Two-Stage Refinement of Transitive Query Translation with English Disambiguation for Cross-Language Information Retrieval: An Experiment at CLEF 2004

Kazuaki Kishida[1], Noriko Kando[2], and Kuang-Hua Chen[3]

[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
[3] National Taiwan University, Taipei 10617, Taiwan
khchen@ntu.edu.tw

Abstract. This paper reports experimental results of cross-language information retrieval (CLIR) from German to French. The authors focus on CLIR in cases where available language resources are very limited. Thus transitive translation of queries using English as a pivot language was used to search French document collections for German queries without any direct bilingual dictionary or MT system for these two languages. The two-stage refinement of query translations that we proposed at the previous CLEF 2003 campaign is again used for enhancing performance of the pivot language approach. In particular, disambiguation of English terms in the middle stage of transitive translation was attempted as a new experiment. Our results show that the two-stage refinement method is able to significantly improve search performance of bilingual IR using a pivot language, but unfortunately, the English disambiguation has almost no effect.

1 Introduction

This paper describes our experiment for cross-language IR (CLIR) from German to French in CLEF 2004. In CLEF 2003, the authors proposed the "two-stage refinement technique" for enhancing search performance of the pivot language approach in situations when only limited language resources are available. In those experiments, German to Italian search runs were executed using only three resources: (1) a German to English dictionary, (2) an English to Italian dictionary, and (3) a target document collection [1]. The target document collection was employed as a language resource for both translation disambiguation and query expansion by applying a kind of pseudo-relevance feedback (PRF) [1].

In CLEF 2004, we attempt to add an English document collection as a language resource for executing German to French search runs via English as a pivot. Thus, unlike CLEF 2003, a disambiguation procedure using a document collection is applied to the English term set in the middle position of transitive query translation. This is expected to reduce irrelevant French words by removing inappropriate English translations.

C. Peters et al. (Eds.): CLEF 2004, LNCS 3491, pp. 135 – 142, 2005.

This paper is organized as follows. In Section 2, the two-stage refinement technique and the English disambiguation method are introduced. Section 3 describes the system we used in the CLEF 2004 experiment. In Section 4, the results are reported.

2 Two-Stage Refinement of Query Translation

2.1 Basic Procedure

One purpose of the two-stage refinement technique is to modify the results of query translation in order to improve CLIR performance. The modification consists in two steps: (1) disambiguation and (2) expansion. In our approach, "disambiguation" means selecting a single translation for each search term in the source language, and "expansion" means executing a standard PRF technique using the set of translations selected in the disambiguation stage as an initial query. Although many researchers have performed the two processes together for CLIR, in our method, both processes are based on a PRF technique using the target document collection, i.e., under the assumption that only limited language resources are available, we use the target collection as a language resource for disambiguation.

We define the following mathematical notations:

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

T'_j: a set of translations in the target language for term s_j, and

$T = T'_1 \cup T'_2 \cup ... \cup T'_m$.

First, the target document collection is searched for the set of terms T. Second, the most frequently appearing term in the top-ranked documents is selected from each set of T'_j ($j = 1,2,...,m$) respectively. That is, we choose a term $\tilde{t}_j$ for each T'_j such that

$$\tilde{t}_j = \arg\max r_t \quad (t \in T'_j), \tag{1}$$

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

$$\tilde{T} = \{\tilde{t}_1, \tilde{t}_2, ..., \tilde{t}_m\}. \tag{2}$$

The disambiguation technique is clearly based on PRF, where some top-ranked documents are assumed to be relevant. The most frequently appearing term in the relevant document set is considered as a correct translation in the context of a given query. While standard disambiguation techniques based on term co-occurrence use statistics on the whole collection (see [2]), our method tries to extract information for disambiguation from a part of the collection that is relevant to the given query. We expect this disambiguation approach to find a correct combination of search terms within the context of the query (the combination is not always important in general, i.e., in the whole document set).

In the next stage, according to Ballesteros and Croft [2], a standard post-translation query expansion by the PRF technique is executed using $\tilde{T}$ in (2) as a query. In this study, we use a standard formula based on the probabilistic model for estimating term weights as follows:

$$w_t = r_t \times \log \frac{(r_t + 0.5)(N - R - n_t + r_t + 0.5)}{(N - n_t + 0.5)(R - r_t + 0.5)}, \quad (3)$$

where N is the total number of documents, R is the number of relevant documents, n_t is the number of documents including term t, and r_t is defined as before (see Equation (1)). The expanded term set is used as a final query for obtaining a list of ranked documents.

2.2 Disambiguation During Transitive Query Translation

The pivot language approach is adopted in this paper, i.e., a search term in the source language is translated into a set of English terms, and each English term is transitively translated into terms in the target language. As many researchers have pointed out, if the set of English terms includes erroneous translations, they will yield many more irrelevant terms in the target language.

One solution is to apply a disambiguation technique to the set of English translations (see Figure 1). If an English document collection is available, we can easily execute our disambiguation method described in the previous section.

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 an 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 separate word.

We downloaded free dictionaries (German to English and English to French) from the Internet [1]. Also, stemmers and stopword lists for German and French were obtained through the Snowball project[2]. Stemming for English was conducted by the original Porter's algorithm [3].

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

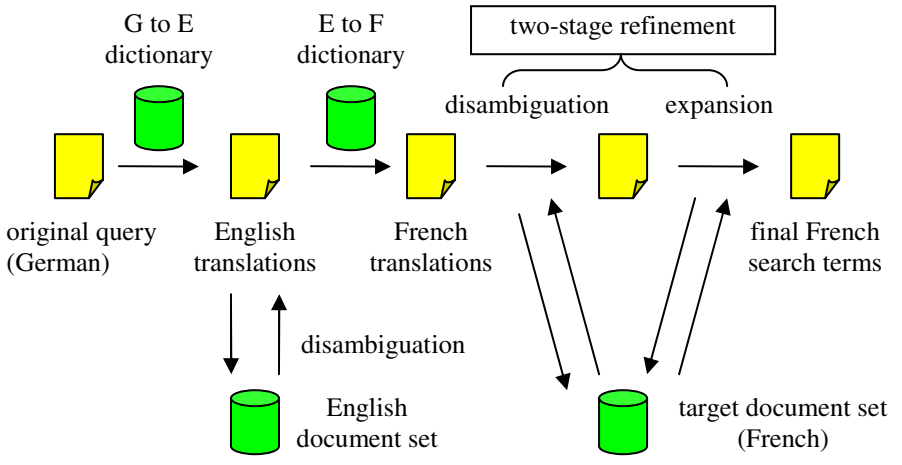

Fig. 1. Two-stage refinement of translation with English disambiguation

3.2 Transitive Translation Procedure

Before executing transitive translation by two bilingual dictionaries, all terms included in the dictionaries were normalized through stemming and lemmatization processes with the same procedure applied to texts of documents and queries. The actual translation process is a simple replacement, i.e., each normalized German term (to which the decomposition process was applied) in a query was replaced with a set of corresponding normalized English words, and similarly, each English word was replaced with the corresponding French words. As a result, for each query, a set of normalized French words was obtained. If no corresponding headword was included in the dictionaries (German–English or English–French), the unknown word was directly sent to the next step without any change. During the transitive translation process, we attempted to apply our disambiguation technique to the set of English words (see Section 2.2).

Next, the translations were refined by our two-stage technique described in the previous section. The number of top-ranked documents was set to 100 in both stages, and in the query expansion stage, the top 30 terms were selected from the ranked list in decreasing order of term weights (Equation (3)).

3.3 Search Algorithm

The standard Okapi BM25 [4] was used for all search runs, and we employed the term weighting formula (3) for all PRF procedures. Let y_t be the frequency of a given term in the query. If the top-ranked term was already included in the set of search terms, the term frequency in the query was changed into $1.5 \times y_t$. If not, the term frequency was set to 0.5 (i.e., $y_t = 0.5$).

3.4 Type of Search Runs

As for dictionary-based transitive query translation via a pivot language, we executed three types of run as follows:

- (a) Two-stage refinement of translation with English disambiguation
- (b) Two-stage refinement of translation without English disambiguation (as in CLEF 2003)
- (c) No refinement

In order to comparatively evaluate the performance of our two-stage refinement method, we decided to use commercial MT software produced by a Japanese company[3]. In this case, first the original German query was entered into the software. The software we used executes German to English translation automatically and then English to French translation (i.e., a kind of transitive translation). The resulting French text from the software was processed according to the procedure described in Section 3.1, and finally, a set of normalized French 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 as in the case of dictionary-based translation.

Similarly, for comparison, we tried to execute French monolingual runs with post-translation query expansion.

We executed five runs in which <TITLE> and <DESCRIPTION> fields in each query were used, and submitted the results to the organizers of CLEF 2004. All runs were executed on the information retrieval system, ADOMAS (Advanced Document Management System) developed at Surugadai University in Japan.

4 Experimental Results

4.1 Basic Statistics

The target French collections include 90,261 documents in total. The average document length is 227.14 words. We also use the Glasgow Herald 1995 as a document set for English disambiguation. The English collection includes 56,742 documents and the average document length is 231.56. Other experimental settings are described in the overview [5].

4.2 Results

Scores of average precision and R-precision are shown in Table 1, and the recall-precision curves of each run are presented in Figure 2. Note that each value in Table 1 and Figure 2 is calculated for 49 topics.

As shown in Table 1, MT significantly outperforms dictionary-based translations, and its mean average precision (MAP) is .3368, which is 85.4% of that given by the monolingual run (.3944). Although the performance of the dictionary-based approach using free dictionaries downloaded from the Internet is lower than that of the MT approach, Table 1 shows that two-stage refinements improve the effectiveness of the

[3] http://www.crosslanguage.co.jp/english/

Table 1. Average precision and R-precision (49 topics)

Run	ID	Average Precision	R-Precision
French Monolingual	NiiFF01	.3944	.3783
MT	NiiMt02	.3368	.3125
Dictionary 1: Two-stage refinement with English disambiguation	NiiDic03	.2690	.2549
Dictionary 2: Two-stage refinement without English disambiguation	NiiDic04	.2746	.2542
Dictionary 3: No refinement	NiiDic05	.1015	.1014

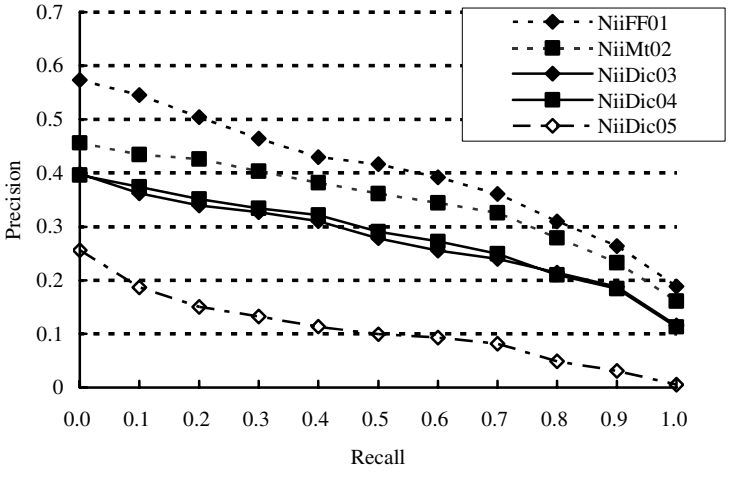

Fig. 2. Recall-precision curves

dictionary-based translation method, similar to our CLEF 2003 experiment. That is, the MAP score of NiiDic05 with no refinement is .1015, and NiiDic03 (with English disambiguation) and NiiDic04 (with no English disambiguation) significantly outperform NiiDic05.

However, English disambiguation appears to have almost no effect. The MAP score of NiiDic03 is .2690, which is slightly inferior to that of NiiDic04 (.2746), and clearly there is no statistically significant difference between them. Figure 3 shows average precision scores of NiiDic03 and NiiDic04 for each topic with some topic numbers. The x-axis represents the average precision score of NiiDic03, and the y-axis indicates that of NiiDic04. Therefore, each dot shows a pair of scores of NiiDic03 and NiiDic04 for a topic. For most topics, the scores of NiiDic03 are almost the same as those of NiiDic04. However, for topics 213, 245, 203, 229 and 206, NiiDic04 outperforms NiiDic03 significantly. On the other hand, scores of NiiDic03 are higher than those of NiiDic04 for topics 231, 233 and 242.

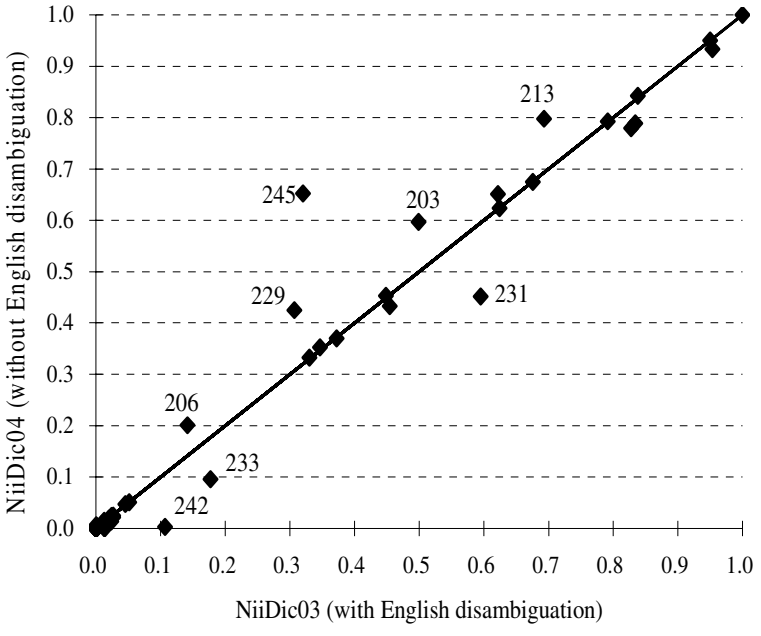

Fig. 3. Topic-by-topic analysis (average precision score)

Table 2 is a list of French translations in NiiDic03 and NiiDic04 for topic 245, in which the difference of average precision scores between NiiDic03 and NiiDic04 is the largest among the 49 topics. The <TITLE> field of topic 245 is "Christopher Reeve," and the text in the <DESC> field is "Finde Dokumente über die Karriere des Schauspielers Christopher Reeve und den Unfall der zu seiner Lähmung führte." It

Table 2. Example of French translations – topic 245

German Terms	NiiDic03 (with English disambiguation)	NiiDic04 (without English disambiguation)
christoph	christoph	christoph
fuhrt	fuhrt	fuhrt
karri	carri	carri
lahm	boiteux	boiteux
lahmung	lahmung	lahmung
reev	reev	reev
rist	instep	instep
schauspiel	jou	jou
uber	uber	uber
unfall	casualt	mésaventur
ung	contrecoeur	contrecoeur

turns out that both methods provide us with the same set of French translations except for “casualt” and “mésaventur.” In topic 245, the shift of just one term has a large effect on the retrieval performance. However, on average, the English disambiguation process did not yield a drastic change in the resulting final set of terms. In view of the fact that disambiguating English translations increases the processing time, we conclude that English disambiguation in the pivot language approach has no advantage for improving information retrieval systems.

5 Conclusions

This paper reported the results of our experiment on CLIR from German to French, in which English was used as a pivot language. Two-stage refinement of query translation was employed to remove irrelevant terms in the target language produced by transitive translation using two bilingual dictionaries successively and to expand the set of translations. In particular, in CLEF 2004, disambiguation of English terms in the intermediate process of transitive translation was attempted. The results showed that:

- our two-stage refinement method significantly improves retrieval performance of bilingual IR using a pivot language, and
- English disambiguation has almost no effect.

Intuitively, English disambiguation is promising because theoretically, removing erroneous English terms should effectively prevent irrelevant terms from spreading in the final set of search terms in the target language. However, our experimental results indicate that English disambiguation is useless. Further research is needed.

References

1. Kishida, K., Kando, N.: Two stages refinement of query translation for pivot language approach to cross lingual information retrieval: a trial at CLEF 2003. In Working Notes for the CLEF 2003 Workshop (2003) 129-136
2. Ballesteros, L., Croft, W.B.: Resolving ambiguity for cross-language retrieval. In Proceedings of the 21st ACM SIGIR conference on Research and Development in Information Retrieval (1988) 64-71
3. Porter, M.F.: An algorithm for suffix stripping. Program. 14 (1980) 130-137
4. 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/
5. Peters, C., Braschler, M., Di Nunzio, G., Ferro, N.: CLEF 2004: Ad hoc track overview and results analysis. In Peters, C., Clough, P., Gonzalo, J., Jones, G., Kluck, M., Magnini, B., Proceedings of Fifth Workshop of the Cross-Language Evaluation Forum (CLEF 2004), Lecture Notes in Computer Science (LNCS), Springer, Heidelberg, Germany. This volume.