Should we Translate the Documents or the Queries in Cross-language Information Retrieval?

J. Scott McCarley
IBM T.J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598
jsmc@watson.ibm.com

Abstract

Previous comparisons of document and query translation suffered difficulty due to differing quality of machine translation in these two opposite directions. We avoid this difficulty by training identical statistical translation models for both translation directions using the same training data. We investigate information retrieval between English and French, incorporating both translations directions into both document translation and query translation-based information retrieval, as well as into hybrid systems. We find that hybrids of document and query translation-based systems outperform query translation systems, even human-quality query translation systems.

1 Introduction

Should we translate the documents or the queries in cross-language information retrieval? The question is more subtle than the implied two alternatives. The need for translation has itself been questioned : although non-translation based methods of cross-language information retrieval (CLIR), such as cognate-matching (Buckley et al., 1998) and cross-language Latent Semantic Indexing (Dumais et al., 1997) have been developed, the most common approaches have involved coupling information retrieval (IR) with machine translation (MT). (For convenience, we refer to dictionary-lookup techniques and interlingua (Diekema et al., 1999) as "translation" even if these techniques make no attempt to produce coherent or sensibly-ordered language; this distinction is important in other areas, but a stream of words is adequate for IR.) Translating the documents into the query's language(s) and translating the queries into the document's language(s) represent two extreme approaches to coupling MT and IR. These two approaches are neither equivalent nor mutually exclusive. They are not equivalent because machine translation is not an invertible operation. Query translation and document translation become equivalent only if each word in one language is translated into a unique word in the other languages. In fact machine translation tends to be a *many-to-one* mapping in the sense that finer shades of meaner are distinguishable in the original text than in the translated text. This effect is readily observed, for example, by machine translating the translated text back into the original language. These two approaches are not mutually exclusive, either. We find that a hybrid approach combining both directions of translation produces superior performance than either direction alone. Thus our answer to the question posed by the title is *both*.

Several arguments suggest that document translation should be competitive or superior to query translation. First, MT is error-prone. Typical queries are short and may contain key words and phrases only once. When these are translated inappropriately, the IR engine has no chance to recover. Translating a long document offers the MT engine many more opportunities to translate key words and phrases. If only some of these are translated appropriately, the IR engine has at least a chance of matching these to query terms. The second argument is that the tendency of MT

engines to produce fewer distinct words than were contained in the original document (the output vocabulary is smaller than the input vocabulary) also indicates that machine translation should preferably be applied to the documents. Note the types of preprocessing in use by many monolingual IR engines: stemming (or morphological analysis) of documents and queries reduces the number of distinct words in the document index, while query expansion techniques *increase* the number of distinct words in the query.

Query translation is probably the most common approach to CLIR. Since MT is frequently computationally expensive and the document sets in IR are large, query translation requires fewer computer resources than document translation. Indeed, it has been asserted that document translation is simply impractical for large-scale retrieval problems (Carbonell et al., 1997), or that document translation will only become practical in the future as computer speeds improve. In fact, we have developed fast MT algorithms (McCarley and Roukos, 1998) expressly designed for translating large collections of documents and queries in IR. Additionally, we have used them successfully on the TREC CLIR task (Franz et al., 1999). Commercially available MT systems have also been used in large-scale document translation experiments (Oard and Hackett, 1998). Previously, large-scale attempts to compare query translation and document translation approaches to CLIR (Oard, 1998) have suggested that document translation is preferable, but the results have been difficult to interpret. Note that in order to compare query translation and document translation, two different translation systems must be involved. For example, if queries are in English and document are in French, then the query translation IR system must incorporate *English*⇒*French* translation, whereas the document translation IR system must incorporate *French*⇒*English*. Since familiar commercial MT systems are "black box" systems, the quality of translation is not known *a priori*. The present work avoids this difficulty by using statistical machine translation systems for both directions that are trained on the same training data using identical procedures. Our study of document translation is the largest comparative study of document and query translation of which we are currently aware. We also investigate both query and document translation for both translation directions within a language pair.

We built and compared three information retrieval systems : one based on document translation, one based on query translation, and a hybrid system that used both translation directions. In fact, the "score" of a document in the hybrid system is simply the arithmetic mean of its scores in the query and document translation systems. We find that the hybrid system outperforms either one alone. Many different hybrid systems are possible because of a tradeoff between computer resources and translation quality. Given finite computer resources and a collection of documents much larger than the collection of queries, it might make sense to invest more computational resources into higher-quality query translation. We investigate this possibility in its limiting case: the quality of human translation exceeds that of MT; thus monolingual retrieval (queries and documents in the same language) represents the ultimate limit of query translation. Surprisingly, we find that the hybrid system involving fast document translation and monolingual retrieval continues to outperform monolingual retrieval. We thus conclude that the hybrid system of query and document translation will outperform a pure query translation system no matter how high the quality of the query translation.

2 Translation Model

The algorithm for fast translation, which has been described previously in some detail (McCarley and Roukos, 1998) and used with considerable success in TREC (Franz et al., 1999), is a descendent of IBM Model 1 (Brown et al., 1993). Our model captures important features of more complex models, such as fertility (the number of French words

output when a given English word is translated) but ignores complexities such as distortion parameters that are unimportant for IR. Very fast decoding is achieved by implementing it as a direct-channel model rather than as a source-channel model. The basic structure of the *English⇒French* model is the probability distribution

$$p(n_i; f_1 \ldots f_{n_i} | e_i, context(e_i)). \quad (1)$$

of the fertility n_i of an English word e_i and a set of French words $f_1 \ldots f_{n_i}$ associated with that English word, given its context. Here we regard the context of a word as the preceding and following non-stop words; our approach can easily be extended to other types of contextual features. This model is trained on approximately 5 million sentence pairs of Hansard (Canadian parliamentary) and UN proceedings which have been aligned on a sentence-by-sentence basis by the methods of (Brown et al., 1991), and then further aligned on a word-by-word basis by methods similar to (Brown et al., 1993). The *French⇒English* model can be described by simply interchanging English and French notation above. It is trained separately on the same training data, using identical procedures.

3 Information Retrieval Experiments

The document sets used in our experiments were the English and French parts of the document set used in the TREC-6 and TREC-7 CLIR tracks. The English document set consisted of 3 years of AP newswire (1988-1990), comprising 242918 stories originally occupying 759 MB. The French document set consisted of the same 3 years of SDA (a Swiss newswire service), comprising 141656 stories and originally occupying 257 MB. Identical query sets and appropriate relevance judgments were available in both English and French. The 22 topics from TREC-6 were originally constructed in English and translated by humans into French. The 28 topics from TREC-7 were originally constructed (7 each from four different sites) in English, French, German, and Italian, and human translated into all four languages. We have no knowledge of which TREC-7 queries were originally constructed in which language. The queries contain three SGML fields (<topic>, <description>, <narrative>), which allows us to contrast short (<description> field only) and long (all three fields) forms of the queries. Queries from TREC-7 appear to be somewhat "easier" than queries from TREC-6, across both document sets. This difference is not accounted for simply by the number of relevant documents, since there were considerably fewer relevant French documents per TREC-7 query than per TREC-6 query.

With this set of resources, we performed the two different sets of CLIR experiments, denoted *EqFd* (English queries retrieving French documents), and *FqEd* (French queries retrieving English documents.) In both *EqFd* and *FqEd* we employed both techniques (translating the queries, translating the documents). We emphasize that the query translation in *EqFd* was performed with the same *English⇒French* translation system as the document translation in *FqEd*, and that the document translation *EqFd* was performed with the same *French⇒English* translation system as the query translation in *FqEd*. We further emphasize that both translation systems were built from the same training data, and thus are as close to identical quality as can likely be attained. Note also that the results presented are not the TREC-7 CLIR task, which involved both cross-language information retrieval and the merging of documents retrieved from sources in different languages.

Preprocessing of documents includes part-of-speech tagging and morphological analysis. (The training data for the translation models was preprocessed identically, so that the translation models translated between morphological root words rather than between words.) Our information retrieval systems consists of first pass scoring with the Okapi formula (Robertson et al., 1995) on unigrams and symmetrized bigrams (with

en, *des*, *de*, and - allowed as connectors) followed by a second pass re-scoring using local context analysis (LCA) as a query expansion technique (Xu and Croft, 1996). Our primary basis for comparison of the results of the experiments was TREC-style average precision after the second pass, although we have checked that our principal conclusions follow on the basis of first pass scores, and on the precision at rank 20. In the query translation experiments, our implementation of query expansion corresponds to the post-translation expansion of (Ballasteros and Croft, 1997), (Ballasteros and Croft, 1998). All adjustable parameters in the IR system were left unchanged from their values in our TREC ad-hoc experiments (Chan et al., 1997),(Franz and Roukos, 1998), (Franz et al., 1999) or cited papers (Xu and Croft, 1996), except for the number of documents used as the basis for the LCA, which was estimated at 15 from scaling considerations. Average precision for both query and document translation were noted to be insensitive to this parameter (as previously observed in other contexts) and not to favor one or the other method of CLIR.

4 Results

In experiment $EqFd$, document translation outperformed query translation, as seen in columns *qt* and *dt* of Table 1. In experiment $FqEd$, query translation outperformed document translation, as seen in the columns *qt* and *dt* of Table 2. The relative performances of query and document translation, in terms of average precision, do not differ between long and short forms of the queries, contrary to expectations that query translation might fair better on longer queries. A more sophisticated translation model, incorporating more nonlocal features into its definition of context might reveal a difference in this aspect. A simple explanation is that in both experiments, *French*⇒*English* translation outperformed *English*⇒*French* translation. It is surprising that the difference in performance is this large, given that the training of the translation systems was identical. Reasons for this difference could be in the structure of the languages themselves; for example, the French tendency to use phrases such as *pomme de terre* for *potato* may hinder retrieval based on the Okapi formula, which tends to emphasize matching unigrams. However, separate monolingual retrieval experiments indicate that the advantages gained by indexing bigrams in the French documents were not only too small to account for the difference between the retrieval experiments involving opposite translation directions, but were in fact smaller than the gains made by indexing bigrams in the English documents. The fact that French is a more highly inflected language than English is unlikely to account for the difference since both translation systems and the IR system used morphologically analyzed text. Differences in the quality of preprocessing steps in each language, such as tagging and morphing, are more difficult to account for, in the absence of standard metrics for these tasks. However, we believe that differences in preprocessing for each language have only a small effect on retrieval performance. Furthermore, these differences are likely to be compensated for by the training of the translation algorithm: since its training data was preprocessed identically, a translation engine trained to produce language in a particular style of morphing is well suited for matching translated documents with queries morphed in the same style. A related concern is "matching" between translation model training data and retrieval set - the English AP documents might have been more similar to the Hansard than the Swiss SDA documents. All of these concerns heighten the importance of studying both translation directions within the language pair.

On a query-by-query basis, the scores are quite correlated, as seen in Fig. (1). On TREC-7 short queries, the average precisions of query and document translation are within 0.1 of each other on 23 of the 28 queries, on both $FqEd$ and $EqFd$. The remaining outlier points tend to be accounted for by simple translation errors, (e.g. *vol*

$EqFd$	qt	dt	$qt + dt$	ht	$ht + dt$
trec6.d	0.2685	0.2819	0.2976	0.3494	0.3548
trec6.tdn	0.2981	0.3379	0.3425	0.3823	0.3664
trec7.d	0.3296	0.3345	0.3532	0.3611	0.4021
trec7.tdn	0.3826	0.3814	0.4063	0.4072	0.4192

Table 1: Experiment $EqFd$: English queries retrieving French documents
All numbers are TREC average precisions.
qt : query translation system
dt : document translation system
$qt + dt$: hybrid system combining qt and dt
ht : monolingual baseline (equivalent to human translation)
$ht + dt$: hybrid system combining ht and dt

$FqEd$	qt	dt	$qt + dt$	ht	$ht + dt$
trec6.d	0.3271	0.2992	0.3396	0.2873	0.3369
trec6.tdn	0.3666	0.3390	0.3743	0.3889	0.4016
trec7.d	0.4014	0.3926	0.4264	0.4377	0.4475
trec7.tdn	0.4541	0.4384	0.4739	0.4812	0.4937

Table 2: Experiment $FqEd$: French queries retrieving English documents
All numbers are TREC average precisions.
qt : query translation system
dt : document translation system
$qt + dt$: hybrid system combining qt and dt
ht : monolingual baseline (equivalent to human translation)
$ht + dt$: hybrid system combining ht and dt

d'oeuvres d'art → *flight art* on the TREC-7 query CL.036.) With the limited number of queries available, it is not clear whether the difference in retrieval results between the two translation directions is a result of small effects across many queries, or is principally determined by the few outlier points.

We remind the reader that the query translation and document translation approaches to CLIR are not symmetrical. Information is distorted in a different manner by the two approaches, and thus a combination of the two approaches may yield new information. We have investigated this aspect by developing a hybrid system in which the score of each document is the mean of its (normalized) scores from both the query and document translation experiments. (A more general linear combination would perhaps be more suitable if the average precision of the two retrievals differed substantially.) We observe that the hybrid systems which combine query translation and document translation outperform both query translation and document translation individually, on both sets of documents. (See column $qt + dt$ of Tables 1 and 2.)

Given the tradeoff between computer resources and quality of translation, some would propose that correspondingly more computational effort should be put into query translation. From this point of view, a document translation system based on fast MT should be compared with a query translation system based on higher quality, but slower MT. We can meaningfully investigate this limit by regarding the human-translated versions of the TREC queries as the extreme high-quality limit of machine translation. In this task, monolingual retrieval (the usual baseline for judging the degree to which translation degrades retrieval performance in CLIR) can be regarded as the extreme high-quality limit of query trans-

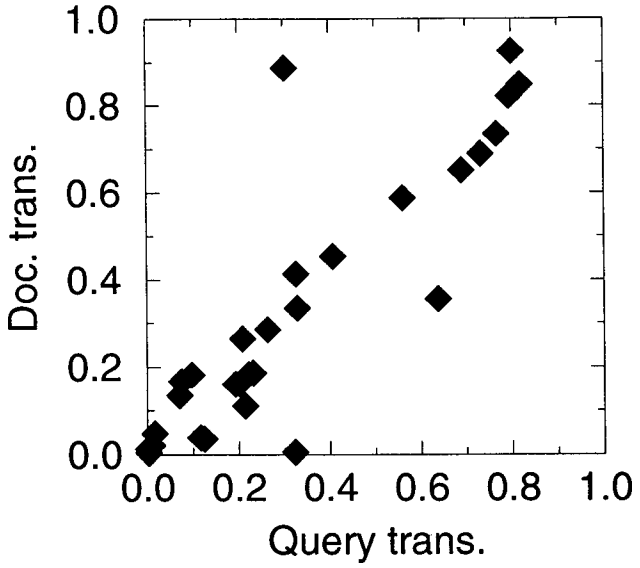

Figure 1: Scatterplot of average precision of document translation vs. query translation.

lation. Nevertheless, document translation provides another source of information, since the context sensitive aspects of the translation account for context in a manner distinct from current algorithms of information retrieval. Thus we do a further set of experiments in which we mix document translation and monolingual retrieval. Surprisingly, we find that the hybrid system outperforms the pure monolingual system. (See columns *ht* and *ht* + *dt* of Tables 1 and 2.) Thus we conclude that a mixture of document translation and query translation can be expected to outperform pure query translation, even very high quality query translation.

5 Conclusions and Future Work

We have performed experiments to compare query and document translation-based CLIR systems using statistical translation models that are trained identically for both translation directions. Our study is the largest comparative study of document translation and query translation of which we are aware; furthermore we have contrasted query and document translation systems on both directions within a language pair. We find no clear advantage for either the query translation system or the document translation system; instead *French*⇒*English* translation appears advantageous over *English*⇒*French* translation, in spite of identical procedures used in constructing both. However a hybrid system incorporating both directions of translation outperforms either. Furthermore, by incorporating human query translations rather than machine translations, we show that the hybrid system continues to outperform query translation. We have based our conclusions by comparing TREC-style average precisions of retrieval with a two-pass IR system; the same conclusions follow if we instead compare precisions at rank 20 or average precisions from first pass (Okapi) scores. Thus we conclude that even in the limit of extremely high quality query translation, it will remain advantageous to incorporate both document and query translation into a CLIR system. Future work will involve investigating translation direction differences in retrieval performance for other language pairs, and for statistical translation systems trained from comparable, rather than parallel corpora.

6 Acknowledgments

This work is supported by NIST grant no. 70NANB5H1174. We thank Scott Axelrod, Martin Franz, Salim Roukos, and Todd Ward for valuable discussions.

References

L. Ballasteros and W.B. Croft. 1997. Phrasal translation and query expansion techniques for cross-language information retrieval. In *20th Annual ACM SIGIR Conference on Information Retrieval.*

L. Ballasteros and W.B. Croft. 1998. Resolving ambiguity for cross-language retrieval. In *21th Annual ACM SIGIR Conference on Information Retrieval.*

P.F. Brown, J.C. Lai, and R.L. Mercer. 1991. Aligning sentences in parallel corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics.*

P. Brown, S. Della Pietra, V. Della Pietra, and R. Mercer. 1993. The mathematics of statistical machine translation : Parameter estimation. *Computational Linguistics*, 19:263–311.

C. Buckley, M. Mitra, J. Wals, and C. Cardie. 1998. Using clustering and superconcepts within SMART : TREC-6. In E.M. Voorhees and D.K. Harman, editors, *The 6th Text REtrieval Conference (TREC-6).*

J.G. Carbonell, Y. Yang, R.E. Frederking, R.D. Brown, Yibing Geng, and Danny Lee. 1997. Translingual information retrieval : A comparative evaluation. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence.*

E. Chan, S. Garcia, and S. Roukos. 1997. TREC-5 ad-hoc retrieval using k nearest-neighbors re-scoring. In E.M. Voorhees and D.K. Harman, editors, *The 5th Text REtrieval Conference (TREC-5).*

A. Diekema, F. Oroumchian, P. Sheridan, and E. Liddy. 1999. TREC-7 evaluation of Conceptual Interlingua Document Retrieval (CINDOR) in English and French. In E.M. Voorhees and D.K. Harman, editors, *The 7th Text REtrieval Conference (TREC-7).*

S. Dumais, T.A. Letsche, M.L. Littman, and T.K. Landauer. 1997. Automatic cross-language retrieval using latent semantic indexing. In *AAAI Symposium on Cross-Language Text and Speech Retrieval.*

M. Franz and S. Roukos. 1998. TREC-6 ad-hoc retrieval. In E.M. Voorhees and D.K. Harman, editors, *The 6th Text REtrieval Conference (TREC-6).*

M. Franz, J.S. McCarley, and S. Roukos. 1999. Ad hoc and multilingual information retrieval at IBM. In E.M. Voorhees and D.K. Harman, editors, *The 7th Text REtrieval Conference (TREC-7).*

J.S. McCarley and S. Roukos. 1998. Fast document translation for cross-language information retrieval. In D. Farwell., E. Hovy, and L. Gerber, editors, *Machine Translation and the Information Soup*, page 150.

D.W. Oard and P. Hackett. 1998. Document translation for cross-language text retrieval at the University of Maryland. In E.M. Voorhees and D.K. Harman, editors, *The 6th Text REtrieval Conference (TREC-6).*

D.W. Oard. 1998. A comparative study of query and document translation for cross-language information retrieval. In D. Farwell., E. Hovy, and L. Gerber, editors, *Machine Translation and the Information Soup*, page 472.

S.E. Robertson, S. Walker, S. Jones, M.M. Hancock-Beaulieu, and M. Gatford. 1995. Okapi at TREC-3. In E.M. Voorhees and D.K. Harman, editors, *The 3d Text REtrieval Conference (TREC-3).*

Jinxi Xu and W. Bruce Croft. 1996. Query expansion using local and global document analysis. In *19th Annual ACM SIGIR Conference on Information Retrieval.*