NTCIR CLIR Experiments at the University of Maryland

Douglas W. Oard and Jianqiang Wang
Digital Library Research Group
College of Library and Information Services
University of Maryland, College Park, MD 20742
{oard,wangjq}@glue.umd.edu

Abstract

This paper presents results for the Japanese/English cross-language information retrieval task on the NACSIS Test Collection. Two automatic dictionary-based query translation techniques were tried with four variants of the queries. The results indicate that longer queries outperform the required description-only queries and that use of the first translation in the edict dictionary is comparable with the use of every translation. Japanese term segmentation posed no unusual problems, which contrasts sharply with results previously obtained for cross-language retrieval between Chinese and English.

1 Introduction

Cross-language information retrieval (CLIR) deals with the problem of retrieving information in languages different from that of the query [8]. Several effective CLIR approaches are now known, but none have yet been tested on large-scale collections that include Asian languages. Several Asian languages lack explicit word boundaries in their written form, and this poses a challenge for CLIR systems about which little is presently understood. We recently ran an experiment using Chinese queries to retrieve English documents from the Text REtrieval Conference (TREC) in order to begin to address this issue [9]. In that work we found that segmentation errors produced a cascading effect through translation that ultimately produced inappropriate term weights, thus depressing retrieval effectiveness. In the NACSIS Test Collection Information Retrieval (NTCIR) experiments reported in this paper we applied the same experiment design to Japanese/English retrieval to explore whether the problem is present to the same degree in this case.

2 Background

There are four fundamental ways to match queries in one language with documents in another:

- **Cross-language matching.** Leave the queries and the documents untranslated and embed translation knowledge in the matching algorithm (e.g., [3]).
- **Query translation.** Translate the query into the documents' language(s) and then perform monolingual retrieval (e.g. [1]).
- **Document translation.** Translate the documents into the supported query language(s) and then perform monolingual retrieval (e.g., [7]).
- **Interlingual matching.** Translate both the queries and the documents into a language-neutral representation and use those representations as a basis for retrieval (e.g., [5]).

In cross-language retrieval between European languages, query translation has proven to be popular because it is efficient (for relatively short queries), and because the common character set sometimes results in helpful cross-language exact string matches when no translation is known for a word (as is commonly the case with proper names, for example). Dictionary-based query translation (term-by-term translation using a term list built from a bilingual dictionary) is easily implemented, and is well known to produce about half the retrieval effectiveness (e.g., average precision) of monolingual systems. Since our primary goal is to understand the additional challenges posed by Asian languages, we elected to use dictionary-based query translation (referred to below as DQT) for our experiments

Figure 1 illustrates the key differences between cross-language retrieval using DQT and the monolingual case. Queries enter from the left, and in what are called "bag-of-words" retrieval systems (i.e., those

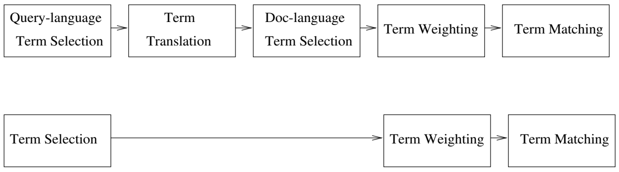

Figure 1: Comparison between cross-language (top) and monolingual (bottom) retrieval

that do not preserve word order information), the first step in both cases is to select terms. In European languages this can involve tokenization on white space, phrase recognition, and (for languages such as German) compound splitting. For Asian languages, the corresponding step is segmentation.

Although both cross-language and monolingual bag-of-words retrieval systems perform term selection, the intended use of the selected terms differs. In monolingual systems, the selected terms will be used directly for matching. The so-called "ranked retrieval" systems that we use seek to place documents that best match the query closest to the top of a ranked list. For this reason, query terms that are highly selective (i.e., that appear in only a few documents) typically receive greater weight.[1] The term matching stage, where weighted query terms are matched with the terms found in the documents, is then used to identify documents that best match the query.

In cross-language retrieval using DQT, two term selection stages are needed. The goal of the first is to discover terms *for which translations are known*, while the goal of the second is to select the best translation(s) from among those that are known to be possible. Some dictionaries present the most common translation (in general usage) first, and in that case a useful heuristic is to choose the first translation (DQT-FT). In other cases, a more conservative heuristic in which every translation is retained for each term (DQT-ET) has proven to be useful. Since detailed information about the development of a particular dictionary can be difficult to obtain, we routinely compare the two term choice strategies when running DQT experiments.

Term weighting serves the same purpose in cross-language retrieval—to give more emphasis to the most useful terms. In experiments with automatically segmented Chinese queries, we discovered that assigning term weights based on the selectivity of a *translated* term caused problems because segmentation errors typically produced terms for which many translations were known, and some of those translations were rare (and hence highly selective) English words [9]. Emphasizing selective terms is helpful when weighting query terms that are provided directly by the user, but our results with Chinese clearly indicate that it can sometimes be dangerous to apply it in the same way to translated terms.

3 Experiment Design

Each of our four query sets was formed by automatically extracting one or more fields from the given topics. The query set was then passed to JUMAN version 2.2 for segmentation[2]. The first column of the output (the component words) was then extracted and passed to Dictionary-based Query Translation (DQT). The DQT code requires a query set and a bilingual dictionary as input and produces, a query set with the translations of each query word into target language as output. We used the freely available "edict' Japanese/English dictionary, which contains 64,433 Japanese terms and a total of 104,705 bilingual term pairs.[3] Some preprocessing was done, including removal of pronunciation information and (after our official submission), and removal of parenthetical clauses (which are generally explanations rather than translations). Our existing DQT code had to be modified to accommodate multibyte characters—we did this by converting the Japanese characters (in both in the dictionary and the query set) into the corresponding hexadecimal representations.

The translated queries were passed to version 3.1p1 of the Inquery information retrieval system, which we obtained from the University of Massachusetts [4]. Inquery is a probabilistic retrieval system based on Bayesian inference networks. In our experiment, we used "sum" operator to form queries. The sum operator calculates the value of the belief that a query is satisfied by a document as the mean of the beliefs associated each query term. The Inquery "kstem"

[1] This measure of selectivity is generally referred to as the "inverse document frequency" (IDF) of a term. For reasons of efficiency, it is more common to associate IDF weights with every occurrence of a term in a document because the value can be computed in advance. Associating IDF weights with the query as we have done here sheds light on the interaction between query translation and IDF weights without altering the retrieval outcome.

[2] We happened to have an installed copy of JUMAN 2.2 available, and our inability to read the Japanese documentation for JUMAN prevented us from installing a more recent version in time for these experiments. JUMAN 3.61 is available at http://pine.kuee.kyoto-u.ac.jp/nl-resource/juman.html

[3] The edict dictionary is freely available in electronic form from Monash University.

stemmer and the standard English Inquery stopword list were used when indexing the English document collection.

4 Results

After submitting the two official runs, we discovered that we had inadvertently omitted 10 of the scored topics from the run in which we used NARRATIVE field to form the queries (umd2). We have corrected this mistake in the experiments reported here. We also performed the dictionary cleanup described above between our official results and the ones reported here. In all, we made eight runs for this paper:

- **DFT** Queries formed with the DESCRIPTION field and translated with DQT-FT (submitted officially as umd1).
- **DET** Queries formed with the DESCRIPTION field and translated with DQT-ET.
- **JFT** Queries formed with the J.CONCEPT field and translated with DQT-FT.
- **JET** Queries formed with the J.CONCEPT field and translated with DQT-ET.
- **NFT** Queries formed with the NARRATIVE field and translated with DQT-FT (submitted officially as umd2).
- **NET** Queries formed with the NARRATIVE field and translated with DQT-ET.
- **TNJDFT** Queries formed with the TITLE, NARRATIVE, J.CONCEPT and DESCRIPTION fields and translated with DQT-FT.
- **TNJDET** Queries formed with the TITLE, NARRATIVE, J.CONCEPT and DESCRIPTION fields and translated with DQT-ET.

Non-interpolated average precision values for these eight runs are shown in Table 1, and Figures 2 and 3 show the 11 point recall-precision graphs for DQT-FT and DQT-ET respectively. By these measures, we achieved the best overall retrieval effectiveness by using DQT-FT with queries formed from all four topic fields. The insignificant change in DFT between our official submission and these results (from 0.0788 to 0.0791) is due solely to dictionary cleanup. The inclusion of the previously omitted queries is thus the obvious explanation for the dramatic increase in NFT between our official submission and these results (from 0.0968 to 0.1204).

	Topic Fields			
DQT	D	J	N	TNJD
ET	0.0704	0.0981	0.0996	0.1337
FT	0.0791	0.1056	0.1204	0.1534

Table 1: Non-interpolated average precision (D=DESCRIPTION, J=J.CONCEPT, N=NARRATIVE, T=TITLE).

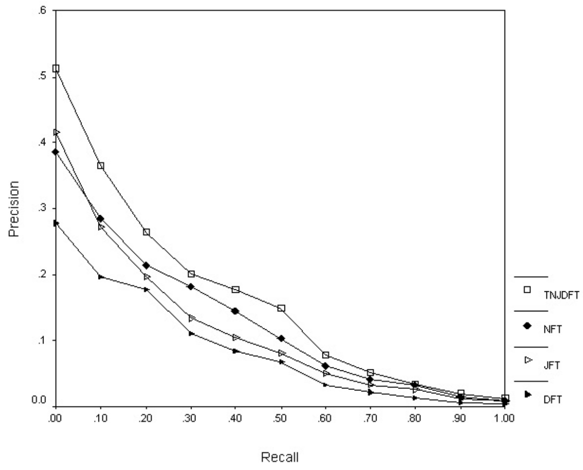

Figure 2: Precision-recall curves with DQT-FT.

We tested our results for statistical significance using paired sample t-tests. The significance values for pairwise comparisons between topic sets when DQT-FT was used are shown in Table 2. Values below 0.05 are generally accepted as significant is studies of this type [6]. In this test, the 39 queries are taken as random samples form a query population, the 11-point average precision for each query is the dependent variable, and the DQT technique and the query set are the independent variables. We found that long queries often outperform short queries. For example, queries formed with all four fields (TNJDFT and TNJDET) perform significantly better than all the other six sets of queries. Queries with NARRATIVE field also significantly outperform the required queries that used only the DESCRIPTION field. However, we didn't observe statistically significant differences (at the 0.05 level) between queries with DESCRIPTION or NARRATIVE fields and queries with J.CONCEPT field.

As Figure 4 illustrates, the results for DQT-FT and DQT-ET were quite similar when averaged over all queries. Statistical significance tests failed to detect

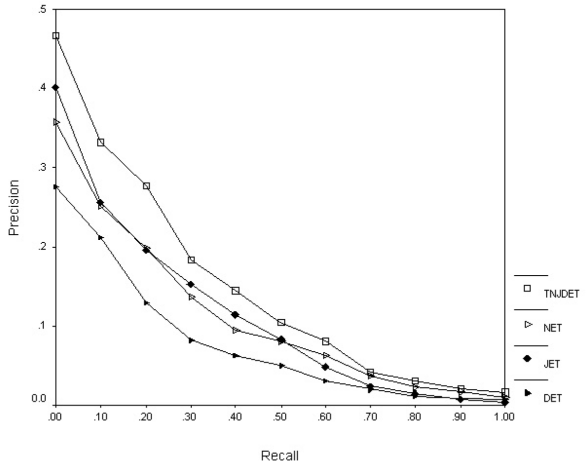

Figure 3: Precision-recall curves with DQT-ET.

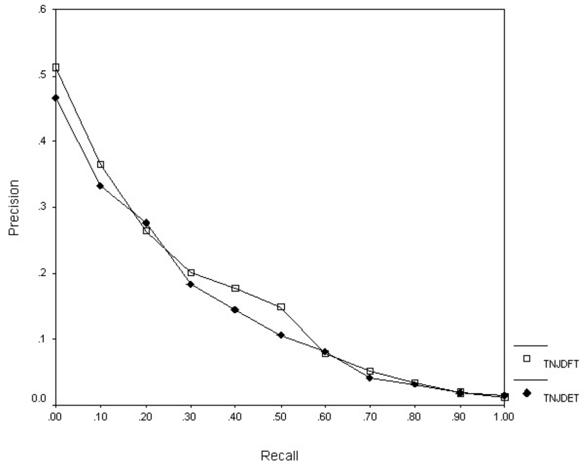

Figure 4: DQT-FT vs. DQT-ET.

Query	D	J	N
J	0.486		
N	0.032	0.278	
TNJD	0.002	0.007	0.012

Table 2: Paired sample t-test significance values for DQT-FT.

a significant difference between DQT-FT and DQT-ET for any of the four query sets. The query-by-query comparison in Figure 5 provides some additional insight, showing that DQT-FT noticeably outperformed DQT-ET on some queries, but noticeably underperformed it on others.

We explored the interaction between segmentation and translation by examining some of the original, segmented and translated queries. Although Japanese in written form is similar in some ways to Chinese, it does have unique characteristics. Chinese texts are mainly composed of hanzi characters, while Japanese texts are composed of kanji, hiragana, and katakana. A character set change provides a reliable cue for term segmentation, so segmentation is inherently easier for Japanese than for Chinese. Furthermore, hiragana, which is common in the queries we examined, often represents function words that are of little use with bag-of-words retrieval techniques. There are few English translations for hiragana in the edict dictionary. So even if the segmenter missegments hiragana, edict would be unlikely to propagate the error through translation. Together, these factors might explain why the cascading errors observed in Chinese were not present in these experiments.

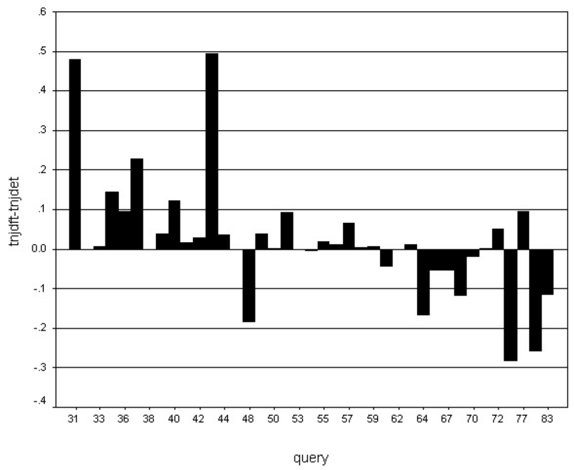

Figure 5: Query-by-query comparison of DQT-FT and DQT-ET.

5 Conclusion

We have tested Japanese/English cross-language information retrieval with automatic dictionary-based query translation. The results reveal that long queries often outperform shorter ones, but that our two query translation techniques perform comparably. Japanese term segmentation does not appear to pose problems that are as severe as those that we have encountered with CLIR between Chinese and English. The existence of multiple character types in Japanese seems to be the fundamental reason for this. In future work we plan to explore additional cross-language retrieval techniques in the context of Asian languages, perhaps including the application of word sense disambiguation approaches such as those studied by Ballesteros and Croft [2].

This first NTCIR evaluation has provided us with valuable experience that has helped us to deepen our understanding of critical issues for cross-language information retrieval using Asian languages. We expect that the test collection will prove to be a valuable legacy, now permitting a broader range of experiments than has previously been possible.

Acknowledgments

This work has been supported in part by DARPA contract N6600197C8540. The authors wish to express their appreciation to Noriko Kando for extending an invitation to join the evaluation and her assistance along the way.

References

[1] Lisa Ballesteros and W. Bruce Croft. 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*, July 1997.

[2] Lisa Ballesteros and W. Bruce Croft. Resolving ambiguity for cross-language retrieval. In C.J. Van Rijsbergen W. Bruce Croft, Alistair Moffat, editor, *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 64–71. ACM Press, August 1998.

[3] Chris Buckley, Mandar Mitra, Janet Walz, and Claire Cardie. Using clustering and SuperConcepts within SMART: TREC 6. In *The Sixth Text REtrieval Conference (TREC-6)*. National Institutes of Standards and Technology, November 1998. http://trec.nist.gov/.

[4] James P. Callan, W. Bruce Croft, and Stephen M. Harding. The INQUERY retrieval system. In *Proceedings of the Third International Conference on Database and Expert Systems Applications*, pages 78–83. Springer-Verlag, 1992.

[5] Susan T. Dumais, Thomas K. Landauer, and Michael L. Littman. Automatic cross-linguistic information retrieval using latent semantic indexing. In *Cross-Language Information Retrieval*. Kluwer Academic, 1998.

[6] David Hull. Using statistical testing in the evaluation of retrieval experiments. In *Proceedings of the 16th ACM/SIGIR Conference on Research and Development in Information Retrieval*, pages 329–338, 1993.

[7] Douglas W. Oard. A comparative study of query and document translation for cross-language information retrieval. In *Proceedings of the Third Conference of the Association for Machine Translation in the Americas*, October 1998.

[8] Douglas W. Oard and Anne Diekema. Cross-language information retrieval. In *Annual Review of Information Science and Technology*, volume 33. American Society for Information Science, 1998.

[9] Douglas W. Oard and Jianqiang Wang. Term segmentation in Chinese/English cross-language information retrieval. In *Proceedings of the Symposium on String Processing and Information Retrieval*, September 1999. to appear. http://www.glue.umd.edu/~oard/research.html.