A Comparative Study of Query and Document Translation for Cross-Language Information Retrieval

Douglas W. Oard[1]

College of Library and Information Services
University of Maryland, College Park, MD 20742
oard@glue.umd.edu
WWW home page: http://www.glue.umd.edu/~oard/

Abstract. Cross-language retrieval systems use queries in one natural language to guide retrieval of documents that might be written in another. Acquisition and representation of translation knowledge plays a central role in this process. This paper explores the utility of two sources of translation knowledge for cross-language retrieval. We have implemented six query translation techniques that use bilingual term lists and one based on direct use of the translation output from an existing machine translation system; these are compared with a document translation technique that uses output from the same machine translation system. Average precision measures on a TREC collection suggest that arbitrarily selecting a single dictionary translation is typically no less effective than using every translation in the dictionary, that query translation using a machine translation system can achieve somewhat better effectiveness than simpler techniques, and that document translation may result in further improvements in retrieval effectiveness under some conditions.

1 Introduction

As international markets and rapidly expanding trans-national information networks interact, an imperative for access to information written in many languages is becoming increasingly apparent. Cross-Language Information Retrieval (CLIR), the detection of relevant documents in one natural language using queries expressed in another, provides an important capability that can help meet that challenge [9]. Two principal lines of CLIR research have emerged, approaches which exploit explicit representations of translation knowledge (such as bilingual dictionaries or machine translation lexicons) and those which seek to extract useful translation knowledge from training corpora using representations such as cooccurrence matrices that are not designed for direct human interpretation. We refer to the approaches in the first group as "knowledge-based" and those in the second group as "corpus-based." Carbonell, et al. have reported excellent results when corpus-based techniques are evaluated on a held-back portion of the corpus from which the translation knowledge was extracted [3].

On the other hand, we have previously investigated the retrieval effectiveness of corpus-based CLIR and found that domain shift effects can adversely affect retrieval effectiveness when translation knowledge is acquired from one corpus and then used for retrieval from a different collection [7]. Knowledge-based techniques that exploit broad coverage resources such as dictionaries and machine translation lexicons appear to be less sensitive to this effect, so their use may be preferred when domain-specific training corpora are not available. Rather than focus further on corpus-based techniques, in this paper we explore the performance of several knowledge-based CLIR approaches.

There are four fundamental strategies for knowledge-based CLIR: direct matching of terms in different languages without translation, translation of each query into every document language, translation of each document into every possible query language, and translation of each query and each document into some common representation. Cognate matching, in which knowledge about related word forms in a pair of languages is encoded directly into the query-document matching algorithm, is an example of the first strategy (c.f., [2]), and controlled vocabulary retrieval using indexing and search terms chosen from a domain-specific multilingual thesaurus is an example of the last strategy (c.f. [12]). Although cognate matching offers a useful way of handling unknown words, the semantically meaningful lexical regularities on which it depends are presently known in only a few language pairs so we have not yet chosen to focus a study on that approach. Similarly, although controlled vocabulary retrieval has proven to be useful in limited domains, achieving broad coverage using a controlled vocabulary is difficult. We have thus chosen to focus this study on query translation and document translation strategies.

Over the past several years, query translation has emerged as the most popular strategy for fully automatic broad coverage CLIR [8]. Query translation can be quite efficient when short queries are presented, but simple query translation approaches suffer a severe penalty in effectiveness, usually achieving about half of the retrieval effectiveness of corresponding monolingual techniques when typical measures such as average precision are used. A number of studies have reported that simple linguistic processing such as limiting candidate translations for query terms to those with the same part of speech, or indexing phrases as well as individual words, can raise this performance to perhaps 75% of the monolingual effectiveness (c.f., [4, 6]). In this paper we describe a query translation technique based on the output of an existing Machine Translation (MT) system and compare that approach with some more efficient dictionary-based query translation techniques.

A document translation strategy in which fully automatic MT is used to translate each document into a single language (the query language) at indexing time may be attractive for interactive applications if the users need to rapidly skim retrieved documents in their preferred language. This is a requirement that query translation strategies can not presently support (with translation rates at a substantial fraction of a minute minute per page on typical workstations). By contrast, a document translation strategy in which the full text of the transla-

tions is available for immediate display could easily provide adequate response times. Document translation may also improve retrieval effectiveness if the MT system is able to exploit linguistic context to choose correct translations more often in documents than in queries. Since queries are sometimes quite short and are often not well formed sentences, there is reason to suspect that the promised improvement may be realized. We have tested this hypothesis by implementing a document translation technique and comparing it with our query translation techniques.

In addition to the query translation and document translation techniques, we also implemented two baseline techniques without any translation component: query construction in the same language as the documents, and the presentation of queries in a language different from that of the documents. We expect the first to provide an upper bound for CLIR effectiveness and the second to provide a lower bound.

The next section presents our experiment design and describes the techniques we have implemented in detail. We have learned that arbitrarily selecting a single translation from a bilingual dictionary can be as effective as more commonly implemented techniques based on retaining every possible translation, that techniques based on loosely coupling machine translation and information retrieval perform somewhat better than simple dictionary-based techniques, and that document translation can outperform query translation under some conditions. Section 3 describes these results in detail, and the paper concludes with a discussion of the implications for further work on cross-language information retrieval.

2   Experiment Design

Earlier CLIR evaluations have been hampered by inadequate test collections, but the Text REtrieval Conference (TREC-6) recently developed the first large-scale multilanguage collection that is designed specifically to support CLIR experiments. We have used the German documents ("SDA/NZZ") from that collection for the majority of our experiments, supplemented where practical by the corresponding English collection ("AP").

The TREC-6 CLIR SDA/NZZ collection contains 251,840 German newswire articles from two Swiss news agencies. The SDA documents are from 1988, 1989 and 1990, and the NZZ documents are from 1994. Some information need specifications (topic descriptions) are available in German, English, and three other languages. Relevance judgments were made for each topic by at the National Institutes of Standards and Technology (NIST) using a pooled assessment methodology in which each of the top one hundred documents from four different monolingual German retrieval systems were evaluated for relevance to a topic description similar to that in Figure 1. Documents not in that set were presumed not to be relevant for the purpose of computing recall and precision. The process was repeated for 22 topics and relevant documents were discovered

for 21 of those topics.[1] Three standard sources for query terms were defined at TREC-6: "title queries" are formed from the one to three words in the "title" field, "short queries" are formed from only the one or two sentences or sentence fragments in the "desc" field, and "long queries" are formed using every word in the "title," "desc," and "narr" fields. We report results below for title queries and long queries.[2]

```
<top>
<num> Number: CL1

<E-title> Waldheim Affair

<E-desc> Description: Reasons for controversy surrounding Waldheim's World
War II actions.

<E-narr> Narrative: Revelations about Austrian President Kurt Waldheim's
participation in Nazi crimes during World War II are argued on both sides.
Relevant documents are those that express doubts about the truth of these
revelations. Documents that just discuss the affair are not relevant.
</top>
```

Fig. 1. The English version of TREC-6 topic CL01.

The TREC-6 CLIR AP collection contains 242,918 articles from the Associated Press newswire service in the United States that were generated in 1988, 1989 and 1990. The collection has been assessed at NIST for the same 22 topics using a pooled assessment methodology based on the top one hundred documents from five different monolingual English retrieval systems. Relevant documents are known in the AP collection for the same 21 topics as for the SDA/NZZ collection.

For text retrieval we ran version 3.1p1 of the Inquery system from the University of Massachusetts on a single SPARC 20 under the Solaris 2.5 operating system. The Inquery "kstem" stemmer and the standard English Inquery stopword list were used when processing the AP documents and when processing English translations of the SDA/NZZ documents. No stemmer or stopword list

---

[1] The TREC-6 CLIR evaluation originally included 25 topics. No relevant documents were discovered in the SDA/NZZ collection for topic CL22, and relevance judgments are not available for topics CL03, CL15 and CL25.

[2] We omit results for short queries on the SDA/NZZ collection because we discovered at TREC-6 that the "desc" field often fails to perform well in monolingual German evaluations, presumably because the topic descriptions were constructed by amplifying rather than repeating earlier information [10].

was used when processing SDA/NZZ documents in the original German, and no techniques for splitting German compounds were implemented.[3]

2.1 Same Language Query (SLQ)

To approximate an upper bound for the performance of any CLIR system, we compared the retrieval effectiveness of our four experimental approaches with the retrieval effectiveness achieved by using queries that are given in the same language as the documents. For example, the CL01 "title query" would be presented as [waldheim affair] when retrieving English AP documents and as [die affaire waldheim] when retrieving German SDA/NZZ documents.

2.2 Dictionary-Based Query Translation (DQT)

By far the most commonly used query translation approach is to replace each query term with appropriate translations that are automatically extracted from an online bilingual dictionary (c.f., [6, 1]). The usual approach is to automatically extract a bilingual term list from the dictionary entries, so for translating queries from English into German for retrieval from the SDA/NZZ collection we used an online bilingual term list developed by Stefan Büdenbender.[4] That term list contains 131,274 bilingual pairs in which each pair consists of one word or phrase in English and the corresponding word or phrase in German. The number of unique words in the list is far smaller than 131,274 because many words appear in several bilingual pairs and the number of unique stems is smaller still because the dictionary contains multiple morphological variants for many of the words. The pairs were initially sorted in lexicographic order based on the English terms and we used the same dictionary to translate queries from German into English for retrieval from the AP collection after resorting the pairs by the German terms.

It is common for a single word to have several translations, some with very different meanings. Bilingual dictionaries typically seek to help users select appropriate translations of individual words by embedding the word in a representative phrase, and this practice was present in the bilingual term list that we used. It is not at all clear how one should design an algorithm to extract only the "appropriate" translations using this information, so we have implemented six simple dictionary-based query translation techniques that together explore the effects of winner-take-all, word-match and stem-match approaches. We illustrate the effect of each technique with a German translation of the English CL01 title query given above.

Single Word (SW) Bilingual term lists provide no obvious basis for selecting a single translation when more than one alternative translation for a word

---

[3] We tried a small German stopword list in our TREC-6 experiments and found that it hurt average precision somewhat in most cases [10].

[4] The bilingual term list is available at http://www.bg.bib.de/∼a2h6bu/

is encoded in the list. In SW we arbitrarily choose the first exact single whole-word match in the list.[5] The list is sorted in alphabetical order, and we expect this technique to perform about as well as any other arbitrary choice of a single word. Words which are not found in the dictionary are retained unchanged, a simple cognate matching strategy that often works well for proper names.

[waldheim affäre]

Single Word, Stemmed (SWS) The bilingual term list we have used contains several morphological variants for each word rather than a single entry for a root form, but it is possible that some required morphological variants may not be present. Accordingly, for the SWS technique we first seek an exact match for each term, and if that fails we stem every word in the bilingual term list and in the query and then try the matching again.[6] If that fails we retain the word unchanged in the hopes of a cognate match.

[waldheim affäre]

Every Word (EW) SW involves arbitrary choices, but information retrieval algorithms are able to accept multiple possibilities. Thus, the more common technique for using bilingual term lists has been to retain every possible translation when more than one alternative is present in the list. In the EW technique we replace each word with every exact single whole-word match in the bilingual term list.

[waldheim affäre angelegenheit ereigneis geschäft handlung sache]

Every Word, Stemmed (EWS) EWS is the stemmed variant of EW, in which we retain every exact single stem match in the dictionary. This is done in a single pass, rather than the two pass approach used in SWS, since that approach seems to better match the idea of "every" possible translation.

[waldheim affäre angelegenheit angelegenheiten ereigneis geschäft handlung sache]

Every Phrase (EP) Like many dictionaries, our bilingual term list contains phrases in addition to single words. Phrases are ignored, however, on the source language side of the bilingual term list in the preceding techniques since only single word matches are used. Because some query words may appear only as part of a phrase, in the EP technique we include the translation any time the query word exactly matches any word in the bilingual term list, regardless of whether that word appears alone or as part of a phrase. In order to prevent an explosion of nuisance matches, words which appear in our stopword list are not translated.[7]

---

[5] An "exact" match is one in which the two character strings are the same length and each character in the two strings matches. A "whole word" is any whitespace-delimited string of characters that appears in the document.

[6] Stemming is an automatic suffix removal technique. We used the Porter stemmer for English that is available from ftp://ftp.vt.edu/pub/reuse/IR.code/ for this purpose.

[7] Stopwords are common words that are of little benefit to information retrieval. We used the standard English stopword list supplied with the Inquery system for this purpose.

```
[waldheim affäre angelegenheit ereigneis geschäft handlung
sache ehrensache familienagelegenheit liebesglück es war eine
abgekartete sache es ging heiss her liebesaffäre liebeserlebnis
techtelmechtel staatsangelegenheit das ist meine sache]
```
Every Phrase, Stemmed (EPS) EPS is the stemmed analogue of EP in which we retain every exact stem match in the bilingual term list, regardless of whether the stemmed word appears alone or as part of a phrase. Again, only one pass is needed.

```
[waldheim affäre angelegenheit angelegenheiten ereigneis
geschäft handlung sache ehrensache familienagelegenheit
liebesglück es war eine abgekartete sache es ging heiss her
liebesaffäre liebeserlebnis techtelmechtel mein
privatangelegenheiten staatsangelegenheit staatsangelegenheiten
bescherung das ist meine sache seine angelegenheiten in ordnung
bringen geschäfte abwickeln]
```

In every case we replace each word in the query with the corresponding word or phrase in every matching bilingual pair to produce a version of the query that can be compared with the documents in the collection. In addition to simple word-to-word mappings, word-to-phrase mappings are possible (and, in fact, common), so translated queries are typically longer than untranslated queries and the translated queries sometimes contain repeated words. Furthermore, the translated queries often contain multiple words with the same stems, and in English (but not in German) these words will be treated by our information retrieval system as if they are identical.

2.3   MT-Based Query Translation (MQT)

Machine translation systems seek to translate documents from one language to another, either as an aid for human translators or for direct use as a fairly rapid and inexpensive rough translation. This provides an obvious approach to query translation, but we are aware of only one prior experiment to use such a technique [11]. In that experiment, Radwan and Fluhr compared the retrieval effectiveness of queries translated from French into English by the SYSTRAN machine translation system with the effectiveness of their EMIR dictionary-based query translation system using a version of the small Cranfield collection for which French queries were available. In that study they found that the EMIR was more effective than their MT-based query translation technique using SYS-TRAN. Our experiments offer some insight into the performance of a MT-based query translation approach on larger test collections.

The Logos machine translation system that we used for our experiments is a commercial product that is designed to assist human translators by automatically preparing fairly good translations of individual documents.[8] The system is

---

[8] Logos Corporation, 111 Howard Boulevard, Suite 214, Mount Arlington, NJ 07856 USA

typically used by translation bureaus and other organizations as the first stage of a machine-assisted translation process, and we have previously used it for cross-language routing experiments [7]. The Logos system includes extensive facilities for adding domain-specific technical terminology and new linguistic constructs, but for the experiments reported here we used only the machine readable dictionaries and semantic rules that are delivered as standard components of the product.

We used the Logos system to translate English queries into German for use with the SDA/NZZ collection and to translate German queries into English for use with the AP collection. Since the Logos system is designed to generate readable translation, it generates only a single "best guess" translation for any input. Thus MT-based query translation is most similar to the DQT-SW technique in which a single candidate translations is retained.

2.4 MT-Based Document Translation (MDT)

Our MT-based document translation approach parallels the design of our MT-based query translation design. We have selected English as a query language and translated each SDA/NZZ document into English as a preprocessing step. We then indexed the translated document collection and used English queries for the retrieval experiments. Essentially the preprocessing step reduces cross-language retrieval to a (possibly degraded) monolingual case. We used four SPARC 20 workstations and a fifth workstation that was upgraded from a SPARC 5 to a SPARC Ultra 1 after about three quarters of the documents had been translated.[9] Translation of the 48 months of newswire stories contained in the SDA and NZZ collections using these machines required approximately 10 machine-months, and successful translations were obtained for 251,572 documents. The remaining 268 documents were omitted from the translated collection.

2.5 Foreign Language Query (FLQ)

Monolingual information retrieval systems sometimes produce useful results because of fortuitous matches between words in different languages, proper names that are rendered in the same way in different languages, and foreign language terms in the documents that happen to be in the query language. For example, the English version of the CL01 title query shown above contains the proper name "Waldheim" which also often appears in relevant German documents. In order to establish a practical lower bound on retrieval effectiveness we have used both untranslated queries and untranslated documents to reveal the effect of these cognate matches.

3 Results

Table 1 summarizes the non-interpolated average precision results for the SDA/NZZ collection using every technique, averaged over the 21 topics for which relevant

---

[9] The translated documents are available to TREC participants from NIST.

documents are known. For title queries the advantage of same language queries over four of the eight CLIR techniques is statistically significant (with 95% confidence), as is the difference between three of the CLIR techniques and foreign language queries, but the available 21 queries are not sufficient to produce statistically significant differences among the CLIR techniques that we have tested. It does appear, however, that DQT-SW is no worse than the more commonly implemented DQT-EW technique, and that the same pattern is evident in the stemmed variant of each technique and with long query as well. These figures are averaged over 21 queries, however, and that obscures query-by-query variations. DQT-SW obtains this average performance by doing quite well on some queries and quite poorly on others. Since the average precision is much closer to zero (the minimum possible) than to one (the maximum), the potential gain for an individual query is far greater than the potential loss. Thus, a few exceptionally good translations could account for this effect. So in one sense, this result points up a weakness in the average precision measure. Viewed from another perspective, however, our result suggests that seeking an improvement over arbitrary choice may be as useful as the more common approach of seeking to cut down on the number of translations selected (c.f., [1]).

Technique	Query Length	
	Title	Long
SLQ	0.2480	0.2396
DQT-SW	0.1749	0.1342
DQT-SWS	0.1542	0.0969
DQT-EW	0.1778	0.1312
DQT-EWS	0.1363	0.0827
DQT-EP	0.1152	0.0165
DQT-EPS	0.1172	0.0182
MQT	0.1668	0.1561
MDT	0.1761	0.2171
FLQ	0.0307	0.0117

Table 1. Non-interpolated average precision for the SDA/NZZ collection, averaged over 21 topics.

Machine translation also seems to be doing well. On long queries, MT-based query translation outperforms every DQT technique. Title queries, which are all three words or less, lack the same effect. That is not surprising, since the machine translation system that we used is designed to perform best on well formed sentences. The effect of greater context is also apparent in the performance of MT-based document translation, which outperforms MT-based query translation on both title and long queries. This difference may, in fact, be understated somewhat because our experience suggests that English to German translations

are noticeably better in many cases than German to English translations with the system that we used. Since MT-based query translation used English to German translations and MT-based document translation used German to English translations, we might have seen an even bigger advantage for MT-based document translation with a more evenly balanced translation performance.

In order to seek confirmation for these results we applied three of our CLIR techniques to the English AP collection. Table 2 summarizes the non-interpolated average precision results for that collection, using DQT-SW, DQT-EW and MT-based query translation. DQT-EP was omitted because we lacked a usable stop-word list in German, the stemmed variants were omitted because both German stemming software and a compound splitting technique would have been needed, and replicating MT-based document translation was impractical within the time frame of this study. For those techniques that we were able to easily implement, the same trends were evident on the AP collection as on the SDA/NZZ collection. Again, DQT-SW was no worse than DQT-EW, and that MT-based query translation performs somewhat better than either of those techniques on long queries. Thus although we have not obtained statistically significant results, we now have some reason to believe that two of our most important observations are repeatable.

Technique	Query Length	
	Title	Long
SLQ	0.3449	0.3958
DQT-SW	0.1982	0.1154
DQT-EW	0.1805	0.0710
MQT	0.1928	0.2455
FLQ	0.0105	0.0132

Table 2. Non-interpolated average precision for the AP collection, averaged over 21 topics.

4   Conclusions

We have conducted an extensive evaluation of eight cross-language information retrieval techniques and found some interesting results. When using a bilingual term list that contains morphological variants rather than root forms, we have seen that matching exact words is better than matching stems. We have seen the same result with a different bilingual term list in a different language pair (English and Spanish) [5], so there is reason to believe that this result will generalize to other bilingual term lists and evaluation collections. Perhaps our most surprising result is that arbitrarily choosing a single translation for each term produces

the same average precision as the more common use of every possibly translation. Although we believe that to be an artifact of the average precision measure, it does offer an interesting perspective from which to think about the design of dictionary-based query translation techniques. Finally, and perhaps most importantly, we have observed that a sophisticated machine translation system can outperform simpler techniques for cross-language information retrieval. Our inability to easily replicate the MT-based document translation experiment on the AP collection (an estimated 10 machine-months of computation would have been required) speaks volumes about the practical limitations of that approach, however. But as machine translation becomes faster, we have demonstrated one way in which those powerful capabilities might be used. It is clearly possible to exploit these same resources that we have used to craft more sophisticated techniques. For example, we could take advantage of redundancy in the bilingual term list to improve our translation choices in the DQT-SW method. And in MT-based query translation and MT-based document translation we could preserve some additional terms in the face of unresolvable ambiguity by coupling the translation and retrieval systems more tightly. Our encouraging results suggest that these would be promising directions for future work.

Acknowledgments

References

1. Ballesteros, L. and Croft, W. B. (1997). Phrasal translation and query expansion techniques for cross-language information retrieval. In *Proceedings of the 20th International ACM SIGIR Conference on Research and Development in Information Retrieval*. http://ciir.cs.umass.edu/.
2. Buckley, C., Mitra, M., Walz, J., and Cardie, C. (1997). Using clustering and SuperConcepts within SMART: TREC 6. In *The Sixth Text REtrieval Conference (TREC-6)*. National Institutes of Standards and Technology. To appear. http://trec.nist.gov/.
3. Carbonell, J., Yang, Y., Frederking, R., Brown, R. D., Geng, Y., and Lee, D. (1997). Translingual information retrieval: A comparative evaluation. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*. http://www.cs.cmu.edu/afs/cs.cmu.edu/user/ralf/pub/WWW/papers.html.
4. Davis, M. and Ogden, W. C. (1997). Quilt: Implementing a large-scale cross-language text retrieval system. In *Proceedings of the 20th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

5. Dorr, B. J. and Oard, D. W. (1998). Evaluating resources for query translation in cross-language information retrieval. In *Proceedings of the First International Conference on Language Resource Evaluation*. http://www.glue.umd.edu/∼oard/.

6. Hull, D. A. and Grefenstette, G. (1996). Querying across languages: A dictionary-based approach to multilingual information retrieval. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. http://www.xrce.xerox.com/people/hull/papers/sigir96.ps.

7. Oard, D. W. (1997a). Adaptive filtering of multilingual document streams. In *Fifth RIAO Conference on Computer Assisted Information Searching on the Internet*. http://www.glue.umd.edu/∼oard/.

8. Oard, D. W. (1997b). Alternative approaches for cross-language text retrieval. In *AAAI Symposium on Cross-Language Text and Speech Retrieval*. American Association for Artificial Intelligence. http://www.glue.umd.edu/∼oard/.

9. Oard, D. W. (1997c). Serving users in many languages: Cross-language information retrieval for digital libraries. *D-Lib Magazine*. http://www.dlib.org.

10. Oard, D. W. and Hackett, P. G. (1997). Document translation for cross-language text retrieval at the University of Maryland. In *The Sixth Text REtrieval Conference (TREC-6)*. National Institutes of Standards and Technology. To appear. http://trec.nist.gov/.

11. Radwan, K. and Fluhr, C. (1995). Textual database lexicon used as a filter to resolve semantic ambiguity application on multilingual information retrieval. In *Fourth Annual Symposium on Document Analysis and Information Retrieval*, pages 121–136.

12. Soergel, D. (1997). Multilingual thesauri in cross-language text and speech retrieval. In *AAAI Symposium on Cross-Language Text and Speech Retrieval*. American Association for Artificial Intelligence. http://www.clis.umd.edu/dlrg/filter/sss/papers/.