NCU in Bilingual Information Retrieval Experiments at NTCIR-6

Yu-Chieh Wu
Department of Computer Science

National Central University
No. 300, Jhonda Rd., Jhongli City, Taoyuan County 32001, Taiwan
bcbb@db.csie.ncu.edu.tw

Kun-Chang Tsai
Department of Computer Science

National Central University
No. 300, Jhonda Rd., Jhongli City, Taoyuan County 32001, Taiwan
turninto@db.csie.ncu.edu.tw

Jie-Chi Yang
Graduate Institute of Network Learning Technology
National Central University
No. 300, Jhonda Rd., Jhongli City, Taoyuan County 32001, Taiwan
yang@cl.ncu.edu.tw

Abstract

In this paper, we present the mono-lingual and bilingual ad-hoc information retrieval experimental results at NTCIR-6. This year we compare two different word tokenization levels for indexing, namely, unigram, and overlapping bigram. The two famous information retrieval models, i.e., language model, and BM-25 were adopted in our study. In the mono-lingual results show that our method achieved the average most runs, while the overlapping bigrams were indexed. The unigram level of words did the almost poor results in all runs. In the bilingual retrieval tasks, we translate the queries through a well-known machine translation tool. The evaluation results of our method were also given in the tail of this paper.

Keywords: text retrieval, Chinese information retrieval, Chinese word segmentation, bilingual information retrieval.

1 Introduction

The goal of information retrieval aims at finding the relevant documents in response to the user's queries. In the ad-hoc retrieval tasks, the query is often different from the conventional short query which might contain several words. Instead, a sufficient description (similar to a document) is used. More precisely, the target of ad-hoc retrieval is to provide the ranked document list that can answer the questions or descriptions according to the user's profile (query).

Generally speaking, the information retrieval models can be classified into the three types, Boolean models (Salton, 1989) [12], vector-space models (INQUERY), probabilistic models (Song and Croft, 1999 [13]; Robertson et al., 2001 [10]). Recently, the language model (Zhai, and Laffterty, 2001 [17]; Lavrenko, and Croft, 2001 [4]) and the Okapi BM-25 (Robertson et al., 2001 [10]; Savoy, 2005 [11]) showed excellent performance in many information retrieval tasks, such as TREC ad-hoc retrieval, and NTCIR-5. Although the BM-25 models had been fully investigated and applied in many Far-East languages [11], it is still not compared with the language model approaches at Asian information retrieval tasks.

In this paper, we focus on reporting the experimental results and system descriptions of our retrieval model at NTCIR-6. The two famous information retrieval models, i.e., language model, and BM-25 were adopted in our study. In the mono-lingual results show that our method achieved the average most runs, while the overlapping bigrams were indexed. The unigram level of words did the almost poor results in all runs. In the bilingual retrieval tasks, we translate the queries through a well-known machine translation tool. Our model achieved 0.2439/0.2284 at C-C-T/D task, 0.2561/0.2539 at J-J-T/D, 0.1786/0.1563 at E-C-T/D, 0.1992/0.1816 for J-C-T/D, and 0.1117/0.0868 for K-C-T/D tasks. The evaluation

results of our method were also given in the tail of this paper.

The remainder of this paper is organized as follows: Section 2 describes our indexing and retrieval technologies, in Section 3, the query operation strategies are given, while in Section 4, we report the experimental results on the two tasks, mono-lingual and bilingual results. In Section 5, we draw the conclusion and future works.

2 System Description

Basically speaking, the information retrieval systems contain two components, namely indexer and the ranker. The former builds the keyword index at offline stage, while the latter retrieves the relevant documents based on comparing the given queries with all documents in the collection. These components are discussed as follows.

2.1 Chinese word segmentation

Different from many western languages, there are no explicit boundaries between words in most far-eastern languages, such Chinese, Japanese, and Korean. Hence, almost the first step of Asian language processing technologies should resolve the word segmentation problems first (this is quite similar to the English word tokenization step).

There are two different ways for segmenting Chinese words, one is to apply a fixed-length word extraction (Wu, Lin, & Yang, 2006 [15]; Wu and Yang, 2007 [16]), and the other is to employed a well-trained word segmentation tool (Levow, 2006 [6]). The fixed length word extraction approach defines the fixed word length, for example one (unigram), two (bigram), etc. Thus, a series of continuous tokens are grouped as a word. Usually, the overlapping bigram level of words was drastically and successfully employed in many Chinese information retrieval researches (Savoy, 2005 [11]; Min et al., 2005 [8]; Chen et al., 2005 [2]; Wu, Lin, & Yang, 2006 [15]; Wu and Yang, 2007 [16]).

The second type of treating words is to adopt a well-trained word segmentation tool [6] [15]. However, developing a robust word tokenizer often requires million of labeled texts, which is a laborious work. When porting to different monolingual retrieval, the word segmenter should be retrained. In other words, another annotated corpus should be re-labeled.

An example of different word segmentation strategy can be found in Table 1. In this paper, we focus on the unigram and bigram levels since in different languages, the word segmentation tools are difficult acquired.

Table 1: An example of different word segmentation strategy

Original sentence	查詢建設國際太空站計畫相關的文章
Unigram	查 詢 建 設 國 際 太 空 站 計 畫 相 關的 文 章
Bigram (over-lapping)	查詢 詢建 建設 設國 國際 際太 太空 空站 站計 計畫 畫相 相關 關的 的文 文章
Word-level	查詢 建設 國際 太空站計畫 相關的 文章

2.2 Indexing and Retrieval

As described above, we use two different retrieval models, Okapi BM-25, and the language model-based. For the language model method, we mainly employed a well-known toolkit-lemur[1] for indexing and retrieval. We additionally developed our own BM-25 models by slightly modification the scoring functions to speed up retrieval and indexing. In the following parts, the two methods are briefly described as follows.

BM-25

The well-known Okapi BM-25 weighting scheme is one of the top-performed ranking methods for document retrieval. More and more information retrieval studies further employed this method to Q/A systems and showed the effectiveness to retrieve documents and even the short passages (Tellex et al., 2003 [14]). In this paper, we replicate this retrieval model with the empirical settings that were observed by previous literatures [10] [11] for parameters. However, to make the retrieval stage more efficiently, we re-write the BM-25 term weighting schema. Equation (1), (2), and (3) list the original BM-25 scoring functions.

$$\text{Document Score}(D) = \sum_{i=1}^{|Q|} W^{(1)} \frac{(k_1+1)tf(t_i,D)}{K+tf(t_i,D)} \frac{(k_3+1)tf(t_i,Q)}{k_3+tf(t_i,Q)} \quad (1)$$

[1] http://www.lemurproject.org/

$$W^{(1)} = \log(\frac{N - \mathrm{DF}(t_i) + 0.5}{\mathrm{DF}(t_i) + 0.5}) \quad (2)$$

$$K = (1-b) + b \times \frac{|\mathrm{D}|}{\mathrm{AVG}(|\mathrm{D}|)} \quad (3)$$

k_1, b, k_3 are constants, which empirically set as 1.2, 0.75, 500 respectively [10] [11]. *tf*(t_i, *D*) represent the term frequency of term t_i in document *D*, and *tf*(t_i, *Q*) represent the term frequency of term t_i in query *Q*. *N* denotes as the number of document in the collection. Equation (2) is merely a variant estimation of the "inverse document frequency" which can be observed after the all words were indexed. The third term in equation (1) should be computed in query stage which could be observed offline. However, the term $\frac{(k_1+1)tf(t_i,D)}{K + tf(t_i,D)}$ involves in measuring the average document length and the length of document D should be taken care. We now re-write equation (3) as follows.

$$\begin{aligned} K &= k_1((1-b) + b \times \frac{|\mathrm{D}|}{\mathrm{AVG}(|\mathrm{D}|)}) \\ &= k_1(1-b) + k_1 \times |D| \times b \times \frac{1}{\mathrm{AVG}(|\mathrm{D}|)} \\ &= c_1 + c_2 \times |D| \end{aligned}$$

Where $c_1 = k_1(1\text{-}b)$ and $c_2 = k_1 * b * 1/\mathrm{AVG}(|D|)$. As mentioned above, k_1, b, k_3 were fixed constants that are unchanged during indexing, while c_2 is also a fixed constant after the statistics AVG(|*D*|) is found. By means of the above equations, we can efficiently compute the first term in equation (1) via conventional indexing techniques. In this way, the first two terms of equation (1) can be stored in the indexed files. Once the query input to the retrieval system, the final ranked list is merely involved in computing the third term in equation (1) once and retrieving the indexed files. In our closed experiments, the retrieval time per query is less than 0.2 second since the computing of equation (3) at online stage is unnecessary.

Language Model

Recently, the language model-based ranking models had been shown the state-of-the-art retrieval performance on ad-hoc retrieval task [17] [2] [4]. This method used the P(Q|D) which estimates the conditional probability given D to rank documents. Usually, P(D|P) employs the unigram language model based on the KL-divergence (relative entropy) estimation. Documents are ranked according to the negative of the divergence of the question language model from the document language model. The KL-divergence estimation involves computing the probability for each question word (see (4)), i.e,

$$\begin{aligned} \mathrm{P}(\mathrm{QW}_i \mid \mathrm{D}) = (1-\lambda) \times \mathrm{P}_{\mathrm{smooth}}(\mathrm{QW}_i \mid \mathrm{D}) \\ + \lambda \times \mathrm{P}(\mathrm{QW}_i \mid \mathrm{Collection}) \end{aligned} \quad (4)$$

where $\mathrm{P}_{\mathrm{smooth}}(\mathrm{QW}_i \mid \mathrm{D})$ is the smoothed question word probability estimation given document D, while $\mathrm{P}(\mathrm{QW}_i \mid \mathrm{Collection})$ denotes the MLE estimation of QW_i from the collection. λ is a parameter that controls the importance of the unknown words in the given passage. In this paper, we adopt the well-known language model-based information retrieval toolkit, lemur [2] to compute the KL-divergence for each passage and the given question. We use the parameter settings that were found to be effective in English and Chinese document retrieval tasks [17] [2] [4] for the language model.

3 Query Operation

In this section, we firstly describe the query expansion techniques used in this paper. In Section 3.2, the query translation modules are discussed.

3.1 Relevance Feedback

The CLIR tasks this year involves in two grained size of query descriptions, one is T-run which shall make use of the "title only" as queries, while the N-run means that narrative descriptions could be adopted, which are usually more detail than T-run. A sample of query document can be found at Figure 1.

For the T-run, we can see that most of the queries are exactly short which easily cause out-of-vocabulary (OOV) problems on the initial retrieval set. One solution is to apply the blind query expansion techniques to extract more lexical words to enhance the coverage of the initial query terms. The Lemur language model had included the so-called "two-stage" pseudo feedback for the query term re-adjustment. For the BM-25 we developed

[2] http://www.lemurproject.org/

integrated the Rocchio formula extracting the top-*M* terms from Top-*N* initial retrieved documents. The Rocchio blind feedback is measured in the following way:

$$\mathrm{QW}_i' = \mathrm{QW}_i + \frac{\beta}{N}\sum_{i=1}^{N} D_i - \frac{\gamma}{U}\sum_{j=1}^{U} D_j \quad (5)$$

Where β, γ (set as 2.0, and 1.0) are parameters that control the impact of pseudo positive documents (usually the Top-*N* initial retrieved documents are set to be the positive examples) and negative articles (while the remaining non Top-*N* documents were treated as negative set). By means of the two disjoint document set, the top-*M* words selected from the top-*N* documents are obtained and appended to the original query words. Next, the new query is again used to retrieve the final relevant documents.

```
<TOPIC>
<NUM>048</NUM>
<ONUM>NTCIR4-048</ONUM>
<SLANG>JA</SLANG>
<TLANG>CH</TLANG>
<TITLE>國際太空站，建設</TITLE>
<DESC>查詢建設國際太空站計畫相關的文章。</DESC>
<NARR>
<BACK>基於和平目的，日本、歐洲、俄羅斯和美國進行國際太空站國際合作計畫。工程在海拔 400 公里高度的地球軌道上進行，太空站執行的研究包含新材料開發、生物科學以及太空與地球觀測等。</BACK>
<REL>詳細報導國際太空站計畫規劃或建造過程的文章視爲相關。討論如何使用國際太空站，及其所進行的實驗項目和碰到的困難，或如何評估這些事項的文章視爲部分相關。文章主要在討論其他主題，只簡單提到國際太空站建造規劃的話視爲不相關。</REL>
</NARR>
<CONC>國際太空站，建設，國際合作計畫，太空基地合作計畫，太空開發，日本宇宙事業開發團（National Space Development Agency of Japan），火箭，太空實驗</CONC>
</TOPIC>
```

Figure 1: A sample NTCIR-6 testing topic for Chinese information retrieval

3.2 Machine Translation

In the bilingual information retrieval tasks, the query and collections may be in different languages. Thus, to make them consistently two translation strategies are frequently used, i.e., document translation, and query translation. The two strategies could apply the same translator for different purpose. One is to translate the whole document set, while the other simply converts the query sentences into another language. Usually, the query translation is much more tractable, and cheaper than the document translation since document set might take several GB size of storage spaces.

In this paper, we simply use several online translation tools[3] to translate the queries into the target languages. All of the queries were auto-converted into the specified target languages through these web sites. An example of query translation can be found in Figure 2. However, even these translation tools can successfully translate most of sentences, for some proper nouns, such as person names, project names, etc. it still need to be disambiguated. In this year, we do not employ the disambiguation methods to solve the proper noun translation problems since it requires some additional training corpus, and learners.

Original Chinese Query	基於和平目的，日本、歐洲、俄羅斯和美國進行國際太空站國際合作計畫。工程在海拔 400 公里高度的地球軌道上進行，太空站執行的研究包含新材料開發、生物科學以及太空與地球觀測等。詳細報導國際太空站計畫規劃或建造過程的文章視爲相關。討論如何使用國際太空站，及其所進行的實驗項目和碰到的困難，或如何評估這些事項的文章視爲部分相關。文章主要在討論其他主題，只簡單提到國際太空站建造規劃的話視爲不相關。國際太空站，建設，國際合作計畫，太空基地合作計畫，太空開發，日本宇宙事業開發團（National Space Development Agency of Japan），火箭，太空實驗
Auto-translated into English	Based on the peaceful purpose, Japan, Europe, Russia and US carry on the international space station international cooperation plan. The project carries in the elevation 400kilometer high earth's orbit, the space station execution research contains the new material development, the biological science as well as the outer space and the Earth observes and so on. The detailed report international space station plan or the construction process article regards as the correlation. How discusses uses the international space station, and its carries on the experimental project and bumps into did the difficulty, how or appraise these items the article regards as the partial correlation. The article main is discussing other subjects, only simply mentioned the international space station construction plan the speech regards as not related. The international space station, the construction, the international cooperation plan, the outer space base corporate plan, the outer space development, the Japanese universe enterprise develops the group (National Space Development Agency of Japan), rocket, outer space experiment

Figure 2: A sample of online web-page translation using systran web service

4 Experiments

[3] http://www.systranbox.com/
http://www.excite.co.jp/world/korean/
http://www.alphaworks.ibm.com/aw.nsf/html/mt

The details about the query dataset and document collections were completely described in NTCIR-CLIR task definition [4]. The experiments were evaluated based on TREC-like procedure. Details can be found in (Kishida et al., 2007). Due to the limited number of submission runs, we submit the two different runs with Model1) bigram+language model, and Model2) unigram+BM-25 for the D and N runs. For the DN run, we used bi-gram+language model strategy, i.e., Model 1. The parameter setting of the parameters for Model1 is the same as Lemur toolkit, while the setting of Model2 is the same as described above.

4.1 Mono-lingual retrieval results

The results for the monolingual retrieval task in Chinese are listed in Table 2. The global comparisons to the monolingual Chinese information retrieval task can be found at Table 3. As seeing in Table 2, it is clearly that the bigram+language model significantly outperformed the BM-25+unigram level, while in Table 3, we can see that our method achieved the "slightly" better than the middle rank. It is very encourage that the use of simple bigram+ language model is very effective. We also continue our experiments with combining the BM-25+bigram level of words, and also found the improvement over than the original unigram level. But the performance of the bigram+BM-25 was still worse than the language model in 1~2% in average precision.

Tabel 2: Relaxed and Rigid relevance scores for mono-lingual Chinese information retrieval task

Relaxed relevance score (Chinese-Chinese)		
Run	R-precision	AvgP
NCUTW-C-C-T-01	0.2861	0.2417
NCUTW-C-C-T-02	**0.3666**	**0.3455**
NCUTW-C-C-D-03	0.2838	0.2389
NCUTW-C-C-D-04	**0.3500**	**0.3385**
NCUTW-C-C-DN-05	0.3550	0.3151
Rigid relevance score (Chinese-Chinese)		
Run	R-precision	AvgP
NCUTW-C-C-T-01	0.2157	0.1696
NCUTW-C-C-T-02	**0.2639**	**0.2439**
NCUTW-C-C-D-03	0.2004	0.1561
NCUTW-C-C-D-04	**0.2573**	**0.2284**
NCUTW-C-C-DN-05	0.2720	0.2266

[4] http://homepage3.nifty.com/kz_401/index.htm

Tabel 3: Global results for the C-C task

C-C-T	Relaxed	Rigid
	AvgP	
Min	0.1468	0.1146
Avg	0.3213	0.2320
Max	0.4090	0.3097
NCUTW	0.3455	0.2639
C-C-D	Relaxed	Rigid
	AvgP	
Min	0.2389	0.1561
Avg	0.3339	0.2378
Max	0.4118	0.3136
NCUTW	0.3385	0.2284

The results for the monolingual retrieval task in Japanese are listed in Table 4. The global comparisons to the monolingual Japanese information retrieval task can be found at Table 5. In the J-J task, our method also reached the middle rank.

Tabel 4: Relaxed and Rigid relevance scores for mono-lingual Japanese information retrieval task

Relaxed relevance score (Japanese-Japanese)		
Run	R-precision	AvgP
NCUTW-J-J-T-01	0.2866	0.3060
NCUTW-J-J-T-02	**0.3438**	**0.3417**
NCUTW-J-J-D-03	0.3229	0.3044
NCUTW-J-J-D-04	**0.3494**	**0.3486**
NCUTW-J-J-DN-05	0.3299	0.3012
Rigid relevance score (Japanese-Japanese)		
Run	R-precision	AvgP
NCUTW-J-J-T-01	0.2252	0.2102
NCUTW-J-J-T-02	**0.2637**	**0.2561**
NCUTW-J-J-D-03	0.2450	0.2174
NCUTW-J-J-D-04	**0.2638**	**0.2539**
NCUTW-J-J-DN-05	0.2254	0.2463

Tabel 5: Global results for the J-J task

J-J-T	Relaxed	Rigid
	AvgP	
Min	0.1955	0.1560
Avg	0.3427	0.2707
Max	0.4393	0.3600
NCUTW	0.3417	0.2561
J-J -D	Relaxed	Rigid
	AvgP	

Min	0.2249	0.1768
Avg	0.3214	0.2479
Max	0.4138	0.3255
NCUTW	0.3486	0.2539

4.2 Bilingual retrieval results

The experimental results of our method on the bilingual retrieval tasks can be found at Table 6, Table 7, Table 8, and Table 9. In this competition, there were some font converting errors within the other bilingual retrieval tasks, such as J-J, E-J, C-J, etc. Thus, we only report the actual performances that were reliable to be presented. In this task, we use the language model+bigram approach as described above. All of the queries were auto-translated into the target languages via the online translation resources.

In these experiments, we found that the unigram+language model approach did not perform well as previous experiments. Most of our method did achieve the slightly worse than middle rank. Since we did not perform the proper noun disambiguation for the bilingual retrieval task, most of our runs were not comparable to those perform disambiguation participants. In the future, we plan to integrate more translation technologies to improve such a simple translation approach.

Tabel 6: MAP scores for C-C task in bilingual retrieval task

Run	Relax	Rigid
NCUTW-C-C-D-02-N3	0.2424	0.1908
NCUTW-C-C-D-02-N4	0.1546	0.1205
NCUTW-C-C-D-02-N5	0.3346	0.2746
NCUTW-C-C-T-01-N3	0.2591	0.2039
NCUTW-C-C-T-01-N4	0.1725	0.1424
NCUTW-C-C-T-01-N5	0.3858	0.344

Tabel 7: Global results for the E-C task

Run	Relax	Rigid
NCUTW-E-C-D-02-N3	0.0937	0.0721
NCUTW-E-C-D-02-N4	0.0739	0.0626
NCUTW-E-C-D-02-N5	0.2063	0.1786
NCUTW-E-C-T-01-N3	0.0741	0.0626
NCUTW-E-C-T-01-N4	0.0927	0.0807
NCUTW-E-C-T-01-N5	0.187	0.1563

Tabel 8: Global results for the J-C task

Run	Relax	Rigid
NCUTW-J-C-D-02-N3	0.1139	0.0851
NCUTW-J-C-D-02-N4	0.1058	0.0826
NCUTW-J-C-D-02-N5	0.2212	0.1816
NCUTW-J-C-T-01-N3	0.1099	0.0857
NCUTW-J-C-T-01-N4	0.1079	0.0881
NCUTW-J-C-T-01-N5	0.2256	0.1992

Tabel 9: Global results for the K-C task

Run	Relax	Rigid
NCUTW-K-C-D-02-N3	0.0467	0.0264
NCUTW-K-C-D-02-N4	0.0739	0.0607
NCUTW-K-C-D-02-N5	0.1151	0.0868
NCUTW-K-C-T-01-N3	0.0392	0.0299
NCUTW-K-C-T-01-N4	0.0747	0.0685
NCUTW-K-C-T-01-N5	0.1206	0.1117

5 Conclusion and Future Remarks

This is our first time to participate the cross language information retrieval task. We mainly developed a BM25 retrieval algorithm and blind query expansion techniques to improve the retrieval performance. We compare two different retrieval models, language model and BM-25 with unigram and bigram grained level of words. The experimental results showed that the bigram of words significantly outperformed the simply unigram segmentation. We trust the unigram level of words is much more nature and simple than the bigram, however the results explains the shortcoming of using unigram and unigram BM-25 models. But the retrieval results could be improved by applying higher order of language models, such as bigram or trigram models.

In the bilingual information retrieval tasks, we here used the online machine translation tool to automatically convert one language into another. We also found that many disambiguation or word miss were mainly caused by the proper nouns, like person names, location names, etc. In the future, we plan to resolve the proper noun disambiguation problems by adopting a phrase chunker [5] and some parallel linguistic corpora. To improve the retrievers, we will focus on developing higher order language model approaches with capturing local dependency relations.

References

[1] Broglio, J., Croft, W. B., Callan, J., & Nachbar, D. (1995). Document retrieval and routing using the INQUERY system. *In Proceedings of the third text retrieval conference.* pp. 500-225.

[2] Chen, J., Li, R., & Li, F. (2005). Chinese information retrieval using Lemur. *In Proceedings of the 5th NTCIR workshop.*

[3] Kishida, K., Chen, K. H., Lee, S., Kuriyama, K., Kando, N., Chen, H. H., & Myaeng, S. H. (2007). Overview of CLIR task at the sixth NTCIR workshop. *In Proceedings of the 6th NTCIR Workshop*.

[4] Lavrenko, V., & Croft, W. B. (2001). Relevant-based language models. *In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. pp. 120-127.

[5] Lee, Y. S., & Wu, Y. C. (2007). A Robust Multilingual Portable Phrase Chunking System. *Expert Systems with Applications*, 33(3): 1-26.

[6] Levow, G. A. (2006). The third international Chinese language processing Bakeoff: word segmentation and named entity recognition. *In Proceedings of the 5th SIGHAN Workshop on Chinese Language Processing*. pp. 108-117.

[7] Liu, X., & Croft, W. B. (2002). Passage retrieval based on language models. *In Proceedings of the 11th international conference on information and knowledge management*. pp. 375-382.

[8] Min, J., Sun, L. & Zhang, J. (2005). ISCAS in English-Chinese CLIR at NTCIR-5. *In Proceedings of the 5th NTCIR workshop meeting on evaluation of information access technologies*.

[9] Ponte, J. M., & Croft, W. B. (1998). A language modeling approach to information retrieval. *In Proceedings of the 21st ACM SIGIR conference on research and development in information retrieval*. pp. 275-281.

[10] Robertson, E., Walker, S., & Beaulieu, M. (2000). Experimentation as a way of life: Okapi at TREC. I*nformation processing & management*. 36(1): 95-108.

[11] Savoy, J. (2005). Comparative study on monolingual and multilingual search models for use with Asian languages. *ACM trans. on Asian language information processing*. 4(2): 163-189.

[12] Salton, G. 1989. *Automatic Text Processing*. New York : Addison Wesley.

[13] Song, F., & Croft, W. B. (1999). A general language model for information retrieval. *In Proceedings of Eighth International Conference on Information and Knowledge Management (CIKM'99)*. pp. 279-280.

[14] Tellex, S., Katz, B., Lin, J. J., Fernandes, A. & Marton, G. (2003). Quantitative evaluation of passage retrieval algorithms for question answering. *In Proceedings of the 26th ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 41-47.

[15] Wu, Y. C., Yang, J. C., & Lin, Q. X. (2006). Description of the NCU Chinese word segmentation and named entity recognition system for SIGHAN Bakeoff 2006. *In Proceedings of the 5th SIGHAN Workshop on Chinese Language Processing*. pp. 209-212.

[16] Wu, Y. C., & Yang, J. C. (2007). Toward Multimedia: A String Pattern-based Passage Ranking Model for Video Question Answering. *In Proceedings of the North American chapter of the Association for Computational Linguistics (HLT-NAACL-07)*. in press.

[17] Zhai, C., & Lafferty, J. (2001). A study of smoothing methods for language models applied to ad hoc information retrieval. In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR). pp. 334-342.