Two Stages Refinement of Query Translation for Pivot Language Approach to Cross Lingual Information Retrieval: A Trial at CLEF 2003

Kazuaki KISHIDA
Surugadai University / National Institute of Informatics, JAPAN
kishida@surugadai.ac.jp

Noriko KANDO
National Institute of Informatics, JAPAN
kando@nii.ac.jp

Abstract

This paper reports experimental results of cross-lingual information retrieval from German to Italian. The authors are concerned with CLIR in the case that available language resources are very limited. Thus transitive translation of queries using English as a pivot language was used to search Italian document collections for German queries without any direct bilingual dictionary or MT system of these two languages. In order to remove irrelevant translations produced by the transitive translation, we propose a disambiguation technique, in which two stages of refinement of query translation are executed. Basically, this refinement is based on the idea of pseudo relevance feedback. In the first stage, for each source query term, we select a translation candidate that appears most frequently in the set of top-ranked documents searched for a set of terms provided via transitive translation of the source query. Next, in the second stage, a standard query expansion based on pseudo relevance feedback is conducted. Our experiment result showed that the two stages refinement method is able to improve significantly search performance of bilingual IR using a pivot language. However, it also turned out that performance of this method is inferior to that of machine translation method.

1 Introduction

1.1 Purpose

This paper aims at reporting our experiment of cross language IR (CLIR) from German to Italian at CLEF 2003. Our fundamental interest is CLIR between languages with very limited translation resource, and then we attempt to explore a new approach for transitive query translation using English as a pivot. It is because translation resource between English and each language is often easily obtained, and it is true not only in European environment but also in East Asian environment. Although East Asian languages are completely different from English, pivot language approach using English is very important because of limited availability of resources for direct translation among them.

Thus the basic premise we made for the experiment in CLEF 2003 is that only very limited language resources are available for executing CLIR runs. For example, it was supposed that there is

- no bilingual dictionary,
- no machine translation (MT) system, and
- no parallel corpus,

between German and Italian directly, and

- no corpus written in the language of query (i.e., no German corpus).

We decided to employ only two relatively small dictionaries of German to English (G to E) and English to Italian (E to I), which are easy to be available through the Internet. As mentioned above, our research purpose is fundamentally to develop an effective method for enhancing performance of CLIR in the situation that language resource is very poor.

1.2 Basic idea

According to this presupposition, our method for CLIR is to be characterized as

- dictionary-based approach (query translation),
- pivot language approach (English is a pivot).

As well-known, it is possible that some extraneous or irrelevant translations are unjustly produced by the dictionary-based approach [1]. Particularly, in the case of pivot language approach, consecutive two steps of translation (e.g., German to English and English to Italian) yield often much more extraneous translation candidates because of double replacements of each word. Therefore, a term disambiguation technique is dispensable.

In the presupposition of our study, the resource to be used for translation disambiguation is only the target document collection (i.e., Italian document sets included in the CLEF test collection). We will propose a disambiguation technique in which pseudo relevance feedback is repeated for refining query translations. The basic procedure is as follows:

- Initial search: the target document collection is searched for all translation candidates produced by dictionary-based replacements via pivot language.
- First feedback (*disambiguation stage*): the set of translation candidates are reduced by using term occurrence statistics within a set of some top-ranked documents obtained by the initial search.
- Second search: the target document collection is searched for the reduced set of translations.
- Second feedback (*query expansion stage*): the reduced set of translations is expanded by using a standard pseudo relevance feedback technique.
- Final search: the target document collection is searched for the extended set of search terms.

The two stages of refinement of translation candidates would enable us to obtain better performance of CLIR in the situation that available language resource is poor. The purpose of this paper is to verify experimentally effectiveness of the two stages refinement technique using the test collection of CLEF 2003.

This paper is organized as follows. In the section 2, we will review some previous works on translation disambiguation techniques and pivot language approach. In the section 3, the technique of two stages refinement of query translation will be introduced. The section 4 will describe our system used in the experiment of CLEF 2003. In the section 5, the result will be reported.

2 Previous works

2.1 Translation disambiguation techniques

In the CLIR field, various ideas or techniques for translation disambiguation have been proposed. Among them, some researchers have explored methods of employing the target document collection for identifying extraneous or irrelevant translations. The typical approach is to use co-occurrence statistics of translation candidates according to an assumption that "the correct translations of query terms should co-occur in target language documents and incorrect translation should tend not to co-occur" (Ballestellos and Croft [2]). Many works have been attempted basically in line with the idea [3-9].

The fundamental procedure is as follows:

- Computing similarity degrees for all pairs of translation candidates based on co-occurrence frequencies in the target document collection,
- Selecting 'correct' pairs of translations according to the similarity degrees.

One of the difficulties for implementing the procedure is that computational complexity in selecting correct translations is increasing as the number of translations becomes large. For alleviating the problem, Gao, et al.[4] have proposed an approximate algorithm for choosing optimal translations.

2.2 Pivot language approach

So many languages are spoken in the world, while the bilingual resources are limited. There is no guarantee that useful resources are always available for the combination of two languages that we need in the real situation. For example, it may be difficult to find bilingual resources in machine-readable form between Dutch and Japanese. One of the solutions is to employ English as an intermediate (pivot) language, since English is an international language and it is reasonably expected that bilingual dictionaries or MT systems with English are prepared for many languages.

The basic approach is transitive translation of query by using two bilingual resources (see Ballesteros [10]). If two MT systems or two bilingual dictionaries of Dutch to English and English to Japanese are available, we can translate Dutch query into Japanese without any direct Dutch-Japanese dictionary. This approach has already been attempted by some researchers [11-15].

In the case of using successively two bilingual dictionaries for query translation, it is crucial to solve translation ambiguity because possibly so many extraneous or irrelevant search terms are generated by the two steps of translation. Suppose that an English term obtained from a bilingual dictionary of from the source language to English was irrelevant translation. Inevitably, all terms listed under the English term in the bilingual dictionary from English to the target language would be also irrelevant. Therefore, much more extraneous translations are to be generated in pivot language approach than in standard single-step translation process.

To the disambiguation, Ballesteros [10] has attempted to apply co-occurrence frequency-based method, query expansion and so on. Meanwhile, Gollins and Sanderson [16] proposed a technique of “lexical triangulation”, in which two pivot languages are used independently and removal of error translation is tried by taking only translations in common from two ways of transitive translation using two pivot languages.

3 Two Stages Refinement of Translations

3.1 Translation disambiguation stage

Translation disambiguation technique based on term co-occurrence statistics may be useful in the situation that our study is presupposing, since the technique makes use of only the target document collection as source of disambiguation. However, as already mentioned, the computational complex is fairly high. Also, it should be noted that term co-occurrence frequencies can be considered as macro-level statistics on the entire document collection. This means that the disambiguation based on the statistics may lead to false combination of translation candidates (see Yamabana et al. [3]). Even if two terms A and B are statistically associated in general (i.e., in the entire collection), the association is not always valid in a given query.

Therefore, in the study, the authors decided to use an alternative disambiguation technique, which is not based on term co-occurrence statistics. First, we define some mathematical notations such that

s_j : terms in the source query ($j = 1,2,...,m$),

T_j : a set of translations in the pivot language for the *j*-th term s_j ,

T'_j : a set of translations in the target language for all terms included in the set T_j .

By transitive translation process using two bilingual dictionaries, it is easy to obtain a set of translated query terms in the target language with no disambiguation,

$$T = T'_1 \cup T'_2 \cup ... \cup T'_m . \quad (1)$$

The procedure of disambiguation we propose is to search the target document collection for the set of terms T , and then to select the most frequently appearing term in the top-ranked documents, from each set of T'_j respectively (see Figure 1). The basic assumption is that ‘correct’ combination of each translation from distinct original search terms tends to occur together in a single document in the target collection. If so, such documents are expected to be ranked higher in the result of search for the set T .

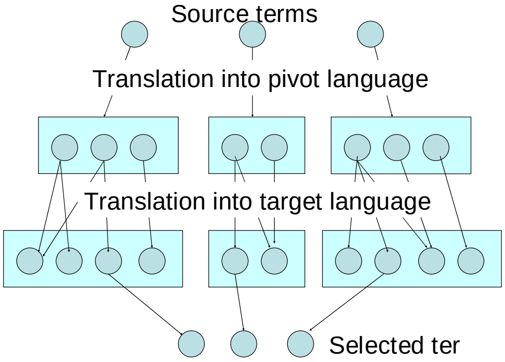
Source terms
Translation into pivot language
Translation into target language
Selected ter

Figure 1 Outline of translation disambiguation

Suppose that we have three sets of translations in the target language as follows:

T'_1 : term A, term B, term C,

T'_2 : term D, term E,

T'_3 : term F, term G.

Also, it is assumed that a combination of term A, D and F is correct and the other terms are irrelevant. In such situation, we can expect reasonably that the irrelevant terms do not appear together in each document because the probability that such irrelevant terms have relations each other is low. Meanwhile, the ‘correct’ combination of term A, D and F would tend to appear in documents more than any combinations including irrelevant translation. Therefore, the documents containing the ‘correct’ combination possibly have higher score for ranking.

For detecting such combination from the result of the initial search for the set T , it would be enough that we use document frequency of each translation in the set of top-ranked documents. That is, we can choose a term $\tilde{t}_j$ for each T'_j ($j = 1,2,...,m$) such that

$$\tilde{t}_j = \arg\max r_t , \quad t \in T'_j \quad (2)$$

where r_t is the number of top-ranked documents including the term t . Finally, we obtain that a set of *m* translations through the disambiguation process

$$\tilde{T} = \{\tilde{t}_1, \tilde{t}_2, ..., \tilde{t}_m\} . \quad (3)$$

Ideally, we should make use of co-occurrence frequencies of all combinations of translation candidates in the set of top-ranked documents. However, the computational cost is expected to be fairly high since we need to compile the statistics dynamically for each search run. A solution for avoiding the complexity is to count only simple frequencies instead of co-occurrence. That is, if the 'correct' combination of translations often appears, naturally the simple frequency of each translation would also become high. Equation (2) is based on this hypothesis.

3.2 Query expansion stage

In the previous stage, translation ambiguity was resolved, and final m search terms in the target language remain. We can consider the stage as a process for improving precision of search. In next stage, enhancement of recall should be attempted since some synonyms or related terms would have been removed in the previous stage.

According to Ballestellos and Croft[1,2], we execute a standard post-translation query expansion using a pseudo relevance feedback (PRF) technique, in which new terms to be added to the query is selected based on its weight w_t, calculated by a formula of standard probabilistic model,

$$w_t = r_t \times \log \frac{(r_t + 0.5)(N - R - n_t + r_t + 0.5)}{(N - n_t + 0.5)(R - r_t + 0.5)}, \qquad (4)$$

where N is the total number of documents, R is the number of relevant documents, and n_t is the number of documents including term t. It should be noted that, in PRF, the set of relevant documents is assumed to be the set of some top-ranked documents by the initial search. Therefore, r_t is defined as the same before (see Equation (2)). We denote the expanded term set by the method as $\tilde{T}'$.

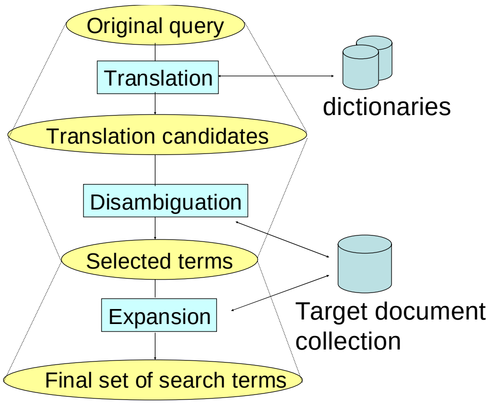

Figure 2 Two stages refinement of translation

To sum up, the method for refining the result of query translation we propose consists of two stages: (a) translation disambiguation and (b) post-translation query expansion. The detailed procedure is as follows (see also Figure 2):

- Obtaining a set of translations T (see Equation (1)) by transitive translation,
- Searching the target document collection for the set T (i.e., initial search),
- Selecting a single translation from each T'_j respectively, according to the document frequency in the top-ranked documents by the initial search, and obtaining a new set $\tilde{T}$ (see Equation (3)) (i.e., disambiguation),
- Searching the target document collection for the set $\tilde{T}$ (i.e., second search),
- Adding terms according to the weight shown as Equation (4) (i.e., query expansion),
- Searching finally the target document collection for the expanded set of terms $\tilde{T}'$ (i.e., third search).

4. System Description

4.1 Purpose of the system

The system enables us to search an Italian document collection for a German query automatically, i.e., it is an automatic CLIR system from German to Italian.

4.2 Text processing

Both of German and Italian texts (in documents and queries) were basically processed by the following steps: (1) identifying tokens, (2) removing stopwords, (3) lemmatization, (4) stemming. In addition, for German text, decomposition of compound words was attempted based on an algorithm of longest matching with headwords included in the German to English dictionary in machine readable form.

4.3 Language resources

We downloaded free dictionaries (German to English and English to Italian) from the Internet[1]. Also, stemmers and stopword lists for German and Italian were available through the Snowball project[2]. Stemming for English was conducted by the original Porter's algorithm [17].

Furthermore, in order to evaluate performance of our two stages refinement method comparatively, we decided to use commercial MT software produced by a Japanese company.

4.4 Transitive translation procedure

[1] http://www.freelang.net/
[2] http://snowball.tartarus.org/

Before executing transitive translation by two bilingual dictionaries, all terms included in the bilingual dictionaries were normalized through stemming and lemmatization steps with the same procedure applied to texts of documents and queries. Actual translation process is a simple replacement, i.e., each normalized German term in a query was replaced with a set of corresponding normalized English words, and similarly, each English word was replaced with the corresponding Italian words. As a result, for each query, a set of normalized Italian words, i.e., T in Equation (1), was obtained. If no corresponding headword was included in the dictionaries (German-English or English-Italian), the unknown word was straightforwardly sent to the next step without any change.

Next, refinement of the set T through two stages described in the previous section was executed. The number of top-ranked documents was set to 100 in both stages, and in the query expansion stage, top-ranked 30 terms in the decreasing order of term weights (Equation (4)) were added.

If the top-ranked term is already included in the set of search terms, $\tilde{T}$, term frequency in the query is changed into $1.5 \times y_t$. If not, the term frequency is set to 0.5 (i.e., $y_t = 0.5$).

On the other hand, in the case of using MT software, first of all, the original German query was input to the software. The software we used is automatically executing German to English translation and then English to Italian translation (i.e., a kind of transitive translation). The resulting Italian text from the MT system was processed according to the procedure described in the section 4.2, and finally, a set of normalized Italian words was obtained for each query. In the case of MT translation, only post-translation query expansion was executed with the same procedure and parameters in the case of dictionary-based translation.

4.5 Search algorithm

The well-known Okapi formula [18] was used for computing each document score in all searches of this study, i.e.,

$$z = \sum_{t \in \Omega} \left(\frac{3.0 x_t}{(0.5 + 1.5 l / \bar{l}) + x_t} \times y_t \times \log \frac{N - n_t + 0.5}{n_t + 0.5} \right),$$

where z is the score of a particular document, x_t is the frequency of occurrence of term t in the document, l is the document length, $\bar{l}$ is an average of the document length over the entire document collection, and y_t is the frequency of occurrence of term t in the query. It should be noted that the value of each y_t is always fixed at the frequency of the corresponding original search terms in the source query. Also, Ω is a set of query term, i.e., in the first search $\Omega = T$, in the second search $\Omega = \tilde{T}$, and in the third search $\Omega = \tilde{T}'$. Finally, the documents were ranked in the decreasing order of the values of z.

4.6 Type of runs executed

In CLEF 2003, we executed three runs (see Table 1), in which only <DESCRIPTION> filed in each query was used.

Table 1 Runs submitted in CLEF 2003

ID	Translation method	Refinement	
		Disambiguation	Expansion
NiiMt01	MT	-	done
NiiDic01	Dictionary	done	done
NiiDic02	Dictionary	-	done

5. Results

5.1 Basic statistics

The Italian collections include 157,558 documents in total. The average document length is 181.86.

5.2 System error

Unfortunately, a non-trivial system error was detected after submission of results, i.e., by a bug in our source code, only a last term within the set of search terms has contributed to the calculation of document scores. Inevitably, search performance of all runs shown in Table 1 was very low.

5.3 Results of runs conducted after submission

Therefore, the authors have corrected the source code and attempted to perform again some search runs after submission of results to the organizers of CLEF. Six types of run were conducted as shown in Table 2, which also indicates each value of mean average precision calculated by using the relevance judgment file. Furthermore, recall-precision curves of the six runs are presented as Figure 3. It should be noted that each value in represented in Table 2 and Figure 3 was calculated for 51 topics to which one or more relevant documents are included in the Italian collections.

Table 2 Mean average precision of runs executed after submission (51 topics)

Translation method		Expansion by PRF	
		done	none
MT		.301	.281
Dictionary	With disambiguation	.207	.181
	Without disambiguation	.190	.143

As shown in Table 2, MT outperforms dictionary-based translation significantly. Also, it turns out that the disambiguation technique based on term frequency moderately improves effectiveness of dictionary-based translation method, i.e., the mean average precision with disambiguation is .207 in comparison with .190 in the case of no disambiguation. Especially, Table 2 indicates that our technique of two stages refinement has a large effect on enhancement of search performance since the mean average precision of search with no disambiguation and no expansion by PRF is only .143, which is significantly lower than .207 in the case of searches through the two stages refinement.

However, we can also point out that there is a large difference of performance between MT and the two stage refinement. The reason may be attributed to difference of quality and coverage between the commercial MT software and free dictionaries downloaded from the Internet. Even if it is true, we need to modify the two stages refinement method so that its performance level is approaching to that of MT system.

For example, in Figure 3, at the levels of recall over 0.7, searches with no disambiguation is reversely superior to those with disambiguation. This may be due to that our disambiguation method selects only one translation and consequently may remove some useful synonyms or related terms. A simple solution is possibly to choose two or more translations instead of using directly Equation (2). Although it is difficult to determine the optimal number of translations to be selected, multiple translations for

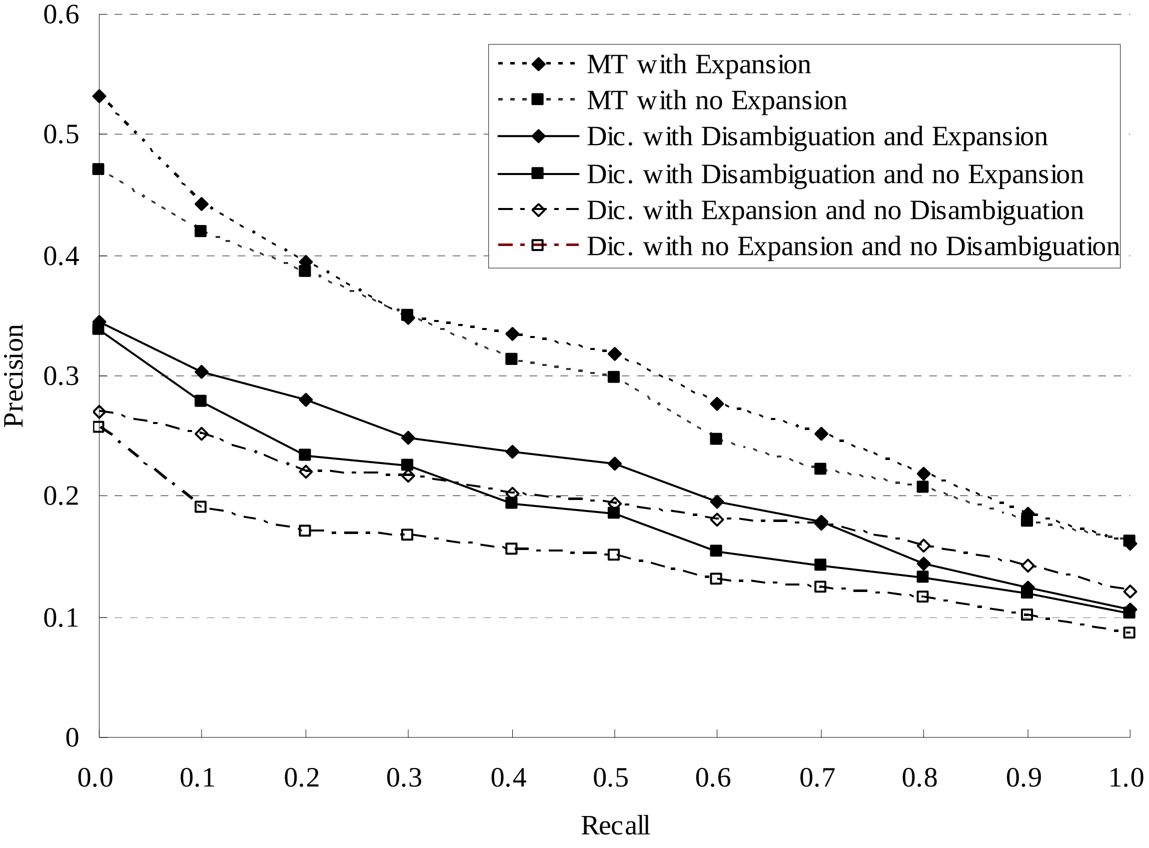

Figure 3. Recall-Precision curves

each source term may improve recall of searches.

6. Concluding Remarks

This paper reported results of our experiment on CLIR from German to Italian, in which English was used as a pivot language. In particular, two stages refinement of query translation was employed for removing irrelevant terms in the target language produced by transitive translation using successively two bilingual dictionaries.

As a result, it turned out that

- our two stages refinement method improves significantly retrieval performance of bilingual IR using a pivot language, and
- the performance is inferior to that by MT-base searches.

By choosing two or more search terms in the disambiguation stage, it is possible that our method becomes more effective.

Reference

[1] Ballesteros, L.A. & Croft, W. B. (1997). Phrasal translation and query expansion techniques for cross-language information retrieval. In Proceedings of the 20st ACM SIGIR conference on Research and Development in Information Retrieval. (pp.84-91).

[2] Ballesteros, L. & Croft, W.B. (1998). Resolving ambiguity for cross-language retrieval. In Proceedings of the 21st ACM SIGIR conference on Research and Development in Information Retrieval (pp.64-71).

[3] Yamabana, K., Muraki, K., Doi, S. & Kamei, S. (1998). A language conversion front-end for cross-language information retrieval. In G. Grefenstette (ed.) Cross-language Information retrieval (pp.93-104). Boston, MA: Kluwer.

[4] Gao, J., Nie, J. Y., Xun, E X., Zhang, J. Zhou, M. & Huang, C. (2001b). Improving query translation for cross-language information retrieval using statistical models. In Proceedings of 24th ACM SIGIR conference on Research and Development in Information Retrieval (pp.96-104).

[5] Lin, C. J., Lin, W. C., Bian, G. W. & Chen, H. H. (1999). Description of the NTU Japanese-English cross-lingual information retrieval system used for NTCIR workshop. In Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition. Tokyo: National Institute of Informatics. http://research.nii.ac.jp/ntcir/workshop/

[6] Sadat, F., Maeda, A. Yoshikawa, M. & Uemura, S. (2002). Query expansion techniques for the CLEF Bilingual track. In C. Peters et al. (Eds.) Evaluation of Cross-Language Information Retrieval Systems: LNCS 2406 (pp.177-184) Berlin: Springer.

[7] Adriani, M. (2002). English-Dutch CLIR using query translation techniques. In C. Peters et al. (Eds.) Evaluation of Cross-Language Information Retrieval Systems: LNCS 2406 (pp.219-225) Berlin: Springer.

[8] Qu, Y., Grefenstette, G. & Evans, D. A. (2002). Resolving translation ambiguity using monolingual corpora: a report on Clairvoyance CLEF-2002 experiments. In Working Notes for the CLEF-2002 Workshop (pp.115-126).

[9] Seo, H. C., Kim, S. B., Kim, B. I., Rim, H. C. & Lee, S. Z. (2003). KUNLP system for NTCIR-3 English-Korean cross-language information retrieval. In Proceedings of the Third NTCIR Workshop on research in information Retrieval, Automatic Text Summarization and Question Answering. Tokyo, National Institute of Informatics. ttp://research.nii.ac.jp/ntcir/workshop/

[10] Ballesteros, L. A. (2000). Cross-language retrieval via transitive translation. In W.B.Croft (Ed.) Advances in Information Retrieval: Recent Research from the Center for Intelligent Information Retrieval (pp.203-234). Boston, MA: Kluwer.

[11] Frantz, M., McCarley, J. S., & Roukos, S. (1999). Ad hoc and multilingual information retrieval at IBM. In Proceedings of the TREC-7, Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/

[12] Gey, F. C., Jiang, H. Chen, A. & Larson, R. R. (1999). Manual queries and machine translation in cross-language retrieval at TREC-7. In Proceedings of the TREC-7, Gaithersburg: MD, National Institute of Standards and Technology. http://trec.nist.gov/pubs/

[13] Hiemstra, D. & Kraaij, W. (1999). Twenty-one at TREC-7: ad-hoc and cross-language track. In Proceedings of the TREC-7, Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/

[14] Chen, A. & Gey, F. C. (2003). Experiments on cross-language and patent retrieval at NTCIR-3 workshop. In Proceedings of the Third NTCIR

Workshop on research in information Retrieval, Automatic Text Summarization and Question Answering. Tokyo, National Institute of Informatics. http://research.nii.ac.jp/ntcir/workshop/

[15] Lin, W. C. & Chen, H.H. (2003). Description of NTU approach to Multilingual Information retrieval. In Proceedings of the Third NTCIR Workshop on research in information Retrieval, Automatic Text Summarization and Question Answering. Tokyo, National Institute of Informatics http://research.nii.ac.jp/ntcir/workshop/

[16] Gollins, T. & Sanderson, M. (2001). Improving cross language information retrieval with triangulated translation. In Proceedings of the 24th ACM SIGIR conference on Research and Development in Information Retrieval (pp.90-95).

[17] Porter, M.F.(1980). An algorithm for suffix stripping. Program, 14(3), pp.130-137.

[18] Roberson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M. & Gatford, M. (1995). Okapi at TREC-3. In Proceedings of TREC-3, Gaithersburg: MD, National Institute of Standards and Technology. http://trec.nist.gov/pubs/