ISCAS at NTCIR-3: Monolingual, Bilingual and MultiLingual IR Tasks

Junlin Zhang,Le Sun,Weimin Qu, Lin Du, Yufang Sun,Yangxing Fan,Zhigen Lin
Chinese Information Processing Center,
Institute of Software,Chinese Academy of Sciences,
P.O.Box 8717,Beijing,100080,P.R.China
E_mail:{zjl,lesun,qwm,ldu,yfsun}@sonata.iscas.ac.cn

Abstract

This paper reports the methods and procedures we took in the CLIR track of NTCIR3. In monolingual subtasks we mainly describe index representation and word segmentation method. We use hybrid model integrating MT approach and dictionary-based approach in bilingual and multilingual subtasks. The method for combining the monolingual retrieval results to produce the final rank list is explained in detail. Finally, we present an improved query translation method in CLIR in order to get better query translation quality.

Keywords: Cross-Language Information Retrieval, Bi-direction Machine Translation

1 Introduction

With the rapid development of Internet and increasing mount of various language resources on Internet, CLIR has become a research hotspot in IR research community. In CLIR, users are allowed to build a query in their native language to search documents written in another language, and utilize the retrieved result effectively. Due to the differences between source and target languages, query translation is usually employed to unify the language in queries and documents. Some different approaches have been proposed for query translation. Dictionary-based approach uses machine-readable dictionaries and selection strategies like select all[1], select best N [2],and randomly select N [3].Corpus-based approaches exploit sentence aligned corpora[4] and document aligned corpora[5]

This paper mainly describes the methods and procedures we took in participating in NTCIR3 CLIR track. We focus our experiments on hybrid approach that integrates both MT and lexical method in CLIR. In order to improve query translation quality, we also present Bi-direction Machine Translation method and describe the main processing flow of this method.

This paper is organized as follows: Section 2 is the overview of our work at NTCIR3. Section 3, Section 4 and Section 5 describe the methods and procedures we took in Monolingual IR subtasks, Bilingual CLIR subtasks and Multilingual CLIR Subtasks respectively. In Section 6 we explain the basic idea of the bi-direction machine translation method. Section 7 is our analysis of some experimental results. We conclude this paper and preview our future work in Section 8.

2 Our Work at NTCIR3

We participated in 12 subtasks in CLIR

track of NTCIR3. Data listed in Table 1 shows the average precision of each subtask we participated in. In order to describe technology and method we adopt more clearly, we classify all these subtasks into three groups:

- Monolingual IR group.
- Bilingual CLIR group.
- Multilingual CLIR group.

We use similar methods and procedures in subtasks that belong to the same group. So we will just describe one or two typical subtasks in each group in detail.

Run Types		Average Precision		
		D	TC	TDNC
Group1	C-C	0.1789	0.2389	0.2932
	E-E	0.2433	0.3166	0.3229
	J-J	0.1917	0.2400	/
Group2	C-E	0.0559	0.1277	0.1330
	C-J	0.0581	0.0615	0.0847
	J-C	0.0563	/	0.1376
	E-C	0.0314	0.0426	0.0768
Group3	C-CJ	0.1012	0.1374	0.1752
	C-JE	0.0408	0.0546	0.0741
	C-CE	0.1371	0.1904	0.2377
	E-CE	0.1651	0.0713	0.0999
	C-JCE	0.0920	0.1263 0.1264	/

Table1.Average Precision of All Subtasks of ISCAS-------Relax

3 Monolingual IR Subtasks

Our NTCIR3 monolingual IR system is based on the C-C IR system[6] previously built for the NTCIR2. We build C-C, E-E and J-J monolingual IR systems by adding specific components in the previous system. As for the document rank method, we adopt the classical VSM (Vector space mode) in our system.

3.1 C-C Subtask

Since word boundaries are not marked in Chinese written text, word segmentation is necessary to break Chinese sentences into indexing terms, which can be words, single characters, two characters, and so on. While breaking sentences into words is necessary for deep natural language processing, character-based indexing could be employed in IR. In practice, treating bi-gram as indexing terms is not only simple to apply but also effective. It may not be efficient in space, however it takes no external linguistic resources to index a collection. Thus this approach can be readily applied to documents in any domain.

In our experiment in NTCIR3, we performed several different C-C runs based on either word-based indexing or bi-gram indexing. Among them, three submitted C-C subtasks that use the D, TC or TDNC field respectively are all based on bi-gram indexing. Other C-C run results will be used in the later CLIR procedures. All the other subtasks which are relevant with Chinese document collection are word based index.

Our segmentation algorithm is called bi-direction maximal match algorithm. It scans the Chinese sentence two times by looking up the maximal match term in a general purpose dictionary: The first time is from left to right and the second time reverse the scan order from right to left. This way we can identify and avoid some type of segmentation ambiguity.

VSM is employed to calculate the similarity between query vector and document vector. The term of vector is word. If $T=\{t_j\}$ is a term set, then query vector v_j of topic j can be express $V_j=(v_{j1},v_{j2},\ldots v_{jn})$,in which v_{jk} denotes the weight of t_k in v_j.The vector $D_i=(d_{i1},d_{i2},\ldots,d_{in})$ denotes a document ,d_{ik} denotes the weight of t_k in d_i.The similarity between v_j and d_i is calculated by following formula

$$s_j = \sum_{k=1}^{n} d_{ik} * v_{jk} \Big/ \sqrt{\sum d_{ik}^2 + \sum v_{jk}^2}$$

3.2 E-E Subtask and J-J Subtask

We use similar method and procedures in E-E subtask and J-J subtask with C-C subtask. However, the Chinese segment unit was replaced by Japanese segment unit which is part of the "juman" software in J-J subtask. In E-E subtask we remove the Chinese segment unit and add a stemming procedure based on the very commonly used "porter" stemming algorithm[7].

4 Bilingual CLIR

The main concern of subtasks in the Bilingual CLIR is query translation. The easiest way to find translations is to look up each query term in a bilingual dictionary. However, We can't neglect problems brought by this method such as coverage, spelling norms.[8] Applying MT in CLIR is also a straightforward approach. Another option to using translation dictionaries is using a parallel or comparable corpus, that is, the same or similar text written in different languages.

In Bilingual CLIR subtasks of NTCIR3, We use the similar technology and procedures. The E-C subtask will be clearly described as a typical subtask. In our experiment, we use the hybrid method integrating MT approach and lexical approach to translate the English query into Chinese. Our process consists of following 3 steps:

Step 1. Use English-Chinese MT system named "read world"[9] to translate English topics into Chinese. Then we search the relevant documents in the Chinese document collection with our Chinese monolingual IR system. This way we get the relevant subset A.

Step 2. We translate each term in English topics by looking up each word in an English-Chinese dictionary. Then we search the relevant documents in the Chinese document collection with our Chinese monolingual IR system. This way we get the relevant subset B.

Step 3.Combine the subset A and subset B according to the rank score of each retrieved relevant document. We regard the top 1000 rank scored documents as most relevant ones with a topic.

We performed 3 E-C CLIR runs that make use of different fields :ISCAS-E-C-D-03, ISCAS-E-C-TC-01 and ISCAS-E-C-TDNC-02. The result is not satisfactory and we analyze some reasons of failure as following:

Proper nouns such as people's name, address name can't be correctly translated because of the coverage of the lexicon. MT system employed by us will keep the original words in the translated query if it can't find the term. For example, some important words like "taoyan", "TakeshiKitano","anti-EINino","EINino","Renalt","Nissan" and the like didn't get any translation at all. Even though some proper nouns are within the coverage of the lexicon, their translations never or seldom appear in the target language document collection. For example, "Kim Dae-Jung" was translated into "基姆達埃瓊斯" and "James Soong" was translated into "詹姆士松".

The failure translation of phrase of MT also contributes to the bad retrieval effectiveness of E-C CLIR. For example," China Airlines" was translated as "中國飛機","TV programs" was translated as "電視程式" and " TV stations" was translated as "電視車站".

Insufficient disambiguation ability of MT system in query translation also causes many wrongly translated words.

5 Multilingual CLIR

In multilingual CLIR group, we regard the C-CJE subtask as the typical one that can clearly show the technology and methods we adopt. The C-CJE subtask involves searching Chinese topics in a collection comprised of mixed Chinese, Japanese and English documents.

Generally speaking, there are two different ways to solve this problem: combining monolingual retrieval results or combining queries in all document languages. The first step in both approaches is to translate the queries from the source language to all document languages .In this subtask, one needs to translate the Chinese topics into English and Japanese. With the first approach, a monolingual run is performed for each document language. Then the monolingual retrieval results are combined to produce the final ranked list of documents. With the second approach, the queries in all document languages are combined first. Then the pooled queries are searched against the mixed document collection. We took the first approach to the C-CJE subtask. Figure 1 shows the diagram of C-CJE subtask.

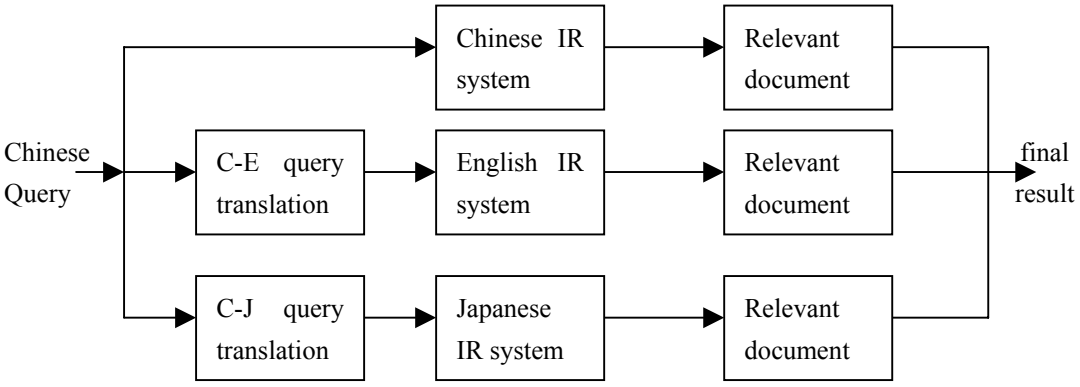

Figure 1. Diagram of C-CJE subtask

We first translated the Chinese topics into English and Japanese. We then performed three monolingual retrieval runs using the subsets of the Chinese Japanese and English documents respectively. The next step is to combine the 3 result sets produced by Chinese, Japanese and English monolingual retrieval systems. Considering the rank score of each relevant document given by each monolingual retrieval system is a local one and incomparable, our combing strategy is to give a comparable rank score to each document by following formula:

$$RankScore(D)=(RS_d—RS_{Min})/(RS_{Max}-RS_{Min})$$

Where $RS_{d:}$ the original rank score given by monolingual retrieval system.

$RS_{Min:}$ the minimal rank score among all the mixed ranked lists of three monolingual retrieval systems.

$RS_{Max:}$ the maximal rank score among all the mixed ranked lists of three monolingual retrieval systems.

Because the number of the documents in Chinese, in Japanese and in English is different, it's reasonable to expect that the number of relevant Chinese, Japanese and English documents should be decided according to the ratio of the different monolingual document collection number for a topic. So we should merge the monolingual Chinese retrieval result, the monolingual Japanese retrieval result and the monolingual English retrieval result in 24:12:1 to produce the final ranked list.

Considering we have gotten the rank lists of C-CJ, C-C, C-E and C-JE subtasks before we process C-CJE subtask, We produce the final rank list of C-CJE by combining the results of C-CJ and C-E or the results of C-JE and C-C in actual merging process of C-CJE subtask. We submit 3 results in C-CJE subtask: ISCAS-C-EJC-TC-01, ISCAS-C-EJC-TC-02 and ISCAS-C-JEC-D-03. The final rank list of ISCAS-C-EJC-TC-01 is the combination of results of C-CJ subtask and C-E subtask .1/10 of the final rank list was selected from C-E result for a topic. The final rank list of

ISCAS-C-EJC-TC-02 is the combination of results of C-JE subtask and C-C subtask. 3/10 of the final rank list was selected from C-JE result for a topic. The final rank list of ISCAS-C-JEC-D-03 is the combination of results of C-JE subtask and C-C subtask. 3/10 of the final rank list was selected from C-JE result for a topic.

In other subtasks of multilingual CLIR group, we use the similar technology and procedure. The ratio of the number of different monolingual retrieval result in the final ranked list is kept as following:

C-CE subtask: Chinese documents occupy 19/20 and English documents occupy 1/20 of final ranked list.

C-CJ subtask: Chinese documents occupy 2/3 and Japanese documents occupy 1/3 of final ranked list.

C-JE subtask: Japanese documents occupy 12/13 and English documents occupy 1/13 of final ranked list.

E-CE subtask: Chinese documents occupy 19/20 and English documents occupy 1/20 of final ranked list.

6 Improved Query Translation Method: Bi-direction Machine Translation Method

Using MT to finish query translation in CLIR is not a new approach. It translates query to reduce the task into monolingual retrieval. However, translation quality of MT system is still not satisfactory. In this paper, we present an improved query translation approach, bi-direction machine translation approach, in CLIR to try to find out the obvious wrongly translated terms and improve the translated query by combining other method.

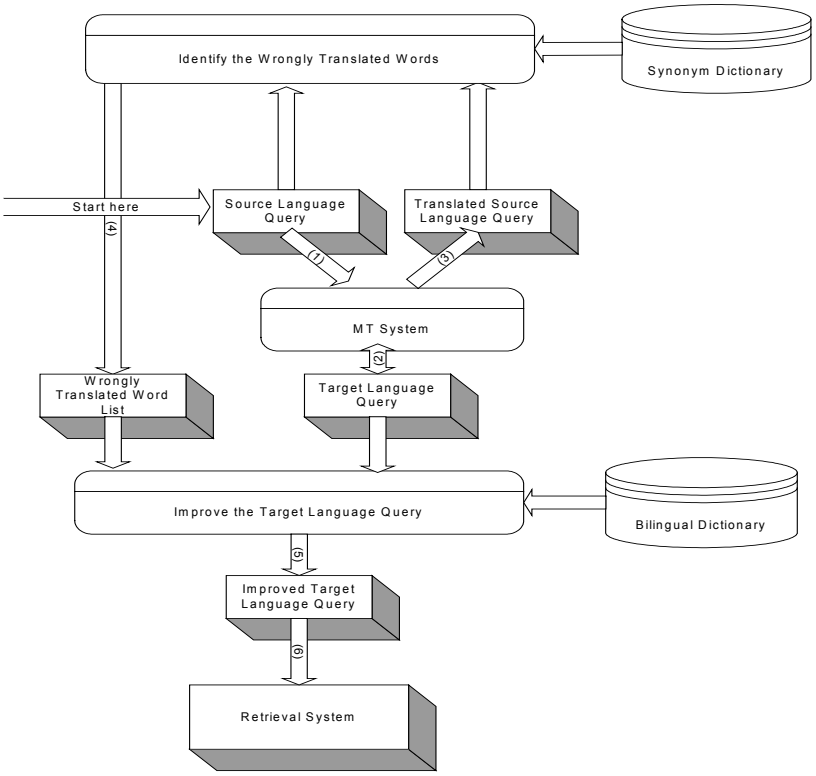

Figure 2 The Processing Flow of Bi-direction Machine Translation

The main processing flow of bi-direction machine translation approach is showed in figure 2. In our method, source language query is first translated into target language query by an MT system and then the target language query is translated back into source language query by the MT system. Next we compare the word bags of two queries. If the original source language query term can find the correspondent synonym in reverse translated query, we think that the word has almost correct translation, or we assure the original query term has not been correctly translated. This way we can ameliorate the translated target language query by keeping the correct translation and adopting other method to replace the wrongly translated words. The easy solution to replace the wrongly translated terms is to look up alternatives in a bilingual dictionary.

7 Analysis of Results

We found that the effectiveness of runs that use TDNC fields was better than the runs which use D field or TC fields in most of our CLIR experiments. We think the content of D field is relatively short sentence and the wrong translation of important terms may play an important role on the final retrieval result. TDNC fields contain much more sentences, so wrong translation of important terms don't have as serious effect as they do in short sentence.

We compare the effectiveness of Japanese query translation and English query translation. We expect the E-C run should get better results than J-C run according to the current state of the effectiveness of E-C and J-C MT system. But the experiments show J-C run is better. Figure 3 show this clearly. We think it is because Chinese and Japanese are much closer in many fields such as similar character or phrase compared with the relationship between Chinese and English.

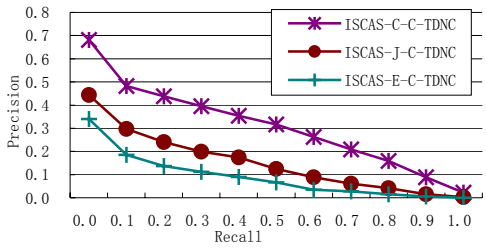

Figure 3 Precision-Recall Curves for J-C and E-C

In C-CJE subtask we submit two runs which use the TC fields: ISCAS-C-EJC-TC-01 and ISCAS-C-EJC-TC-02. However, the final rank lists are merged from different resources. The final rank list of ISCAS-C-EJC-TC-01 is the combination of results of C-CJ and C-E. 1/10 of the final rank list was selected from C-E result for a topic. The final rank list of ISCAS-C-EJC-TC-02 is the combination of results of C-JE and C-C. 3/10 of the final rank list was selected from C-JE result for a topic. From figure 4 we can see that the average precision is almost same. This illustrates the merging strategy are stable enough.

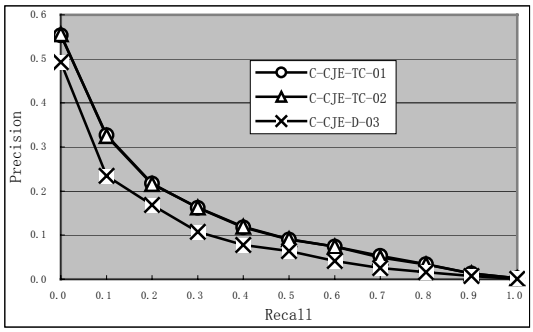

Figure4 Precision-Recall Curve for C-CJE

8 Conclusion and future work

We use hybrid method combining MT method and lexical method in the CLIR track of NTCIR3. Our results show that runs which use TDNC fields can get better effectiveness in CLIR than runs which use TC or D field and Japanese query translation can do better than English query translation by this method. We also describe the main processing of the

bi-direction machine translation method in CLIR. Our future work will focus on finding or building the resources we need to make experiment to testify the effectiveness of this method.

Acknowledgments

This work is supported by China 863 project(Grant No. 2001AA114040) and the National Science Fund of China under contact 69983009.

Reference

[1] Davis, M.W. (1997) "New Experiments in Cross-Language Text Retrieval at NMSU's Computing Research Lab." Proceedings of TREC 5,39-1~39-19

[2] Hayashi,Y.Kikui,G.and Susaki,S.(1997) "TITAN:A Cross-linguistic Search Engine for the WWW." Working Notes of AAAI-97 Spring Symposiums on Cross-Language Text and Speech Retrieval, pp.58-65

[3] Kowk, K.L. (1997) "Evaluation of an English-Chinese Cross-Lingual Retrieval Experiment." Working Notes of AAAI-97 Spring Symposiums on Cross-Language Text and Speech Retrieval, 110-114

[4] Davis, M.W. and Duning, T. (1996) "A TREC Evaluation of Query Translation Methods for Multi-lingual Text Retrieval." Proceedings of TREC-4, 1996.

[5] Sheridan,P. and Ballerini, J.P. (1996) "Experiments in Multilingual Information Retrieval Using the SPIDER System." Proceedings of the 19th ACM SIGIR,58-65

[6] Yibo Zhang,le Sun,Lin Du,Youbing Jin, Yufang Sun,"ISCAS'text retrieval in NTCIR wrkshop 2." Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization,pp.146-153

[7] William B. and Ricardo B.Y. Information Retrieval Data Structures & Algorithms. Prentice Hall PTR. 1993

[8] Gregory Grefenstette "The Problem of Cross-Language Information Retrieval." Cross Language Information Retrieval. Kluwer Academic Publishers.1998

[9]Read World. http://www.readworld.com/tran