Hybrid Approach of Query and Document Translation with Pivot Language for Cross-Language Information Retrieval

Kazuaki Kishida[1] Noriko Kando[2]

[1] Surugadai University, 698 Azu, Hanno, Saitama 357-8555, Japan
kishida@surugadai.ac.jp
[2] National Institute of Informatics (NII), Tokyo 101-8430, Japan
kando@nii.ac.jp

Abstract

This paper reports experimental results of cross-language information retrieval (CLIR) from German to French, in which a hybrid approach of query and document translation was attempted, i.e, combining results of query translation (German to French) and of document translation (French to German). In order to avoid too high complexity of computation for translating a large amount of texts in documents, we executed pseudo-translation, i.e., a simple replacement of terms by a bilingual dictionary (for query translation, a machine translation system was used). In particular, since English was used as an intermediary language for both translation directions of German and French, English translations at the middle stage were employed as document representations in order to reduce the number of translation steps. By omitting a translation step (English to German), the performance was improved. Unfortunately, our hybrid approach could not show better performance than a simple query translation. This may be due to the low performance of document translation, which was carried out by a simple replacement of terms using a bilingual dictionary with no term disambiguation.

Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval]: Query formulation, Relevance feedback

Keywords

Cross-language information retrieval, Query translation, Document translation, Hybrid approach

1 Introduction

This paper describes our experiment of cross-language IR (CLIR) from German to French languages in the CLEF 2005 campaign. Our focus in this experiment is on examining search performance of a hybrid approach combining query translation and document translation, in which English is employed as an intermediary language for translation.

Some researchers have already attempted to merge two results from query and document translation for enhancing effectiveness of CLIR. An intention of combining them is to enlarge possibility of matching successfully subject representations of the query with those of each document. One problem for implementing this approach is that the document translation is usually a cost-intensive task, but we can alleviate it by using simpler translation techniques, e.g., “pseudo translation” [1] in which each term is simply replaced with its corresponding translations by a bilingual dictionary. It is worth while investigating the performance of the hybrid approach in the case of employing such a simpler document translation technique that is more practical for use in real situation.

This paper is organized as follows. In section 2, the hybrid approach combining two results from query and document translation is discussed. Section 3 describes our system used in the experiment of CLEF 2005. In section 4, the results are reported.

2 Hybrid Approach of Query and Document Translation

2.1 Combination of query and document translation

In order to execute CLIR, we have to match subject representations between a query and each document by translating either the query or documents. In general, queries tend to be translated [2]. This may be due to its easiness for implementation, i.e., no special device is needed for executing CLIR rather than a tool for translating the query text. In contrast, document translation has not often been adopted as the strategy for CLIR partly because a very large amount of processing is needed for translating all documents in the whole database.

However, some researchers have reported that a hybrid approach of query and document translation bring us better search performance in CLIR. For example, McCarley [3] has attempted to use an average of two document scores which are computed from query translation and document translation respectively in order to rank documents for output. Fujii and Ishikawa [4] have translated documents that are searched based on query translation, and tried to re-rank them according to results of the document translation. In NTCIR-4, Kang et al. [1] tried to execute Korean to Chinese and Korean to Japanese bilingual retrieval using the hybrid approach.

An advantage of the hybrid approach is to increase possibility for identifying correctly documents having the same subject content with the query. Suppose that a term A is included in a given search query and its corresponding term in the language of documents is B. If a tool for translation from the query language to the document language can not translate A into B correctly, the system would fail to find documents containing the term B by this query translation. However, if another tool for translation in reverse direction, i.e., the document language into the query language, can identify the term A from the term B, matching between the query and documents including the term B becomes successful.

For implementing the hybrid approach, it is important to solve a problem that document translation is a cost-intensive task. For example, it may take too long time for translating all documents by commercial software for machine translation (MT). McCarley [3] applied a statistical translation technique for alleviating this problem. In contrast, Kang et al.[1] have employed “pseudo” translation technique, in which each term in documents is simply replaced with its translations by using a bilingual dictionary. Although the replacement is not exactly equal to MT, it is so fast and enables us to have translations of a large amount of document texts within a reasonable time.

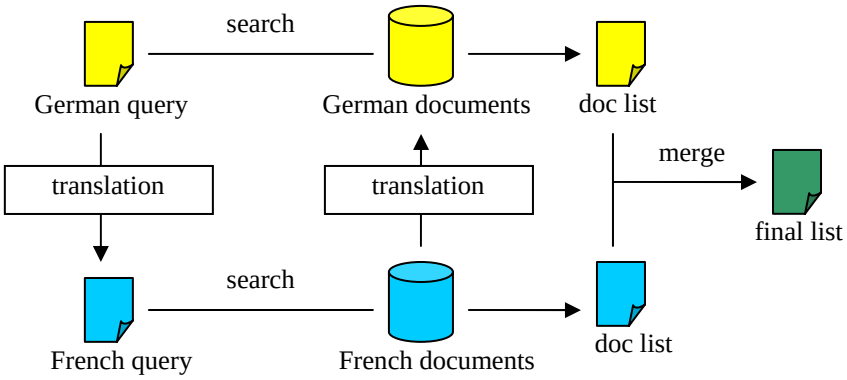

Fig. 1. Procedure of hybrid approach (1)

2.2 Hybrid approach with a pivot language

In our hybrid approach, queries in German are translated by a commercial MT system, and each French term included in documents is replaced with its corresponding German words using bilingual dictionaries. After the translation, two scores are computed for each document from the results of query and document translation respectively. Finally, we calculate a final score for ranking the document by using a simple linear formula such that

$$z = wx + (1 - w)y\,, \qquad (1)$$

where x is a score computed from a results of query translation, y is a score from document translation, and w is a weight (in this paper, we always set that $w = 0.7$). The procedure is graphically shown in Figure 1.

Both the translation methods employed in this experiment, i.e., MT and dictionary-based method, make use of a pivot language. The MT software translates German sentences into English ones, and translates the results into French sentences. Similarly, each term included in French documents are replaced with corresponding English translations by a French to English dictionary, and these English translations are replaced with German terms by an English to German dictionary. An appropriate translation resource is not always available for a pair of languages that actual users require. But, in this case, it is possible that we find translation tools between English and these languages since English is an international language. Therefore, the pivot language approach via English is considered to be useful in real situations, although two steps of translation in this approach often yield erroneously more irrelevant translations, particularly in the case of dictionary-based transitive translation, because all final translations obtained from an irrelevant English term in the middle stage are usually irrelevant [5].

One method for alleviating this problem may be to limit the dictionary-based translation to only conversion of French terms into English ones. In order to compute document scores from documents translated into English, German queries have to be translated into English. In the case of pivot language approach, an English version of the query is automatically obtained in the middle stage of translation from German to French (see Figure 2). Therefore, the number of translation operations is just three as shown in Figure 2. In contrast, the standard hybrid approach in Figure 1 using a pivot language needs four translation operations, i.e., (1) German query to English query, (2) English query to French query, (3) French documents to English documents and (4) English documents to German documents. Removing an operation of dictionary-based translation may contribute to reduction of erroneous translations, and the search performance is expected to be improved.

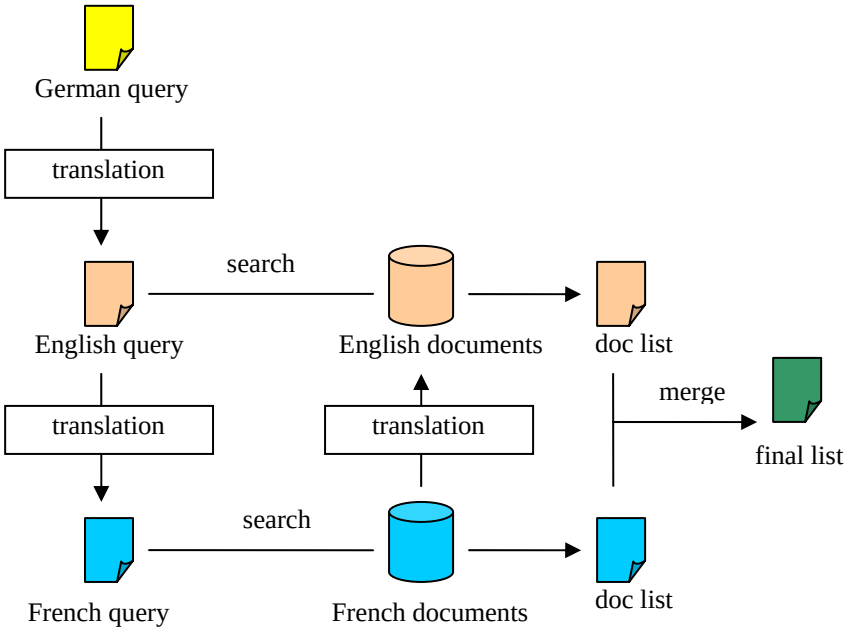

Fig. 2. Procedure of hybrid approach (2)

3 System Description

3.1 Text Processing

Both German and French texts (in documents and queries) were basically processed by the following steps: (1) identifying tokens, (2) removing stopwords, (3) lemmatization, and (4) stemming. In addition, for German text, decomposition of compound words was attempted based on a simple algorithm of longest matching with

headwords included in the German to English dictionary in machine-readable form. For example, a German word, "Briefbombe," is broken down into two headwords listed in the German to English dictionary, "Brief" and "Bombe," according to a rule that only the longest headwords included in the original compound word are extracted from it. If a substring of "Brief" or "Bombe" is also listed in the dictionary, the substring is not used as a separated word.

We downloaded free dictionaries (German to English and English to French) from the Internet[1]. Stemmers and stopword lists for German and French were also available through the Snowball project[2]. Stemming for English was conducted by the original Porter's algorithm [6].

3.2 Translation Procedure

We used a commercial MT system produced by a Japanese company[3] for query translation, and French or English sentences that it output were processed according to the procedures described above. In the case of document translation, each German sentence was processed, and its words and decomposed elements of compound words were simply replaced with corresponding English terms using a German to English dictionary with no term disambiguation. If no corresponding headword was included in the dictionary for a German term, it was entered into the set of English terms with no change as an unknown term. In order to moreover obtain French translations, a set of the English translations is converted using an English to French dictionary by the same procedure with that for obtaining English translations. It should be noted that all terms included in these dictionaries were normalized through stemming and lemmatization processes with the same procedure applied to texts of documents and queries. Therefore, by the dictionary-based translation, a set of normalized English or French terms was obtained.

3.3 Search Algorithm

The standard Okapi BM25 [7] was used for all search runs, and for pseudo-relevance feedback, we employed a term weighting formula,

$$w_t = r_t \times \log \frac{(r_t + 0.5)(N - R - n_t + r_t + 0.5)}{(N - n_t + 0.5)(R - r_t + 0.5)}, \qquad (2)$$

where N is the total number of documents, R is the number of top-ranked documents that is assumed to be relevant, n_t is the number of documents including term t, and r_t is the number of documents including term t in the top-ranked R documents. In this experiment, we always set that $R = 30$ and ten terms were selected based on their weights in Eq. (2). Let y_t be the frequency of a given term in the query. If a newly selected term was already included in the set of search terms, the term frequency in the query y_t was changed into $1.5 \times y_t$. If not, the term frequency was set to 0.5 (i.e., $y_t = 0.5$). The PRF procedure was carried out for all search runs in this experiment.

3.4 Merge of Document Lists

For merging two document lists generated by different strategies (i.e., query and document translation), we used the formula in Eq.(1). More precisely, the procedure is as follows.

(a) Using a result of query translation, document scores are computed, and documents up to 10,000th position in the ranked list are selected in maximum.
(b) Similarly, using a result of document translation, document scores are computed again, and documents up to 10,000th position in the ranked list are selected in maximum.
(c) Final scores for documents selected in (a) and (b) are computed based on Eq.(1) and all documents are re-ranked (If a document was not included in either of lists in (a) or (b), its score is set to zero in the list).

[1] http://www.freelang.net/
[2] http://snowball.tartarus.org/
[3] http://www.crosslanguage.co.jp/english/

3.5 Type of Search Runs

We executed five runs in which <TITLE> and <DESCRIPTION> fields in each search topic were used, and submitted the results to the organizers of CLEF 2005. All runs were executed on the information retrieval system, ADOMAS (Advanced Document Management System) developed at Surugadai University in Japan. The five runs are as follows.

- **Hybrid-1**: merging two results of French translations for query and of German translation for documents.
- **Hybrid-2**: merging two results of French translations for query and of English translation for documents as shown in Figure 1.
- **Query translation**: using only query translation from German to Italian with no document translation as shown in Figure 2.
- **Document translation**: using only document translation from French to German with no query translation
- **Monolingual**: searching the French document collection for the French topics (not translation).

In order to comparatively evaluate the performance of our hybrid approach, search runs using only query translation and only document translation were attempted. In addition, for checking effectiveness of these CLIR runs, monolingual search was also executed.

4 Experimental Results

4.1 Basic Statistics

The target French collection includes 177,452 documents in total. The average document length is 232.65 words. In the case that the document collection was translated into English, the average document length in the English collection amounts to 663.49 and that in the German collection translated from the original French one is 1799.74. Since we did not incorporate any translation disambiguation into our process as mentioned above, each translated document became so long.

Table 1. Average precision and R-precision (average over all 50 topics)

Run	ID	Average Precision	R-Precision
French Monolingual	SrgdMono01	.3910	.3998
Hybrid-1: German doc translation	SrgdMgG02	.2492	.2579
Hybrid-2: English doc translation	SrgdMgE03	.2605	.2669
Query translation	SrgdQT04	.2658	.2642
Document translation	SrgdDT05	.1494	.1605

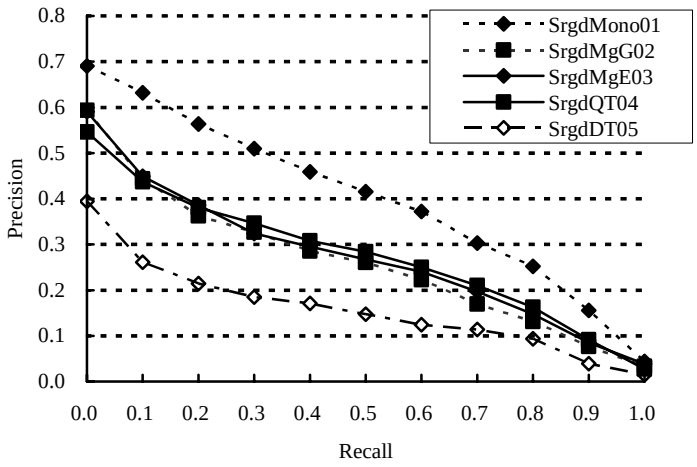

Fig. 3. Recall-precision curves

4.2 Results

Scores of average precision and R-precision are shown in Table 1, and recall-precision curves of these runs are presented in Figure 3. Note that each value in Table 1 and Figure 3 is calculated for all 50 topics that be prepared for evaluating search runs.

As shown in Table 1, the hybrid approach using English documents translated from the original collection (hybrid-2, SrgMgE03) outperforms another hybrid approach using German documents (hybrid-1, SrgdMgG02), i.e., the scores of mean average precision (MAP) are 0.2605 for hybrid-2 and 0.2492 for hybrid-1. Although the degree of difference is not large, dominance of the hyper-2 approach is consistent with our logical expectation.

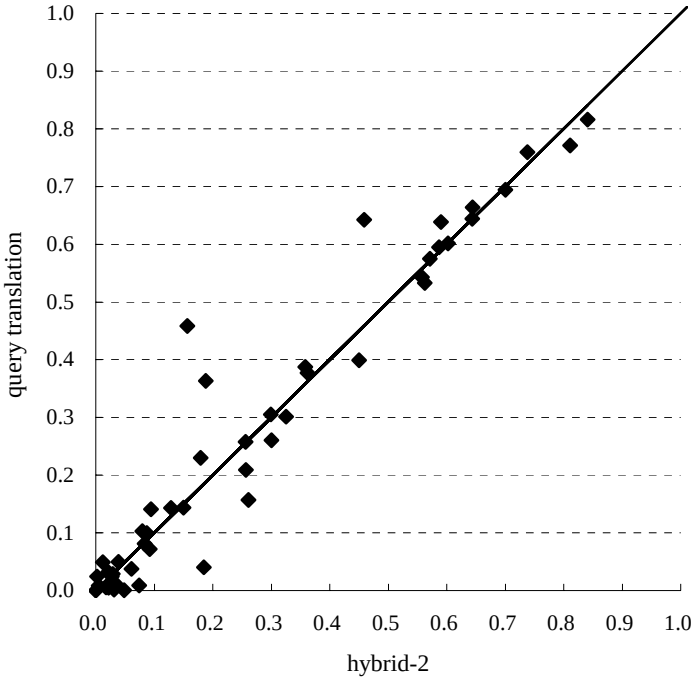

Fig. 4. Topic-by-topic analysis (average precision score)

Unfortunately, the hyper approach could not show better performance than a simple query translation approach (SrgdQT04), i.e., its score of MAP is 0.2658, which is slightly greater than that of SrgdMgE03. This may be due to the low performance in document translation approach, e.g., the MAP score of document translation from French to German (SrgdDT05) is only 0.1494. That is, by combining results from document translation with that from query translation, ranking of relevant documents in the list generated by query translation approach became lower in some topics. Of course, in other topics, the performance was improved as shown in Figure 3, which is a topic-by-topic plot of two scores of average precision for hyper-2 and query translation approach. However, we should consider that our hybrid approach did not show better effectiveness due to the low performance in document translation approach. The reason of the low performance may be (1) quality of free dictionaries downloaded from the Internet and (2) omission of translation disambiguation. We have to solve these problems for improving the performance of our hybrid approach.

5 Concluding remarks

This paper reports the results of our experiment on German to French bilingual retrieval, for which a hybrid approach combining results of query translation and document translation was used. For avoiding too high complexity of computation for translating a large amount of documents in the database, we applied pseudo-translation, i.e., a simple replacement of terms by using a bilingual dictionary. In contrast, machine translation software was used for translation of queries which are usually short.

Since a pivot language approach was applied in the translation process by both MT system and bilingual dictionaries, we attempted to reduce the number of translation steps by employing English translations from the

original French collection as a result of document translation. Actually, it is empirically shown that this approach outperforms slightly the standard hybrid approach using German translations as representations of documents. Unfortunately, our hybrid approach could not show better effectiveness than a simple query translation approach partly because the performance of document translation is poor. We have to develop techniques for enhancing effectiveness of document translation approach.

References

1. Kang, I. S., Na, S. H. Na , Lee, J. H.: POSTECH at NTCIR-4: CJKE Monolingual and Korean-related Cross-Language Retrieval Experiments. In NTCIR Workshop 4 Meeting Working Notes, 2004, p.89-95
2. Kishida, K.: Technical issues of cross-language information retrieval: a review. Information Processing & Management, 41 (2005), 433-455
3. Scott McCarley, J.: Should we translate the documents or the queries in cross-language information retrieval? In Proceedings of the 37th conference on Association for Computational Linguistics (1999) 208-214
4. Fujii, A, Ishikawa, T. Japanese-English cross-language information retrieval integrating query and document translation methods. The Transactions of the Institute of Electronics, Information and Communication Engineers. J84-D-II (2001) 362-369 (*In Japanese*)
5. Kishida, K., Kando, N.: Two-Stage refinement of query translation in a pivot language approach to cross-lingual information retrieval: an experiment at CLEF 2003. In C. Peters, J. Gonzalo, M. Braschler and M. Kluck (Eds), Comparative Evaluation of Multilingual Information Access Systems. LNCS 3237, Springer Verlag, pp.253-262.
6. Porter, M.F.: An algorithm for suffix stripping. Program. 14 (1980) 130-137
7. Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M.M., Gatford, M.: Okapi at TREC-3. In Proceedings of TREC-3. National Institute of Standards and Technology, Gaithersburg (1995) http://trec.nist.gov/pubs/