A Comparison of Query Translation Methods for English-Japanese Cross-Language Information Retrieval

Gareth Jones*, Tetsuya Sakai, Nigel Collier†, Akira Kumano and Kazuo Sumita
Human Interface Laboratory
Research and Development Center, Toshiba Corporation
1 Komukai Toshiba-cho, Saiwai-ku, Kawasaki 210-8582, Japan
email:{Tetsuya.Sakai,Akira.Kumano,Kazuo.Sumita}@toshiba.co.jp

Abstract

In this paper we report results of an investigation into English-Japanese Cross-Language Information Retrieval (CLIR) comparing a number of query translation methods. Results from experiments using the standard BMIR-J2 Japanese collection suggest that full machine translation (MT) can outperform popular dictionary-based query translation methods and further that in this context MT is largely robust to queries with little linguistic structure.

1 Introduction

Interest in Cross-Language Information Retrieval (CLIR) has grown rapidly in recent years [1] [2] [3]. While most existing studies have concentrated on CLIR between English and one or more European languages, there is a need to develop methods for CLIR between European and Asian languages. In this paper we report results of an experimental investigation into English-Japanese CLIR. In particular we concentrate on the comparison of various query translation methods.

Three principal methods have been proposed and explored for query translation: *dictionary-based* [1] [2], *parallel-corpora* [3] and *full machine translation*. The simplest method, dictionary term lookup (DTL), involves looking up each query term in a bilingual dictionary and replacing it with all possible term translations, while full machine translation (FMT) uses all available linguistic resources to calculate a single best possible translation of the whole query. At a linguistic level these two approaches can be regarded as extremes of translation complexity. It has been argued in the literature that the shortness and lack of linguistic structure in typical search queries means that FMT is unsuitable and that dictionary methods should be favoured. In this investigation we seek to test this hypothesis. While not discounting the potential utility of parallel corpora methods where suitable resources are available, we do not consider them further in our current investigation.

In this paper we describe English-Japanese CLIR experiments using the standard BMIR-J2 Japanese text collection [4]. Our results suggest that FMT can perform substantially better than DTL methods and is generally robust to a lack of linguistic structure in queries.

*Current address: Department of Computer Science, University of Exeter, U.K. email: G.J.F.Jones@exeter.ac.uk

†Current address: Department of Information Science, University of Tokyo, Japan email: nigel@is.s.u-tokyo.ac.jp

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
SIGIR '99 8/99 Berkley, CA, USA
© 1999 ACM 1-58113-096-1/99/0007...$5.00

2 Translation Methods

Five translation methods were investigated in this study:

- *ALL*: ALL corresponds to simple DTL. Each English word in the query is replaced by all possible translations that appear in the bilingual dictionary.
- *POS*: Each English word is tagged with its part-of-speech (POS). It is then replaced by translations from a bilingual dictionary which correspond to this POS. If there is no POS match for a word then the ALL method is used.
- *DEF*: The English word is replaced by a single default translation from the bilingual dictionary.
- *SNS*: The SNS methods performs linguistic analysis to disambiguate an English word and then outputs a list of all possible synonyms. This is similar to FMT, but no final choice is made of the single most likely translation of a word.
- *FMT*: Full Machine Translation (FMT) performs linguistic analysis to disambiguate an English word and also synonym selection to choose the word in Japanese that is most appropriate to the style and context of the request.

FMT query translation is carried out using the Toshiba ASTRANSAC system. ASTRANSAC uses the *transfer translation method* following a standard approach of morphological analysis, syntactic analysis, semantic analysis and selection of translation words. Analysis is predominantly top-down and uses ATNs (Augmented Transition Networks) on a context-free grammar. In our simulation we used a 65,000 term common word bilingual dictionary and a 14,000 term proper noun dictionary which we considered relevant to the new events covered in the document collection used in this experiment. The other translation methods used relevant components from ASTRANSAC and its bilingual dictionary.

Translation			*ALL*		*POS*		*DEF*		*SNS*		*FMT*	
Translator		Mono	*T1*	*T2*	*T1*	*T2*	*T1*	*T2*	*T1*	*T2*	*T1*	*T2*
Prec.	5 docs	0.576	0.204	0.196	0.268	0.232	0.248	0.244	0.264	0.244	0.364	0.312
	10 docs	0.504	0.194	0.206	0.240	0.244	0.244	0.242	0.254	0.242	0.332	0.288
	15 docs	0.461	0.192	0.196	0.232	0.228	0.244	0.240	0.257	0.241	0.316	0.281
	20 docs	0.426	0.183	0.184	0.221	0.213	0.224	0.226	0.235	0.224	0.292	0.264
	Av Prec	0.449	0.160	0.155	0.211	0.193	0.207	0.203	0.215	0.201	0.284	0.242
	% change	—	-64.4	-65.5	-53.0	-57.0	-53.9	-54.8	-52.1	-55.2	-36.7	-46.1
LF	Av Prec	0.478	0.219	0.223	0.266	0.256	0.264	0.264	0.271	0.263	0.339	0.304
	% change	—	-54.2	-53.3	-44.4	-46.4	-44.8	-44.8	-43.3	-45.0	-29.1	-36.4
hack –	Av Prec	0.449	0.164	0.151	0.200	0.193	0.190	0.200	0.201	0.203	0.264	0.239
no LF	% change	—	-63.5	-66.4	-55.5	-57.0	-57.7	-55.5	-55.2	-54.8	-41.2	-46.8

cw: $K1 = 0.4$, $b = 0.5$

Table 1: Precision values for CLIR experiments with BMIR-J2.

3 Information Retrieval

Our retrieval experiments use the Toshiba NEAT Japanese Information Retrieval System [5]. Documents are indexed using both morphological segmentation and character-based analysis. In response to a search request a list of articles ranked by request-article matching score is returned.

The NEAT System uses the BM25 probabilistic combined weight (*cw*), shown to be effective for Japanese language retrieval in [5]. The BM25 constants $K1$ and b are selected empirically for the BMIR-J2 collection.

4 BMIR-J2 Japanese Retrieval Test Collection

The BMIR-J2 collection consists of 5080 articles taken from the Mainichi Newspapers in the fields of economics and engineering, and a total of 50 natural language search requests. The average number of relevant documents for each query is 33.6. BMIR-J2 was designed so that some search requests can be satisfied very easily, while for some others it is very difficult to retrieve the relevant documents using the request. BMIR-J2 is described in detail in [4].

4.1 Cross-Language Retrieval Extensions

A CLIR BMIR-J2 collection was constructed by manually translating the Japanese BMIR-J2 requests into English. The effectiveness of the various query translation methods for CLIR was then investigated.

An underlying assumption in this approach is that the initial manual translation is accurate, and that it can be unambiguously translated back to the original Japanese query. To investigate the possible effects of alternative translation, the query set was independently translated by two bilingual Japanese and English speakers (*T1* & *T2*).

Two further query sets ("hack") were formed by editing the English requests to remove linguistic structure and hence make them more closely resemble typical user search requests. This process mainly involved removing function words from the natural language sentences.

5 Retrieval Experiments

Table 1 shows the results of our retrieval experiments. All results show precision at ranked cutoff of 5, 10, 15 and 20 documents, and standard TREC average precision. Clearly for such a small collection the specific figures are neither reliable nor significant, reported results should thus be regarded only as indicative.

The results show that for both of our translators FMT outperforms the other translation methods. Simple ALL translation performs worst, but there is little difference between the other approaches. Retrieval performance in all cases is improved by application of Local Feedback (LF) [2] using Okapi style reweighting and expansion [5].

Perhaps surprisingly, this performance pattern is repeated for the "hack" queries. Analysis of the *FMT* queries showed that translation of 2/3 of the translated "hack" queries were identical to the natural language queries, and that for most of the other cases only small changes were observed. Lack of space prevents us describing this behaviour in more detail here.

6 Conclusions and Further Work

Results reported in this paper indicate that English-Japanese CLIR can be performed successfully, and that query translation via *FMT* is potentialy robust in CLIR applications, challenging the frequent assumption that DTL is the better option. Our further work will focus on detailed investigation of translation differences and exploring other techniques such as exploiting co-occurrence information [2] to further improve performance.

References

[1] D. A. Hull and G. Grefenstette. Querying across languages: A dictionary-based approach to multilingual information retrieval. In *Proceedings of ACM SIGIR 96*, pages 49–57, Zurich, 1996. ACM.

[2] L. Ballesteros and W. B. Croft. Resolving Ambiguity for Cross-Language Retrieval. In *Proceedings of the ACM SIGIR 98*, pages 64–71, Melbourne, 1998. ACM.

[3] J. Carbonell, Y. Yang, R. E. Frederking, R. D. Brown, Y. Geng, and D. Lee. Translingual Information Retrieval: A Comparative Evaluation. In *Proceedings of the 15th International Joint Conference of Artificial Intelligence*, pages 708–714, Nagoya, 1996. IJCAI.

[4] T. Kitani et al. Lessons from BMIR-J2: A Test Collection for Japanese IR Systems. In *Proceedings of ACM SIGIR 98*, pages 345–346, Melbourne, 1998. ACM.

[5] G. J. F. Jones, T. Sakai, M. Kajiura, and K. Sumita. Experiments in Japanese Text Retrieval and Routing using the NEAT System. In *Proceedings of ACM SIGIR 98*, pages 197–205, Melbourne, 1998. ACM.