ISCAS in English-Chinese CLIR at NTCIR-5

Jinming Min Le Sun Junlin Zhang
Institute of Software, Chinese Academy of Sciences
P.O.Box 8718, Beijing, 100080, P.R.China
jinming03@iscas.cn

Abstract

We participated in the Chinese single language information retrieval(SLIR) C-C task and English-Chinese cross-language information retrieval(CLIR) E-C tasks in NTCIR5. Our project concentrates on the two aspects of the CLIR research: 1) We test various IR models especially language models for Chinese SLIR using the training corpus provided by the NTCIR organizer, and different smoothing methods have been studied for Chinese SLIR; 2) Our C-E CLIR task is based on the dictionary-based translation approach, and a new context-based translation algorithm using web corpus is proposed to solve the out-of-vocabulary(OOV) problem in CLIR.

Keywords: *CLIR, OOV, Query Translation, Disambiguation.*

1 Monolingual IR for Chinese Language

Usually the process of CLIR experiment has been divided into two steps: 1)Queries translation from the Source-Language to the Target-Language; 2)Single language information retrieval(SLIR) in the target language. Our objective is evaluate the English-Chinese CLIR task which has been shown particularly low performance in previous NTCIR campaign [2], and one effective evaluation method is to compare the CLIR performance with that of SLIR which the manual translation for queries are used from the source language to the target language. And in this project, the manual translation of queries has been economized as the NTCIR organizer has provided the official queries in both Chinese and English. For the comparability of the CLIR and SLIR, the target corpus and SLIR model including IR model and stemming algorithm and stopword list should be promised congruity.

Before directly starting the E-C CLIR experiment, we test various SLIR models and choose the most appropriate model for Chinese language as our official SLIR model in all NTCIR5 task including C-C SLIR and E-C CLIR. In order to test various IR models, we adopt the training document collections provided by NTCIR-5 organizer. Our focus is mainly on the Chinese language SLIR, and the training Chinese document collections are as same as the corpus which has been used in the NTCIR4 companion. To see the detailed introduction of the corpus, please refer to [2]. Also the judgment file for training corpus and queries are provided by the NTCIR5 organizer, so several retrieval evaluation parameters can be calculated at the end of the experiments and it will be shown at section 1.4.

The first section is organized as follows. Section 1.1 will give an overview of the training document collections and the test document collections that we used in this project. Section 1.2 describes our pretreatment process and searching strategy in use. Section 1.3 will explain the main features of various IR models. Finally, the evaluation results of our SLIR training experiments will be given in section 1.4.

1.1 Overview of NTCIR5 training and test documents

For testing various IR models in Chinese SLIR, We choose the NTCIR4 Chinese document collections as our training corpus and NTCIR5 Chinese document collections as our test corpus. And to compare the proformance of the same IR model in different languages, a relatively small English corpus ISCAS01 which is produced manually in our previous research project is introduced here . The following tables show the corpus that we use in the experiment, and the coding of all the Chinese corpus is Big5. Table 1 shows the Chinese training corpus, Table 2 shows the English training corpus, Table 3 shows the NTCIR5 test Chinese document collections.

Table 1. NTCIR5 Chinese Training Corpus

Sources	No. of Docs
CIRB020(United Daily News)	249,508
CIRB011(China Times)	132,173
Total	381,681

Table 2. NTCIR5 English Training Corpus

Sources	No. of Docs
ISCAS01 (Random English Web Documents)	3,204

Table 3. NTCIR5 Chinese Evaluation Documents

Files	No. of Docs
United Daily News	466,564
United Express	92,296
Ming Hseng News	169,739
Economic Daily News	172,847
Total	901,446

As the Table 1 and Table 3 are analyzed, the number of the Chinese documents in NTCIR5 has been added 136.18% comparing with that of NTCIR4. The larger corpus scale can make our experiments more analogous to the real problem. In this whole project, we use an Intel Pentium IV 2.8G computer(memory: 512 MB, swap: 1 GB, disk: 2 $\times$ 80 GB), and operating system is GNU/Linux, and the Linux kernel is version 2.6.

1.2 Pretreatment and searching strategies

Before building index for Chinese documents, word segmentation has been applied on the corpus. A free Java Chinese word segmenter from the mandarintools website[1] is used, and this segmenter is based on a large Chinese word list and the the maximal matching algorithm to segment the Chinese sentences into words. Also a Chinese stopword list, which is produced by us through statistical training on a large Chinese web corpus for one of our previous projects, has been introduced to get rid of the high frequency Chinese words. And this Chinese stopword list contains 89 words all in together.

Through our previous research [11] in language modeling, we plan to test the language modeling retrieval method in the Chinese SLIR. Before choosing the IR models for formal tasks in NTCIR5, we test various classical IR models including four different language modeling methods. Our simple language modeling method is based on the Kullback-Leiback divergence formula as follow, where $p(x), q(x)$ are two probability mass functions which denote the parameters of the query unigram language model and parameters of the document unigram language model respectively:

$$D(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

[1] http://www.mandarintools.com

Simple language model method are tested when combing with three different smoothing methods [9] as Jelinek-Mercer method, Dirichlet method, Absolute discount method. Document language model's fundamental goal is to estimate the $p(w|d)$, where d is the single document and w is a single word. The three smoothing methods are described as the following formulas, where the C is the collection language model, $\lambda\ \mu\ \delta$ is paramete that can be adjusted citesmooth:

- Jelinek-Mercer method:

$$p_s(w|d) = (1-\lambda)p_{ml}(w|d) + \lambda p(w|C)$$

- Dirichlet method:

$$p_s(w|d) = \frac{c(w;d) + \mu p(w|C)}{\sum_w c(w;d) + \mu}$$

- Absolute discount method:

$$p_s(w|d) = \frac{max(c(w;d) - \delta, 0)}{\sum_w c(w;d)} + \frac{\delta|d|_\delta}{|d|} p(w|C)$$

In next section we will test these smoothing methods when they are applied on the Chinese SLIR.

1.3 Comparison of various IR models

Our IR system is based on the Lemur toolkit [3] project by Center for Intelligent Information Retrieval at Umass and Language Technologies Institute at CMU. To evaluate the various IR models, we adopt three evaluation parameters: set AP[2](non-interpolated), set BP[3], average P[4] at 1000 which means the AP when 1000 documents have been retrieved. For the explanation of these parameters, see the following formulas and each topic's breakeven precision is its precision at the rank that is equal to the total Number of Relevant Documents(NRD) for that topic at this point, precision is equal to recall(ie if there are 5 total relevant docs, it's breakeven precision would be its precision at 5 docs).

$$Set\ AP\,(non-interpolated) = \frac{Sum\ of\ AP\ of\ Each\ Topic}{The\ Number\ of\ Topics} \quad (1)$$

$$Each\ topic's\ AP(non-interpolated) = \frac{Sum\ of\ P\ for\ Topic\ at\ Each\ Rank}{NRD\ for\ Topic} \quad (2)$$

$$Set\ BP = \frac{Sum\ of\ BP\ for\ Topic}{Number\ of\ Topics} \quad (3)$$

[2] average precision
[3] breakeven precision
[4] precision

The various IR models which are adopted in our experiments are as follows: simple TFIDF retrieval(**simple-tfidf**), and the score of document d against query q is given by $s(d,q) = \sum_{i=1}^{n} tf_d(x_i)tf_q(y_i)idf(t_i)^2)$, where x_i, y_i is the element word t_i frequency in the document and query respectively; TFIDF feedback retrieval(**fb-tfidf**), where the simplified Rocchio feedback algorithms has been used in the relevance feedback; simple Okapi retrieval(**simple-okapi**), which is the same classical Okapi model as TREC 3 [8]; Okapi feedback retrieval(**fb-okapi**), where Okapi model is combined with the relevance feedback; Simple language model(KL-divergence) retrieval, using JM smoothing(**simple-kl-jm**); Simple language model(KL-divergence) retrieval, using Dirchlet smoothing(**simple-kl-dir**); Simple language model (KL-divergence) retrieval example, using absolute discounting smoothing(**simple-kl-abs**); Language modeling(KL-divergence) feedback retrieval,using collection mixture method and Dirichlet smoothing(**mixfb-kl-dir**), where the collection mixture method means that a word is picked according to the collection language model, when a feedback document is "generated"; The re-ranks res-simple-tfidf with the tf-idf method(**rerank-simple-tfidf**), where rerank means that one can re-rank a subset of docs when the scoring method is computationally complex; Thes re-ranks res-simple-tfidf with KL-divergence and Dirichlet smoothing; At last, Two-Stage language models [10] method is applied in the experiment, where in the first stage the document language model is smoothed using a Dirichlet prior with the collection language model as the reference model, in the second stage the smoothed document language model is further interpolated with a query background language model(**twostage**).

1.4 Evaluation results of IR models

Table 4 shows the experiment results of various IR models when they are applied on the large Chinese training corpus and small English training corpus. We can draw the following conclusions from the analysis of the experiment datas: 1) Relevance feedback is a two-edge technology in improving precision of the IR system. When combined with the classical tfidf model, IR model with relevance feedback get highest set AP in the English training corpus. But when combined with Okapi model, the relevance feedback actually does harm to the precision. That's the state of the English training corpus, in Chinese language the relevance feedback does improve the precision in all IR models, so effects of the relevance feedback in IR system is still unclear in our experiments. 2) As we have expected before experiments, language modeling retrieval method is the most promising model in the near future, our best retrieval result is got from the mixfb-kl-dir model in Chinese IR. And by the more, the smoothing methods play an very important role in the language modeling IR system. Different smoothing methods usually affect the retrieval precision greatly and dirichlet smoothing method produce higher performance for both English and Chinese in our experiments. 3) Different IR models get almost comparable precision results in both Chinese and English, except the high performance of simple-tfidf in the English training corpus. This conclusion is accorded with the statistical feature of all of our IR systems.

For the official NTCIR C-C task, we adopt the same pretreatment method as we have used in the training experiments. Using different parts of the official queries, two groups of results as ISCAS-C-C-D-01 and ISCAS-C-C-T-01 are submitted to the NTCIR5 organizer. We lastly choose the mixfb-kl-dir model for our NTCIR5 official C-C retrieval task, the evaluation result of the NTCIR5 C-C task will be shown at table 5.

2 English-Chinese CLIR

CLIR research has been divided into three main approaches as machine translation based approach, parallel corpus based approach and dictionary based approach. Taking acount of the easy acquire of the translation resources, dictionary based approach is the most promising method when applied to the realistic industry system. But the Machine Readable Ditionary(MRD) usually doesn't cover all the English query words and this is called Out-Of-Vocabulary(OOV) problem in CLIR research, and the web corpus is the biggest corpus that ever emerged in spit of it's hardness to use. Usually we can find the English OOV words online and Chinese counter-part by human, but how to extract those words automatically is still a big challenge to researchers. Through analysis of the web documents, we developed a new method of extracting English OOV's Chinese counterpart online, and the result of the translation show that we have succeeded in a reliable extent. And in all, We adopt the dictionary-based query translation approach for E-C CLIR task. After we have collect the translation for all English query words, a search engine based disambiguation algorithm is used in our CLIR translation system. At last, the translated queries are sent to the SLIR system which adopt the same IR model that we use in the C-C SLIR task.

This section will organized as follows, in section 2.1 we will introduce our pretreatment method of the English queries and the dictionary based query translation is described including the dictionary resource that we use, in section 2.2 a new OOV translation method is given and, in section 2.3 we introduce our search engine based disambiguation algorithm, in section 2.4 we give our evaluation results of the experiments.

Table 4. IR Models Comparison Experimental Results

IR Models	Set Average Prec (non interpolated)		Set Breakeven Prec		Average Prec at 100	
	E	C	E	C	E	C
simple-tfidf	0.2556	0.2075	0.2715	0.3292	0.0678	0.8100
fb-tfidf	**0.2736**	0.2289	**0.2833**	0.3517	**0.0723**	0.8448
simple-okapi	0.2377	0.2317	0.2544	0.3491	0.0656	0.8834
fb-okapi	0.2060	0.2441	0.2186	**0.3627**	0.0603	**0.8978**
simple-kl-jm	0.2458	0.2458	0.2602	0.3415	0.0617	0.8405
simple-kl-dir	0.2566	0.2319	0.2756	0.3506	0.0637	0.8721
simple-kl-abs	0.1849	0.2193	0.2076	0.3370	0.0505	0.8669
mixfb-kl-dir	0.2725	**0.2524**	**0.2883**	**0.3701**	0.0691	**0.9036**
rerank-simple-tfidf	0.2556	0.2076	0.2715	0.3295	0.0678	0.8114
rerank-simple-kl-dir	0.2578	0.2240	0.2756	0.3366	0.0637	0.8724
twostage	0.2288	0.2029	0.2609	0.3188	0.0575	0.8395

Table 5. NTCIR5 C-C Task Evaluation Result

NTCIR5 Task	Average Precision	R-Precision	Precision at 100 docs
ISCAS-C-C-D-01(Relax)	**0.4632**	**0.4591**	0.3020
ISCAS-C-C-D-01(Rigid)	0.3963	0.3910	0.2026
ISCAS-C-C-T-01(Relax)	0.4244	0.4233	0.2876
ISCAS-C-C-T-01(Rigid)	0.3480	0.3503	0.1850

2.1 Pretreatment

We first parse all the English queries, and the NTCIR complete query is consisted by TITLE,DESC,NARR and CONC part. The DESC and NARR parts are usually complete English sentences. As we preprocess the English corpus before, we use the stopword list that are provided by the SMART IR system to get rid of the high frequency words. For the look-up in the bilingual dictionary, the porter stemming algorithms [7] is used to stem all the English query words as the bilingual Machine Readable Dictionary doesn't contain any plural information. Our basic translation resource is the E-C bilingual dictionary which is provided by LDC, and this dictionary contains 110,843 entries of the English words and Chinese counter-part. After pre-translation of the English queries, we filter the English words that can't be translated by the ecdict dictionary, and these words are marked as Out-Of-Vocabulary such as "Belgrade","Nago","Albuquerque","ILOVEYOU". And the pretreatment and OOV translation process is shown in Figure 1.

2.2 A new OOV translation method

Web corpus is a rich source for translation of OOV, much research has been done in the past years [5, 6, 12]. The OOV word can be seen as a special word which can decide the classification of the whole article usually. So if we get two articles in different languages containing the same OOV word, and these articles can be seen as comparable article in two kinds of language. This enlighten us to use the OOV words to download bilingual comparable corpus on the web using search engine. First we use the ldc-ec-dict to produce an OOV list. Through analysis of large volumes of web documents by human, we find the best translation resource for English OOV words is the mixed Chinese and English document on the web, for usually this kind of documents contain both the English OOV words and their counter-part in Chinese. Another important finding is that the English OOV words' contextual words are usually comparable with those of their Chinese counter-part in semantics. Going to the next step, we find that these contextual words in two languages usually accord with the available bilingual dictionary entries. So the basic assumption in our experiments is that the English OOV words and their Chinese counter-part usually concur with same semantical concept which make up of common translation pairs.

Our English OOV translation method are described as the following steps:

- Step 1: send one English OOV word to the web search engine to crawl related 100 documents which contain itself, and the documents are pretreated by parsing, stemming, getting rid of stopwords and every sentence are treated as a chunk;
- Step 2: the English contextual words of the OOV word are calculated in the step 1's output documents using the following formula, we choose the n words with highest score(x,y) in

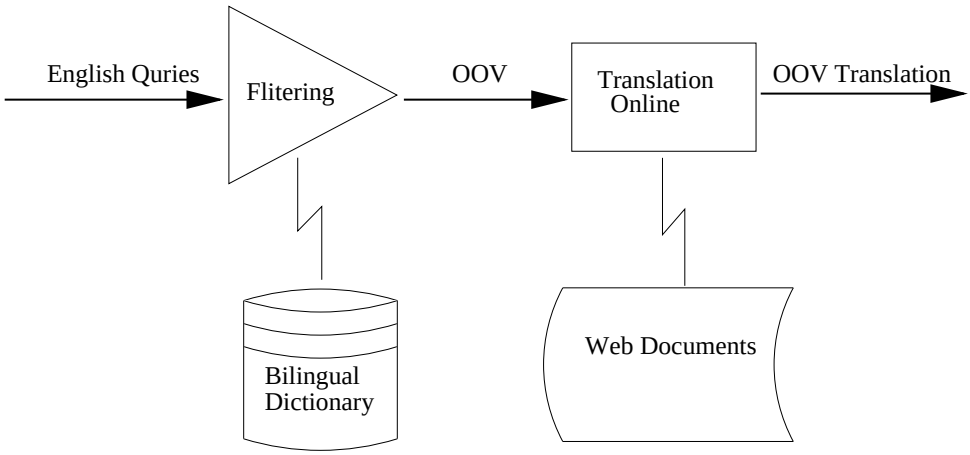

Figure 1. Pretreatment and OOV Translation Process

the formula as the contextual words. In the formula, x indicates every word which concur with the OOV word y in the same chunk, p(x,y) is the probability of the x and y concurring in the whole 100 documents, p(x) or p(y) is the independent probability of x or y appearing in the whole documents, distance(x,y) is the average distance of x and y in various chunk and max_distance is the length of longest sentences in the documents.

$$score(x,y) = [p(x,y) \times \log \frac{p(x,y)}{p(x)p(y)}] \times [\log \frac{max\_distance}{distance(x,y)} + 1] \quad (4)$$

- Step 3: send the English OOV word to the web search engine but confine the target document language as Chinese to crawl 100 related documents which contain itself;
- Step 4: the crawled Chinese and English mixture documents in step 3 are pretreated by segmenter and getting rid of stopword;
- Step 5: the contextual words as step 2's output are translated by bilingual dictionary and a long list of Chinese words has been produced;
- Step 6: the Chinese words list are treated as OOV words in the documents of mixed languages and the step 2 has been re-processed, then a candidate OOV translation words list has been produced;
- Step 7: the candidate translation words in step 6 has been filtered by a large Chinese word list which is produced by extracting the ecdict Chinese words.

The above output in step 7 are treated as our translation candidates of English OOV. After extracting all the translation of OOV online, we add the English OOV words and their extracted Chinese counter-part into the LDC dictionary, then the new dictionary are used to translate all the queries again. So we usually get no new OOV words again for we have extract all the corresponding translations for OOV. But we also can't promise that the extracted OOV translations are correct because all the process is done automatically without human intervention. Through the second dictionary translation process, a large translation candidate group is produced. For example we have the original English query $e_1, e_2, ..., e_n$, and the e_1 has the corresponding Chinese translation $c_1^1, c_1^2, ..., c_1^m$, and the e_n has translation $c_n^1, c_n^2, ...c_n^m$, so we produce the candidate Chinese candidate queries using the following algorithms: For $1 \leq i \leq n$, We choose the first English query word e_i; For $1 \leq j \leq m$ we choose the translation c_i^j, where m is the largest Chinese candidate words number. By this simple O(n*m) algorithms we get a large translation candidate group.

2.3 Search engine based disambiguation

When the above algorithm is adopted, m×n translation candidates are produced. Surely it's not possible to use all the translation candidates as our chosen Chinese queries in the SLIR which is now the second step in the whole E-C CLIR task. We adopt a web search engine based algorithm [4] in our translation system. Many disambiguation methods have been invented for CLIR research, and most of them utilize the words concurrence information by training corpus. The common training corpus exists two obvious drawbacks as: 1) Large training corpus is not always available and the construction of large corpus needs many human resources; 2) Any particular training corpus lacks in the fraction of coverage and won't satisfy the open domain problem. On the contrary, the web corpus is the biggest corpus ever which is always easy to assess and

it's fraction of coverage is compact. So our disambiguation method is also based on the words concurrence information on the web corpus, and the common web search English "Google" is chosen as our portal to web corpus.

The words concurrence information on the web corpus is hard to acquire, but the common web corpus usually will provide the retrieved documents number for a group of queries words. This information has been proved useful in replacing the words concurrence information. Given a source English query which contains the English words $e_1, e_2, ...e_n$, the corresponding Chinese candidate queries are $g_1, g_2, ...g_{m*n}$. The number of the candidate queries is $m * n$ which we have demonstrated in the previous section. All the candidate groups are sent to Google as the queries for web retrieval, and the retrieved web documents number is recorded by our system automatically. At last we choose the candidate group which get the highest number of retrieved documents as the official translation candidate in our SLIR step.

2.4 Results of evaluation

We take part in the NTCIR5 E-C task and submit two groups of experimental results using our described CLIR method above. But the result doesn't accord to our expectation in advance, we will give an analysis in the following part. The results of the official E-C CLIR task evaluation is as Table 5:

As the results show, the ISCAS-E-C-T-01(Relax) has exceeded ISCAS-E-C-D-01(Relax) 73.49% in average precision, 65.28% for the R-Precision, 61.58% for the Precision at 100 docs. As the former introduction describes, we use the same queries translation method and SLIR model in all the official experiments. Through careful analysis of the queries translation method, we find that these experiment's drawback lies on the OOV filtering. We use the porter stemming algorithms for the preprocess of the English queries, and we find that many wrong stemmed words give us OOV filtering system big trouble. For example, the NTCIR5 English Query NO.1's title contain the word "online", after the stemming process we get the new word "onlin", then our OOV filtering system take it as an OOV and try to find it's Chinese translation online. Of course we'll get wrong information from the web extracting. As we have analyzed, our OOV translation method are mostly useful when the OOV is a real "OOV" which contain important semantical meanings. But normal words or words with high frequencies are usually can't satisfy our requirements. So through our system, the common fake OOV words usually greatly do great harm to our retrieval precision. As the results show, the queries containing the description part are usually common English sentences which contain more common words than title part of the queries. So when our method are applied on the title part, we get the expected results relatively. These can explain the big difference in performance of the two official E-C CLIR task.

One of our experiment's goals is to achieve comparable precision in the English-Chinese CLIR with the Chinese SLIR. For the convenient in the compare between CLIR and SLIR, our all official IR models are the same mixfb-kl-dir model as we have introduced in the previous section. As our experiments show, we only obtained 36.55%(0.1551 vs. 0.4244) of the performance level achieved by a monolingual search for the Chinese language in ISCAS-C-C-T-01(Relax). We ascribe these low performance as the following reasons in our methods and system:

- The OOV filtering errors which is caused by the stemming algorithm;
- Lack of Chinese new word boundary detecting in the OOV translation;
- Less competitive disambiguation algorithm is used in the experiment;

3 Conclusions and Future Work

In our work of NTCIR5, we test various IR models based on the NTCIR4 Chinese Corpus and the small English corpus produced by ourselves. Our SLIR experiment results shows relevance feedback is a two-edge technology in IR research, although it has shown good remarks in most situation. So we plan to have more comprehensive experiments on different corpus which is various on the scale and language. The send finding of the SLIR is that when language model retrieval method is applied on the Chinese language, it has very great improve in the precision. Also the smoothing technology plans an very important role in the language model retrieval method. Different smoothing methods take various effects on the precision, and the diriletct smoothing method produce best retrieval precision in our experiments.

The E-C CLIR task has been shown very low precision in the previous NTCIR evaluation results [2], and our CLIR system focus on solving the OOV problem between Chinese and English. In fact, our context based OOV translation method is unrelated with the special language feature, so it can be introduced to the other language pairs. The web corpus can help solve the translation resource poverty problem greatly, and our OOV translation method can solve OOV problem in a definite scale but not all of it. Also we don't combine with the Chinese new words boundary detecting algorithms to produce high accuracy translation, so this is also a promising direction of the CLIR research between Chinese and English, but the Chinese language still lies many open problems needed to be

Table 6. NTCIR5 E-C Task Evaluation Result

NTCIR5 Task	Average Precision	R-Precision	Precision at 100 docs
ISCAS-E-C-D-01(Relax)	0.0894	0.0959	0.0734
ISCAS-E-C-D-01(Rigid)	0.0786	0.0866	0.0504
ISCAS-E-C-T-01(Relax)	**0.1551**	**0.1585**	0.1186
ISCAS-E-C-T-01(Rigid)	0.1360	0.1420	0.0772

solved, such as the new word detection and Chinese words segmentation.

Disambiguation is another big obstacle in CLIR research, much research [1] has been done in statistical way. Usually these research has the same basic assumption that if the source language query words concur then their corresponding translation also occur in the target documents. We use the common web search engine the calculate the retrieved documents numbers of the translation candidate, and this method is also based on the same assumption. One of the drawback of this method is due to it's long time consuming, and when applied on the real problem, this method may be not so effective as the experimental environment. So better disambiguation has to be done.

Through our experiments, E-C CLIR experiment still can't achieve the precision as other CLIR task. In the near future, we plan to combine some other OOV methods [12, 6] and develop new disambiguation algorithm based on the corpus learning. Also through the analysis of the failure of the E-C CLIR experiments, our plan for the new track of the NTCIR evaluation research are as follows:

- Various OOV translation will be combined in the queries translation;
- A stemming error detection mechanism should be added to the OOV filtering system;
- The effects of the language model retrieval method in the Chinese SLIR need more deep analysis;
- New disambiguation algorithms will be added to the translation system, and it should satisfy a limited time complexity.

4 Acknowledgments

This research was supported by the Beijing New Star Plan of Technology&Science Fund of China under Grant NO.60203007 and the National Natural Science Foundation of China under Grant NO.H020820790130.

References

[1] L. Ballesteros and W. Croft. Resolving ambiguity for cross-language retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 64 – 71, Melbourne, Australia, 1998.

[2] K. Kishida, K. hua Chen, et al. Overview of clir task at the fourth ntcir workshop. In *NTCIR Workshop 4 Meeting:Evaluation of Information Access Technologies*, pages 1–59. National Institute of Informatics, June 2004.

[3] The lemur toolkit for language modeling and information retrieval. http://www.lemurproject.org/.

[4] A. Maeda, F. Sadat, et al. Query term disambiguation for web cross-language information retrieval using a search engine. In *International Workshop on Information Retrieval with Asia Languages,Proceedings of the fifth international workshop on on Information retrieval with Asian languages*, pages 25 – 32, Hong Kong, China, 2000.

[5] C. McEwan, I. Ounis, and I. Ruthven. Building bilingual dictionaries from parallel web documents. In *Proceedings of the Twenty-Fourth European Colloquium on Information Retrieval Research (ECIR 02)*, pages 303 – 323, London, UK, 2002.

[6] H. Meng, S. Khudanpur, G.-A. Levow, et al. Mandarin english information (mei) investigating translingual speech retrieval. In *Proc. of NAACL Workshop on Embedded Machine Translation*, pages 23 – 30, London, UK, 2000.

[7] P. M.F. An algorithm for suffi x stripping. *Program*, 3(14):1980, 1980.

[8] S. E. Robertson, S. Walker, et al. Okapi at trec-3. In *Text REtrieval Conference (TREC) TREC-3 Proceedings*, pages 109 – 128, 1995.

[9] C. Zhai and J. Lafferty. 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*, pages 334 – 342, New Orleans, Louisiana, United States, 2001.

[10] C. Zhai and J. Lafferty. Two-stage language models for information retrieval. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 49 – 56, Tampere, Finland, 2002.

[11] J. Zhang, L. Sun, et al. A trigger language model-based ir system. In *The 20th International Conference on Computational Linguistics(COLING2004)*, pages 680 – 686, Geneva, Switzerland, 2004.

[12] Y. Zhang and P. Vines. Detection and translation of oov terms prior to query time. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 524 – 525, Sheffi eld, United Kingdom, 2004.