Exploring the use of Machine Translation resources 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@toshiba.co.jp

Abstract

This paper describes a detailed investigation into the use of Machine Translation techniques for query translation in English-Japanese Cross-Language Information Retrieval (CLIR). Experimental results are reported for the standard BMIR-J2 Japanese text retrieval collection extended to a CLIR task. Results indicate that retrieval performance increases as greater linguistic processing is used in the translation process, and that local feedback via term reweighting and query expansion is effective for this task.

1 Introduction

The recent development in technologies for global access to online information sources has led to a rapid expansion in interest in Cross-Language Information Retrieval (CLIR) research. A fundamnetal question for CLIR is, what method should be used for text translation? This paper contributes to this ongoing debate by providing a detailed exploration into the use of Machine Translation (MT) techniques for query translation in English-Japanese CLIR.

Much of the previous work in CLIR such as [1] [2] has focused on CLIR for European language pairs and has avoided many of the challenges faced in processing European-Asian language pairs. In the latter case particular difficulties arise because the language pairs are not cognates. The three major practical challenges in CLIR are: *coverage*: providing sufficient bilingual knowledge; *disambiguation*: how to identify conceptually different forms from the set of possible translations of a query word; and *synonym selection*: how to identify conceptually equivalent forms of a translation.

Experimental results reported in this paper for CLIR retrieval using the standard BMIR-J2 Japanese information retrieval test collection [3] extended for CLIR experimentation suggest that using full machine translation can perform substantially better than attempting to use simple dictionary term lookup or individual components of a machine translation system.

The remainder of this paper is organised as follows: Section 2 provides some brief background to translation methods in CLIR research; Section 3 describes the MT system used in our work and Section 4 gives relevant details of the Information Retrieval methods used. Section 5 provides experimental results, and finally Section 6 provides conclusions to our current study and gives pointers to further work.

2 Translation Methods in Cross-Language Information Retrieval

This section gives a brief outline of translation techniques which have been applied to CLIR highlighting the advantages and disadvantages of each. Query translation methods for CLIR fall into three categories:

- **Dictionary Term Lookup**: Each query term is replaced by all its possible translations from a bilingual dictionary. The main disadvantage of DTL is the high degree of ambiguity often introduced into the translated query which significantly degrades retrieval performance [1].
- **Full Machine Translation**: All available linguistic resources are used to calculate a single best possible translation of the whole query. It has been argued in the literature that the shortness and lack of linguistic structure in typical search queries and domain dependence issues [1] mean that FMT is unsuitable for CLIR and that dictionary methods should be favoured.
- **Parallel Corpora Methods**: Two parallel (or usually more strictly comparable) document collections in the query and document languages are used to translate the query. These methods have shown some promise [2] [5]. However they rely on the availablility of suitable parallel document collection for the language pair. While not discounting the potential utility of parallel corpora methods, we do not consider them further in this paper.

Since the DTL amd MT approaches can be regarded as extremes of translation complexity, we investigate these extremes and the utility of using the output from the various of components of an MT system. The next section describes the MT system used in our investigation.

3 The Toshiba ASTRANSAC System

For our experiments query translation was performed using a modified version of the Toshiba ASTRANSAC English-Japanese MT system based on the *transfer method* [6]. Five translation methods were investigated in this study.

[*]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

- *ALL*: ALL corresponds to simple DTL. Of the five translation methods *ALL* has the lowest level of analysis.
- *POS*: Part-of-speech (POS) is used to tag each English word. The word is then replaced only by words from the bilingual dictionary with the same POS as the tag. If there is no match for the English word tag, it is replaced by all possible translations as in ALL.
- *DEF*: Each English word is replaced by a single default translation from the bilingual dictionary which is the most common translation.
- *SNS*: SNS 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 find the Japanese word which is most appropriate to the style and context of the request.

4 Information Retrieval

Our retrieval experiments use the Toshiba NEAT Japanese Information Filtering System with the BM25 probabilistic retrieval model [7] . The BM25 cw term weight is calculated as follows,

$$cw(i,j) = \frac{cfw(i) \times tf(i,j) \times (K1+1)}{K1 \times ((1-b) + (b \times ndl(j))) + tf(i,j)} \qquad (1)$$

where $cw(i,j)$ represents the weight of term i in document j, $cfw(i)$ is the standard collection weight (often referred to as inverse document frequency weight), $tf(i,j)$ is frequency of term i in document j, and $ndl(j) = dl(j)/\text{Av}\ dl$ for all docs is the normalised length of document j. $K1$ and b are empirically selected tuning constants for a particular collection.

In Local Feedback the top ranked retrieved documents are assumed to be relevant and standard relevance feedback methods applied [4]. Terms can be reweighted by replacing the the $cfw(i)$ component of Equation 1 by the Robertson/Sparck Jones relevance weight $rw(i)$ [8],

$$rw(i) = \log \frac{(r(i)+0.5)(N-n(i)-R+r(i)+0.5)}{(n(i)-r(i)+0.5)(R-r(i)+0.5)}$$

where $r(i)$ is the total number of relevant documents containing term i, $n(i)$ is the total number of documents containing i, R is the total number of relevant documents for this query, and N is the total number of documents. Query expansion terms can be selected by taking the top ranked terms from a list calculated using the standard Robertson selection value (rsv) $rsv(i) = r(i) \times rw(i)$ [8],

5 Experimental Investigation

5.1 Test Collection

Our experiments use the BMIR-J2 Japanese retrieval collection [3]. 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 main search requests. A CLIR test collection was constructed by manually translating the Japanese BMIR-J2 requests into English. 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 Translator 1 (*T1*) and Translator 2 (*T2*).

Local Feedback		no	yes
Prec.	5 docs	0.576	0.588
	10 docs	0.504	0.540
	15 docs	0.461	0.499
	20 docs	0.426	0.462
Av Precision		0.449	0.478

Table 1: Precision values for monolingual BMIR-J2.

Translation		*ALL*	*POS*	*DEF*	*SNS*	*FMT*
Prec.	5 docs	0.204	0.268	0.248	0.264	0.364
	10 docs	0.194	0.240	0.244	0.254	0.332
	15 docs	0.192	0.232	0.244	0.257	0.316
	20 docs	0.183	0.221	0.224	0.235	0.292
Av Precision		0.160	0.211	0.207	0.215	0.284
% change		-64.4	-53.0	-53.9	-52.1	-36.7

Table 2: Precision values for BMIR-J2 CLIR for T1 without local feedback.

Translation		*ALL*	*POS*	*DEF*	*SNS*	*FMT*
Prec.	5 docs	0.196	0.232	0.244	0.244	0.312
	10 docs	0.206	0.244	0.242	0.242	0.288
	15 docs	0.196	0.228	0.240	0.241	0.281
	20 docs	0.184	0.213	0.226	0.224	0.264
Av Precision		0.155	0.193	0.203	0.201	0.242
% change		-65.5	-57.0	-54.8	-55.2	-46.1

Table 3: Precision values for BMIR-J2 CLIR for T2 without local feedback.

5.2 Retrieval Experiments

This section presents our retrieval results. Results are shown using precision at ranked cutoff of 5, 10, 15 and 20 documents, and TREC average precision, and using standard Recall-Precision graphs. The empirical cw parameters were selected as $K1 = 0.4$ and $b = 0.5$.

Table 1 shows monolingual retrieval performance for BMIR-J2. Tables 2 and 3 show corresponding retrieval performance for the various translation schemes for T1 and T2 respectively without use of local feedback. Overall it can seen that retrieval performance is degraded for all CLIR methods relative to the monolingual baseline. The worst performance is observed for the ALL translation method, and the best for the FMT approach. The other methods making use of part of the FMT process produced similar retrieval peformance somewhere between the extremes set by the ALL and FMT approaches.

Looking across the tables it can be seen that the performance trends are similar for our two translators, giving some verification to previous studies using only a single query translation, but suggesting that small variations in performance for different retrieval methods be verified by repeating experiments with another independent translation.

Figures 1 and 2 show equivalent Recall-Precision curves T1 excluding and including local feedback respectively[1] . It can be seen that again results for FMT are significantly better than the other methods, and results for ALL the worst. Application of local feedback improves all results, but makes no difference to the relative performance of the different translation methods. Figure 3 shows the percentage gain of FMT over ALL in terms of initial average precision values on a query-by-query basis. This reveals that the average performance is largely affected by outliers [2] .

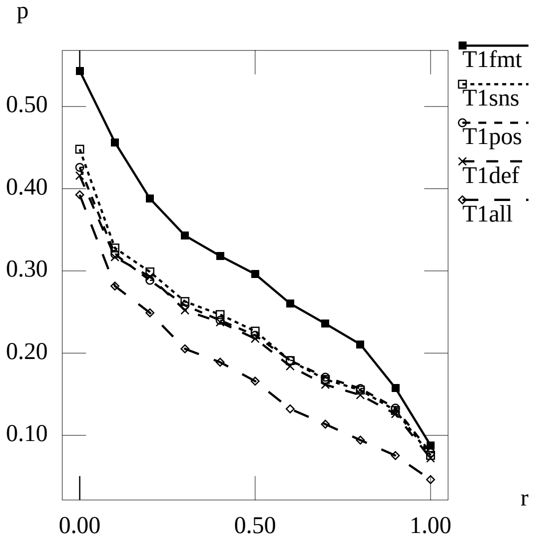

Figure 1: R-P curves for T1 without LF

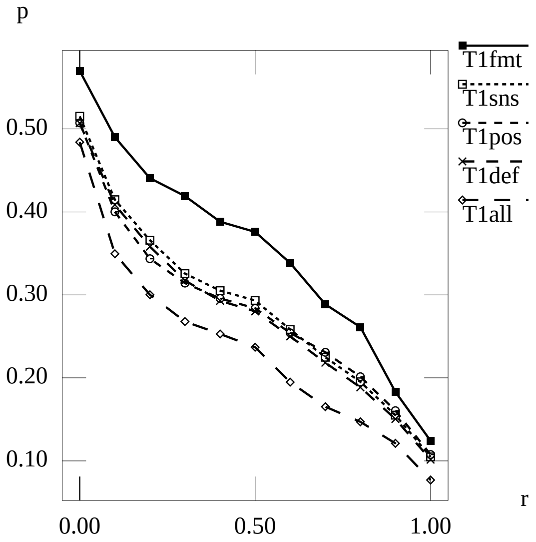

Figure 2: R-P Curves for T1 using LF

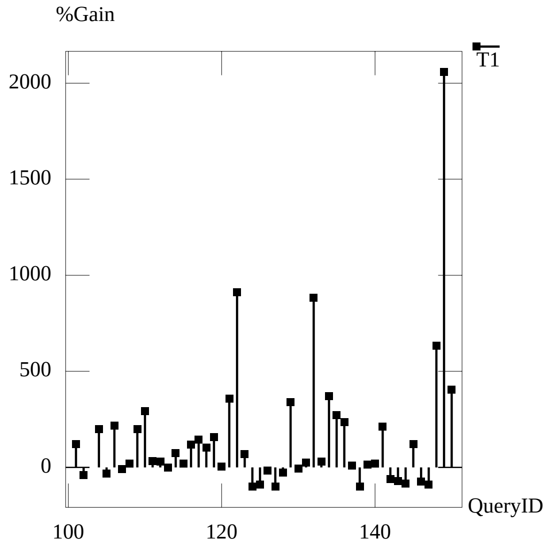

Figure 3: Query-by-query comparision of FMT vs ALL for T1 without LF

6 Conclusions and Further Work

This paper has described an investigation into the use of MT technologies for English-Japanese CLIR. The experimental results presented here indicate that using all available translation resources can produce superior CLIR performance to popular DTL based methods, at least for non-congnate language pairs. Further research with larger collections and alternative language pairs is needed to verify the results of our experiments on BMIR-J2 collection.

References

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

[2] P. Sheridan and J. P. Ballerini. Experiments in Multilingual Information Retrieval using the SPIDER system. In *Proceedings of ACM SIGIR 96*, pages 58–65, Zurich, 1996. ACM.

[3] 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.

[4] L. Ballesteros and W. B. Croft. Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval. In *Proceedings of ACM SIGIR*, pages 84–91, Philadelphia, 1997. ACM.

[5] J. Carbonell, Y. Yang, R. E. Frederking, R. D. Brown, Y. Geng, and D. Lee. Translingual Information Retrieval: A Comparative Evaluation. In *Proceedings of IJCAI 97*, pages 708–714, Nagoya, 1996.

[6] H. Hirakawa, H. Nogami, and S. Amano. EJ/JE machine translation system ASTRANSAC - extensions towards personalization. In *Proceedings of the Machine Translation Summit III*, pages 73–80, 1991.

[7] 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*, pages 197–205, Melbourne, 1998. ACM.

[8] S. E. Robertson. On term selection for query expansion. *Journal of Documentation*, 46:359–364, 1990.

[1] The final paper will include results for T2 as well.

[2] We will have a closer look at the outliers in the final paper.