helpful to select the most proper β to obtain the best performance of sentence retrieval.

With the change of β and the fixed α ($\alpha = 0.9$), the performances of cluster-based language model based on One-Sentence-Multi-Topics are shown in Figure 3.

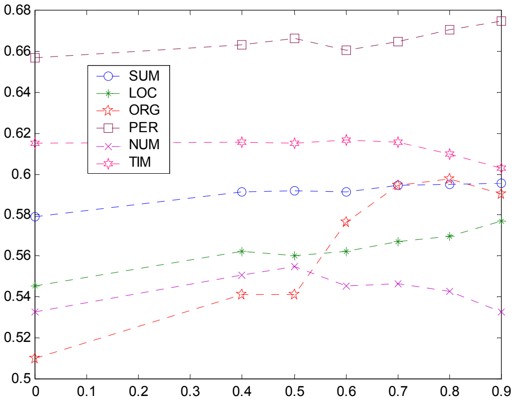

Figure 3 MRR5 Performances of Cluster-based Language Model Based on One-Sentence-Multi-Topics with the Change of β

In Figure 3, the performances of TIM and NUM type questions decreased with the increasing of the parameter β (from 0.6 to 0.9), while the performances of LOC, PER and ORG type questions increased. This phenomenon showed that the performance of sentence clustering based on One-Sentence-Multi-Topics for TIM and NUM type questions is not as good as that for LOC, PER and ORG type questions. This is in fact reasonable. The number and time words frequently appeared in the sentence, which does not represent a cluster/topic when they appear. While PER, LOC and ORG entities can represent a topic when they appeared in the sentence.

Similarly, with the change of β and the fixed α (α=*0.9*), the performances of cluster-based language model based on One-Sentence-One-Topic are shown in Figure 4.

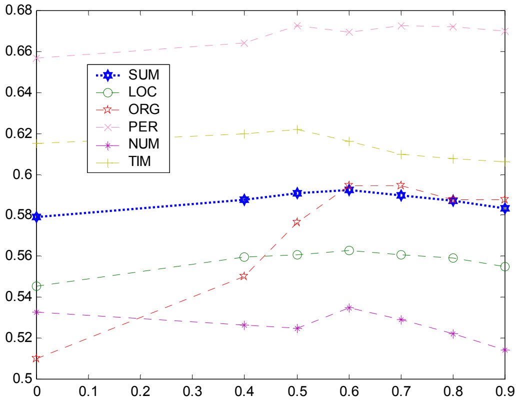

Figure 4 MRR5 Performance of Cluster-based Language Model Based on One-Sentence-One-Topic with the Change of β

In Figure 4, the performances of TIM, NUM, LOC and SUM type questions decreased with the increasing of β (from 0.6 to 0.9). This phenomenon shows that the performances of sentence clustering based on One-Sentence-One-Topic are not satisfactory for most of question types. But, compared to the baseline system, the cluster-based language model based on this kind of sentence clustering can still improve the performances of sentence retrieval in Chinese question answering.

Conclusion 2: The performance of the proposed sentence clustering based on One-Sentence-Multi-Topics for PER, LOC and ORG type questions is higher than that for TIM and NUM type questions. Thus, for PER, LOC and ORG questions, we should choose the larger β value (about 0.9) in cluster-based language model based on One-Sentence-Multi-Topics. While for TIM and NUM type questions, the *value of β should be smaller (about 0.5). But, the performance of sentence clustering based on One-Sentence-One-Topic for all questions is not ideal, so the value for cluster-based language model based on One-Sentence-One-Topic should be smaller (about 0.5) for all questions.*

5 Conclusion and Future Work

The input of a question answering system is natural language question which contains richer information than the query in traditional document retrieval. Such richer information can be used in each module of question answering system. In this paper, we presented a novel cluster-based language model for sentence retrieval in Chinese question answering which combines the sentence model, the cluster/topic model and the collection model.

For sentence clustering, we presented two approaches that are One-Sentence-Multi-Topics and One-Sentence-One-Topic respectively. The experimental results showed that the proposed cluster-based language model could improve the performances of sentence retrieval in Chinese question answering significantly.

However, we only conduct sentence clustering for questions, which have the property that their answers are named entities in this paper. In the future work, we will focus on all other type questions and improve the performance of the sentence retrieval by introducing the structural, syntactic and semantic information into language model.

Reference

J. Ponte, W. Bruce Croft. A Language Modeling Approach to Information Retrieval. In the Proceedings of ACM SIGIR 1998, pp 275-281, 1998.

C. Zhai, J. Lafferty. A Study of Smoothing Techniques for Language Modeling Applied to ad hoc Information Retrieval. In the Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval, 2001.

Ittycheriah, S. Roukos. IBM's Statistical Question Answering System-TREC 11. In the Eleventh Text Retrieval Conference (TREC 2002), Gaithersburg, Maryland, November 2002.

Hui Yang, Tat-Seng Chua. The Integration of Lexical Knowledge and External Resources for Question Answering. In the Proceedings of the Eleventh Text REtrieval Conference (TREC'2002), Maryland, USA, 2002, page 155-161.

Andres Corrada-Emmanuel, W.Bruce Croft, Vanessa Murdock. Answer Passage Retrieval for Question Answering. In the Proceedings of the 27th Annual International Conference on Research and Development in Information Retrieval, pp. 516 – 517, 2004.

Ellen M. Voorhees. Overview of the TREC 2004 Question Answering Track. In Proceedings of the Twelfth Text REtrieval Conference (TREC 2004), 2004.

Vanessa Murdock, W. Bruce Croft. Simple Translation Models for Sentence Retrieval in Factoid Question Answering. In the Proceedings of the SIGIR 2004 Workshop on Information Retrieval for Question Answering, pp.31-35, 2004.

Thomas Hofmann. Probabilistic Latent Semantic Indexing. In the Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval, 1999.

A. Berger and J. Lafferty. Information Retrieval as Statistical Translation. In the Proceedings of ACM SIGIR-1999, pp. 222—229, Berkeley, CA, August 1999.

A. Echihabi and D.Marcu. A noisy-channel approach to question answering. In the Proceeding of the 41st Annual Meeting of the Association for Computational Linguistics, Sappora, Japan, 2003.

Leif Azzopardi, Mark Girolami and Keith van Rijsbergen. Topic Based Language Models for ad hoc Information Retrieval. In the Proceeding of IJCNN 2004 & FUZZ-IEEE 2004, July 25-29, 2004, Budapest, Hungary.

Jian-Yun Nie. Integrating Term Relationships into Language Models for Information Retrieval. Report at ICT-CAS.

Jianfeng Gao, Jian-Yun Nie, Guangyuan Wu and Guihong Cao. 2004b. Dependence language model for information retrieval. In SIGIR-2004. Sheffield, UK, July 25-29.

Youzheng Wu, Jun Zhao, Bo Xu. Chinese Named Entity Recognition Model Based on Multiple Features. In the Proceeding of HLT/EMNLP 2005, Vancouver, B.C., Canada, pp.427-434, 2005.

Youzheng Wu, Jun Zhao, Xiangyu Duan and Bo Xu. Building an Evaluation Platform for Chinese Question Answering Systems. In Proceeding of the First National Conference on Information Retrieval and Content Security. Shanghai, China, December, 2004.(In Chinese)