NTCIR-3 Cross-Language IR Experiments at ULIS

Atsushi Fujii[†,††] Tetsuya Ishikawa[†]
[†] University of Library and Information Science
1-2 Kasuga, Tsukuba, 305-8550, Japan
[††] CREST, Japan Science and Technology Corporation
fujii@ulis.ac.jp

Abstract

This paper describes our retrieval system for the NTCIR-3 CLIR task, focusing on Japanese and English. We integrate query and document translation methods to improve retrieval accuracy, and perform clustering to improve browsing efficiency. In query translation, to derive possible translations for user queries, we use dictionaries and perform a transliteration method, which generates translations for out-of-dictionary loanwords based on the Japanese phonogram system. We also use a probabilistic model to resolve translation/transliteration ambiguity to improve the query translation accuracy. We show the effectiveness of our system with respect to the NTCIR-3 J-J, E-E, J-E, and J-JE subtasks.

Keywords: *Cross-language information retrieval, Query Translation, Transliteration, NTCIR-3*

1 Introduction

We participated in the NTCIR-3 cross-language information retrieval (CLIR) task, where Japanese, English, Korean, and Chinese topics (queries) and newspaper articles were used to evaluate the performance of participating IR systems. The NTCIR-3 CLIR task is subdivided into the following three categories:

- Multilingual CLIR (MLIR), in which queries in one of the four languages (i.e., J, E, K, and C) are submitted to search collections in more than one language for documents relevant to user information need,
- Bilingual CLIR (BLIR), which resembles the MLIR task, but a target collection consists of a single language,
- Single Language IR (SLIR), which is a monolingual IR task for Japanese, Korean, and Chinese.

Among the above various subtasks, we focused on Japanese and English, and performed the J-JE MLIR, J-E BLIR, and J-J/E-E SLIR tasks. This paper describes our CLIR system and reports its performance with respect to the NTCIR-3 CLIR test collection.

It should be noted that the BLIR task here is usually termed "cross-language/lingual information retrieval (CLIR)" in past literature. Thus, hereafter, we shall interchangeably use the terms "CLIR" and "BLIR".

Section 2 explains methodological background for CLIR. Section 3 describes the overall design of our system. Section 4 explains comparative experiments using the NTCIR-3 CLIR collection.

2 Background

Since our research and development must be contextualized in terms of past research literature, we discuss existing methods for CLIR. However, in this section we focus mainly on CLIR, because a) CLIR and MLIR share essential issues to be considered, and b) the number of recent CLIR methods is greater than those for MLIR.

Since by definition queries and documents are in different languages, queries and documents need to be standardized into a common representation so that monolingual retrieval techniques can be used. From this point of view, existing CLIR methods can be classified into the following three fundamental categories.

The first method translates queries into the document language [1, 8, 12]. The second method translates documents into the query language [11, 13]. The third method projects both queries and documents into a language-independent space by way of thesaurus classes [9, 16] and latent semantic indexing [2, 10]. We shall call those methods, "query translation", "document translation" and "interlingual representation" methods, respectively.

A number of integrated methods have also been proposed for CLIR.

McCarley [11] showed that a hybrid system, where the relevance degree of each document (i.e., the score) is the mean of those obtained with query and document translation methods, outperformed systems based on either query or document translation methods.

Fujii and Ishikawa [6] proposed a *two-stage* method, which also integrates the query and document translation methods. In the first stage, the query translation method is used to retrieve a limited number of foreign documents. Then, in the second stage, retrieved documents are machine translated into the user language. Thus, the computational cost required for the MT-based document translation can be minimized. Finally, those documents are re-ranked based on the score, combining those individually obtained with the first and second stages. Preliminary experiments using the NTCIR-1 and 2 Japanese/English CLIR collections showed that the two-stage method outperformed the query translation method.

3 System Description

3.1 Overview

Figure 1 depicts the overall design of our system, which consists of an engine and two postprocessing modules, that is, re-ranking and clustering modules. The engine retrieves documents in response to user queries, and outputs those documents in the source (user) language. We used the same system in the NTCIR-2 IR task [7].

In the case of CLIR, queries are first translated into the document language, to search a monolingual collection (in either of Japanese or English) for relevant documents. Then, retrieved documents are translated into the user language.

In the case of MLIR, both the source and translated queries are used to search a multi-lingual collection (in both Japanese and English) for relevant documents. Then, only retrieved documents that are not in the user language are translated into the user language.

In principle, we need only the engine to realize CLIR and MLIR in the sense that users can retrieve/browse foreign documents through their native language. However, to improve the quality of our system, two alternative postprocessing modules can optionally be used.

Following the two-stage method we previously proposed [6], the re-ranking module re-ranks documents retrieved by the engine to improve the retrieval accuracy. In this case, the engine and re-ranking module correspond to the first and second stages, respectively. Alternatively, the clustering module divides retrieved documents into a certain number of groups, so as to improve the browsing efficiency.

However, in the NTCIR-3 tasks, we did not evaluate the performance of the two postprocessing modules and do not further explain these modules.

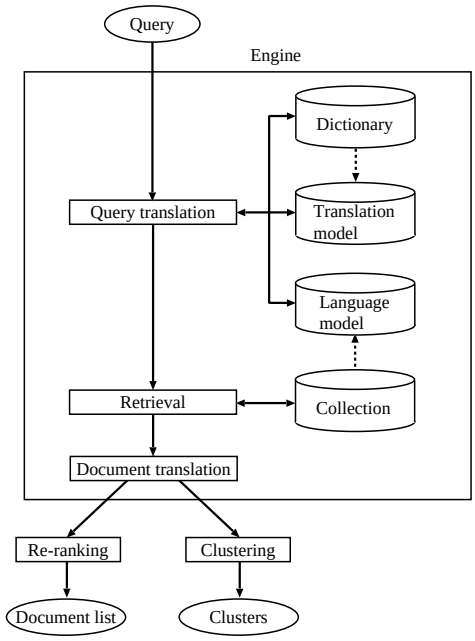

Figure 1. The design of our CLIR system (solid and dashed arrows denote on-line and off-line processes, respectively).

3.2 Engine

The engine consists of query translation, retrieval and document translation modules.

The query translation module is based on our previous method [5, 8], which was also used in the first two NTCIR Japanese/English CLIR tasks [4, 7]. We use the Nova dictionary[1] to derive possible word/phrase translations, and resolve translation ambiguity using a probabilistic method. The Nova dictionary includes approximately one million Japanese-English translations related to 19 technical fields as listed below:

> aeronautics, biotechnology, business, chemistry, computers, construction, defense, ecology, electricity, energy, finance, law, mathematics, mechanics, medicine, metals, oceanography, plants, trade.

In addition, for words unlisted in the dictionary, transliteration is performed to identify phonetic equivalents in the target language.

We represent the user query and one translation candidate in the document language by U and D, respectively. From the viewpoint of probability theory, our task is to select D's with greater probability, $P(D|U)$, which can be transformed as in Equation (1) through the Bayesian theorem.

$$P(D|U) = \frac{P(U|D) \cdot P(D)}{P(U)} \qquad (1)$$

[1] Developed by NOVA, Inc. http://www.nova.co.jp/

In practice, $P(U)$ can be omitted because this factor is a constant with respect to the given query, and thus does not affect the relative probability for different translation candidates.

We estimate $P(D)$ by a word-based bi-gram language model produced from the target collection. We estimate $P(U|D)$, commonly termed a translation model, based on the word frequency obtained from the Nova dictionary, because Japanese-English corpora with sufficient volume of alignment information is expensive.

The retrieval module is based on an existing probabilistic retrieval method [15], which computes the relevance score between the translated query and each document in the collection. The relevance score for document i is computed based on Equation (2).

$$\sum_t \left(\frac{TF_{t,i}}{\frac{DL_i}{avglen} + TF_{t,i}} \cdot \log \frac{N}{DF_t} \right) \qquad (2)$$

Here, $TF_{t,i}$ denotes the frequency that term t appears in document i. DF_t and N denote the number of documents containing term t and the total number of documents in the collection. DL_i denotes the length of document i (i.e., the number of characters contained in i), and $avglen$ denotes the average length of documents in the collection.

For both Japanese and English collections, we use content words extracted from documents as terms, and perform a word-based indexing. For the Japanese collection, we use the ChaSen morphological analyzer[2] to extract content words. However, for the English collection, we extract content words based on parts-of-speech as defined in WordNet [3].

The document translation module consists of the the Transer Japanese/English MT system, which uses the same dictionary used for the query translation module. In practice, since machine translation is computationally expensive and degrades the time efficiency, we perform machine translation on a phrase-by-phrase basis. In our case, phrases are sequences of content words in documents. This method is practical because even a word/phrase-based translation can potentially improve on the efficiency for users to find relevant foreign documents from the whole retrieval result [14].

4 Evaluation

We used the NTCIR-3 CLIR test collection to evaluate the performance of our system with respect to the J-J, E-E, J-E, and J-JE tasks.

Topics contain a number of fields, such as title, description, narrative and concept, irrespective of the language. Each system participated in the CLIR task was allow to submit more than one retrieval result using different methods. However, at least one result must be obtained with only the description field. Thus, we used descriptions (i.e., phrases and sentences) tagged with <DESCRIPTION> in topic files as test queries.

Relevance assessment was performed based on four ranks of relevance, that is, highly relevant (S), relevant (A), partially relevant (B) and irrelevant (C).

Table 1 shows non-interpolated average precision values, averaged over all the test queries, for different tasks. While in the case of "Rigid" documents judged S and A were regarded as correct answers, in the case of "Relax" documents judged B were also regarded as correct answers.

We cannot compare average precision values across tasks, because the number of topics is different depending of the task. However, since most of the English topics are translations of the Japanese topics, we can observe the general tendency.

Looking at Table 1, the average precision values of J-J and E-E did not significantly differ, irrespective of the degree of relevance. In addition, average precision values of CLIR (J-E) and MLIR (J-JE) were roughly 70-80% of that achieved with SLIR (J-J and E-E). This tendency was also observable in past literature.

Table 1. Non-interpolated average precision values, averaged over all the test queries, for different tasks.

			Avg. Precision	
Task	#Topics	#Documents	Rigid	Relax
J-J	42	236,664	0.2541	0.3427
E-E	32	22,927	0.2641	0.2967
J-E	32	22,927	0.1969	0.2149
J-JE	45	259,591	0.1970	0.2609

5 Summary

We described our system for Japanese/English CLIR, where an engine consists of query translation, retrieval, and document translation modules. In addition, re-ranking and clustering modules can optionally be used to enhance the system performance.

The NTCIR-3 collection was used to perform comparative experiments, where we evaluated the retrieval accuracy of our system. We found that the retrieval accuracy for CLIR and MLIR was approximately 70-80% of that obtained with monolingual retrieval.

Future work would include extending our system to Korean and Chinese, so that our system can be seen as multi-lingual IR system in the strict sense.

[2] http://chasen.aist-nara.ac.jp/

References

[1] L. Ballesteros and W. B. Croft. Resolving ambiguity for cross-language retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 64–71, 1998.

[2] J. G. 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 on Artificial Intelligence*, pages 708–714, 1997.

[3] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[4] A. Fujii and T. Ishikawa. Cross-language information retrieval at ULIS. In *Proceedings of the 1st NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pages 163–169, 1999.

[5] A. Fujii and T. Ishikawa. Cross-language information retrieval for technical documents. In *Proceedings of the Joint ACL SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 29–37, 1999.

[6] A. Fujii and T. Ishikawa. Applying machine translation to two-stage cross-language information retrieval. In *Proceedings of the 4th Conference of the Association for Machine Translation in the Americas*, pages 13–24, 2000.

[7] A. Fujii and T. Ishikawa. Evaluating multi-lingual information retrieval and clustering at ULIS. In *Proceedings of the 2nd NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization*, 2001.

[8] A. Fujii and T. Ishikawa. Japanese/English cross-language information retrieval: Exploration of query translation and transliteration. *Computers and the Humanities*, 35(4):389–420, 2001.

[9] J. Gonzalo, F. Verdejo, C. Peters, and N. Calzolari. Applying EuroWordNet to cross-language text retrieval. *Computers and the Humanities*, 32:185–207, 1998.

[10] M. L. Littman, S. T. Dumais, and T. K. Landauer. Automatic cross-language information retrieval using latent semantic indexing. In G. Grefenstette, editor, *Cross-Language Information Retrieval*, chapter 5, pages 51–62. Kluwer Academic Publishers, 1998.

[11] J. S. McCarley. Should we translate the documents or the queries in cross-language information retrieval? In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 208–214, 1999.

[12] J.-Y. Nie, M. Simard, P. Isabelle, and R. Durand. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 74–81, 1999.

[13] D. W. Oard. A comparative study of query and document translation for cross-language information retrieval. In *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas*, pages 472–483, 1998.

[14] D. W. Oard and P. Resnik. Support for interactive document selection in cross-language information retrieval. *Information Processing & Management*, 35(3):363–379, 1999.

[15] S. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 232–241, 1994.

[16] G. Salton. Automatic processing of foreign language documents. *Journal of the American Society for Information Science*, 21(3):187–194, 1970.