Description of the NTU Japanese-English Cross-Lingual Information Retrieval System Used for NTCIR Workshop

Chuan-Jie Lin, Wen-Cheng Lin, Guo-Wei Bian, and Hsin-Hsi Chen

Department of Computer Science and Information Engineering

National Taiwan University

Taipei, TAIWAN, R.O.C.

E-mail: {cjlin, denislin, gwbian}@nlg2.csie.ntu.edu.tw, hh_chen@csie.ntu.edu.tw

Abstract

This paper describes a Japanese-English Cross-Language Information Retrieval (CLIR) System for the evaluation in NTCIR project. We extend our work on Chinese-English CLIR to deal with this problem. Several query translation strategies, including select-all, select-first, and co-occurrence models by different corpora, and several query generation strategies, including topic, description, narrative, and concept fields, are considered. The experiment results show that (1) the combination of topic and concept fields outperform other combinations; and (2) select-first is the best, co-occurrence by NACSIS next, then co-occurrence by TREC-6 and LOB, and select-all is the worse in EO1 task. The evaluation in EO2 task is similar. Co-occurrence by TREC6 and select-first on topic and concept combination rank 1 and 2, respectively.

1. Introduction

Cross language information retrieval (CLIR) (Oard and Dorr, 1996; Oard, 1997) deals with the use of queries in one language to access documents in another. Due to the differences between source and target languages, query translation is usually employed to unify the language in queries and documents. Some different approaches have been proposed for query translation. Dictionary-based approach exploits machine-readable dictionaries and selection strategies like select all (Hull and Grefenstette, 1996; Davis, 1997), randomly select N (Ballesteros and Croft, 1996; Kwok 1997) and select best N (Hayashi, Kikui and Susaki, 1997; Davis 1997). Corpus-based approaches exploit sentence-aligned corpora (Davis and Dunning, 1996) and document-aligned corpora (Sheridan and Ballerini, 1996). These two approaches are complementary. Dictionary provides translation candidates, and corpus provides context to fit user intention. Coverage of dictionaries, alignment performance and domain shift of corpus are major problems of these two approaches. Hybrid approaches (Ballesteros and Croft, 1998; Bian and Chen, 1998; Davis 1997) integrate both lexical and corpus knowledge.

Natural Language Processing Laboratory led by Professor Hsin-Hsi Chen at Department of Computer Science and Information Engineering, National Taiwan University has studied the multilingual information management for several years. A general multilingual information architecture is proposed in the paper (Bian and Chen, forthcoming). We integrate CLIR and MT together (Bian and Chen, 1998b). In this way, users can express their information need and read the requested information in their familiar languages. The paper (Bian and Chen, 1997) presents several important issues in an on-line and real-time document translation. Besides the translation ambiguity issue in query translation (Bian and Chen, 1998a), we also touch on the target polysemy (Chen, *et al.*, 1999) and name translation issues (Chen, *et al.*, 1998).

This paper will extend our work on Chinese-English CLIR to Japanese-English CLIR. We use hybrid model, integrating dictionary-based and corpus-based approach, to resolve translation ambiguity problem. We employ dictionaries and co-occurrence statistics trained from target language documents to deal with translation ambiguity. This method considers the content around the translation equivalents to decide the best target word. The resources that we use are a bilingual dictionary and several target language corpora. A bilingual dictionary provides the translation equivalents of each query term, and the word co-occurrence information trained from target language text collection is used to disambiguate the translation. The translation of a query term can be disambiguated using the co-occurrence of the translation equivalents of this term and other terms.

The rest of this paper is organized as follows. Section 2 discusses the preprocessing of Japanese queries. Section 3 deals with the selection of translation equivalents. Section 4 touches on the evaluation and discussion. Finally Section 5 concludes the remarks.

2. Preprocessing Japanese Queries

To retrieve English documents, the Japanese queries are translated into English ones. Because Japanese queries are composed of characters without word boundaries, Japanese queries have to be segmented and morphologically analyzed. JUMAN (Kurohashi and Nagao, 1999), which is developed in Kyoto University, is adopted as the morphology analyzer. We employ JUMAN to decide word boundaries and transform each word into its root form.

We also adopt the Japanese-English dictionary EDICT (Breen, 1998) as the bilingual dictionary. Each entry in EDICT has two columns written in Kanji and Kana. When

Table 1. The Mutual Information Values for Each Word Pairs in "環 境 汚 染 問 題 "

		環 境		汚 染		問 題	
		environment	circumstance	pollution	contamination	problem	question
環 境	**environment**			**3.425872**	3.202728	0.463110	-0.205647
	circumstance					1.400663	
汚 染	**pollution**	**3.425872**				3.290200	
	contamination	3.202728				3.229575	
問 題	**problem**	0.463110	1.400663	**3.290200**	3.229575		
	question	-0.205647					

we look up a word written in Kanji, we simply search at the Kanji column to see if it exists. While looking up a word all in Kana, we search at the Kana part, and return all the possible translation equivalents. These translation equivalents are left for selection in the next step.

Because we have no stop list information of Japanese, we adopt a heuristic rule to avoid the noises. Many of the stop words, such as auxiliary verbs and prepositions, are all written in Hiraganas. Therefore we do not translate the words all in Hiraganas. That will filter out many of the stop words.

For example, the sentence "ソーラーエネルギーを用いた自動車に対する期待が高まっている。" is transformed as below:

ソーラーエネルギー /##/
を /##/
用いる /to use/to make use of/
自動車 /automobile/
に /##/
対する /##/
期待 /expectation/anticipation/hope/
が /##/
高まる /to rise/to swell/to be promoted/
いる /##/

where ## denotes no translation equivalents. Here "ソーラーエネルギー" failed to find its translation equivalents, because "ソーラー " and "エネルギー" are not separated. This is the borrowed-word segmentation problem.

3. Selecting Translation Equivalents

After getting all the translation equivalents of each query term, we have to select the most appropriate ones. The translation of a query term is disambiguated using the co-occurrence of the translation equivalents of this term and other terms.

Mutual information (Church, *et al.*, 1989) is used to measure the degree of correlation between two words. The mutual information $MI(x,y)$ is defined as follow:

$$MI(x,y) = \log_2 \frac{p(x,y)}{p(x)p(y)}$$

where x and y are words; $p(x)$ and $p(y)$ are probabilities of words x and y; $p(x,y)$ is their co-occurrence probability.

Our CO-Model employs the mutual information to select appropriate translation equivalents. The co-occurrence information is trained from a monolingual corpus. The number of word pairs co-occurring within a window of size 3 is counted to compute the probability $p(x,y)$. The advantage of such CO-Model is that neither bilingual corpus nor aligned corpus is needed.

After translation equivalents of each query term are retrieved from bilingual dictionary, MI values are used to select the best one. To determining an appropriate translation equivalent of a query term, we compare the MI values of each translation equivalent of this query term and all the other translation equivalents of other query terms within a sentence. The translation equivalent with the highest MI value is regarded as the best translation equivalent of this query term. All the selected translations comprise the final English query.

For example, the phrase "環境汚染問題" consists of three terms, each having two translation equivalents, as illustrated below:

環 境 /environment/circumstance/
汚 染 /pollution/contamination/
問 題 /problem/question/

The MI value between each pair of equivalents is show in Table 1. For the term 環境, the highest MI score with other terms is 3.425872, i.e., the pair <environment, pollution>. Therefore, 'environment' is selected as 環境's translation. Similarly, 'pollution' and 'problem' are select as汚染's and問題's translations, respectively. The final translated phrase is 'environment pollution problem'.

Several query translation strategies, including select-all, select-first, and co-occurrence models by different corpora, are considered. They are depicted in the first row of Tables 2 and 3. Select-First simply selects the first equivalent listed in the bilingual dictionary. Select-All selects all the equivalents. CO-LOB, CO-NACSIS, and CO-TREC6 use the co-occurrence information, and the mutual information tables are trained from LOB, NACSIS, TREC6 corpora, respectively. We try to measure the effects of different corpora.

4. Evaluation and Discussion

The original Japanese queries have four fields: Title, Description, Narrative, and Concept. As shown in Table 2, we constructed five sets of queries with various combinations of fields.

Table 2. Run Description in Terms of Query Translation Strategies and Query Generation Strategies

	Select First	Select All	CO-LOB	CO-NACSIS	CO-TREC6
d	tstar1	tstar2	tstar3	tstar17	tstar13
dc	tstar4	tstar5	tstar6	tstar18	tstar14
tc	tstar7	tstar8	tstar9	tstar19	tstar15
tdc	tstar10	tstar11	tstar12	tstar20	tstar16
tdnc	tstar21	tstar22			tstar23

Table 3: Average precisions (non-interpolated) for all relevant documents of EO1 task

	Select First	Select All	CO-LOB	CO-NACSIS	CO-TREC6
d	0.0789 (20,19)	0.0750 (21,20)	0.0613 (23,23)	0.0790 (19,21)	0.0712 (22,22)
dc	0.2322 (7, 2)	0.1944 (15,16)	0.2174 (10,12)	0.2315 (8,10)	0.2264 (9,11)
tc	0.2437 (2, 3)	0.2023 (13,17)	0.2339 (6, 9)	0.2424 (3, 7)	0.2443 (1, 8)
tdc	0.2402 (4, 1)	0.2007 (14,15)	0.2115 (12, 6)	0.2398 (5, 4)	0.2143 (11, 5)
tdnc	0.1766 (17,14)	0.1799 (16,18)			0.1723 (18,13)

There are 23 runs in our experiments. The precedence of these 23 runs requested by NTCIR project organizers are shown as follows. The sequence is determined by our expectation. Long query is preferred to short query, and select-first model is preferred to co-occurrence model, and select-all model.

1. tstar10, 2. tstar4, 3. tstar7, 4. tstar20,
5. tstar16, 6. tstar12, 7. tstar19, 8. tstar15,
9. tstar9, 10. tstar18, 11. tstar14, 12. tstar6,
13. tstar23, 14. tstar21, 15. tstar11, 16. tstar5,
17. tstar8, 18. tstar22, 19. tstar1, 20. tstar2
21. tstar17, 22. tstar13, 23. tstar3

Table 3 shows the average precision (non-interpolated) for all relevant documents of EO1 task. The parentheses followed by the average precision indicate the ranks of the runs in our experiments. The first number denotes the ranks of evaluation and the second number denotes the ranks of our expectation. To evaluate the above tasks, we employ SMART information retrieval system (Salton and Buckley, 1988).

After investigating the performance of our experiments, we found that Select-First performs the best, then CO-NACSIS, CO-TREC6, CO-LOB, and Select-All in order. In fact, the performances of Select-First, CO-NACSIS, and CO-TREC6 are similar. It is consistent to our original expectation to some degree.

We are a little surprised that CO-Model does not perform better than Select-First model. We did an investigation in much detail. We learned that CO-Model needs a large-scale and balanced corpus for training. NACSIS Corpus is not large enough. MI scores of many pairs are not found while selecting translation equivalents. LOB Corpus is balanced, but it is not large enough either. TREC6 Corpus is large. CO-Model works fine with the queries translated from TREC-6 topics 301-350 in Chinese-English CLIR experiments (Chen, *et al.*, 1999). However, CO-TREC6 does not outperform Select-First this time. We found that most of the NACSIS queries are in the domain of electronics. Many of the query terms are domain-specific. TREC6 Corpus is not balanced enough to cover the terminology. That is why CO-TREC6 does not perform better than Select-First. We believe that if the NACSIS Corpus is large enough, CO-NACSIS will outperform Select-First.

Another investigation is done from the query generation side. In the respect of query generation, runs of which queries consist of the topic field and the concept field perform the best. The ranking order is TC > TDC > DC > TDNC > D, where T, D, C, and N are the abbreviations of the Topic, Description, Concept, and Narrative fields, respectively, and the concatenation of the abbreviations denotes the combination of fields. Generally speaking, the performance of queries including concept field is better. It is not surprising since the concept field includes English keywords. Long queries are usually better than short queries. However, narrative field may introduce noise, so that performance is dropped when this field is included.

Table 4 shows the evaluation results in task EO2. Partially relevant documents are also considered to be correct. The ranks of 15 runs (of 23 runs) are not changed in EO2. The runs of ranks 1-6 are reordered. CO-TREC6, Select-first, and CO-NACSIS on topic-and-concept combination rank 1, 2, and 3, respectively.

5. Concluding Remarks

This paper adopts SMART information retrieval system to evaluate the strategies of query generation and query translation in Japanese-English CLIR. The experiments show that topic and concept combination is a better choice for query generation. Select-first and co-occurrence on appropriate corpus demonstrate better performance on query translation.

Table 4: Average precisions (non-interpolated) for all relevant documents of EO2 task.

	Select First	Select All	CO-LOB	CO-NACSIS	CO-TREC6
d	0.0728 (19, 19)	0.0663 (21, 20)	0.0580 (23, 23)	0.0728 (19, 21)	0.0650 (22, 22)
dc	0.2179 (7, 2)	0.1761 (15, 16)	0.2050 (10, 12)	0.2163 (8, 10)	0.2059 (9, 11)
tc	0.2299 (1, 3)	0.1875 (13, 17)	0.2202 (4, 9)	0.2286 (2, 7)	0.2271 (3, 8)
tdc	0.2197 (5, 1)	0.1851 (14, 15)	0.1920 (12, 6)	0.2193 (6, 4)	0.2019 (11, 5)
tdnc	0.1750 (16, 14)	0.1650 (17, 18)			0.1607 (18, 13)

References

Ballesteros, L. and Croft, W.B. (1996) "Dictionary-based Methods for Cross-Lingual Information Retrieval." *Proceedings of the 7th International DEXA Conference on Database and Expert Systems Applications*, 791-801.

Ballesteros, L. and Croft, W.B. (1998) "Resolving Ambiguity for Cross-Language Retrieval." *Proceedings of 21st ACM SIGIR*, 64-71.

Bian, G.W. and Chen, H.H. (1997) "An MT-Server for Information Retrieval on WWW." *Working Notes of the AAAI Spring Symposium on Natural Language Processing for the World Wide Web*, 1997, 10-16.

Bian, G.W. and Chen, H.H. (1998a) "A New Hybrid Approach for Chinese-English Query Translation." *Proceedings of the First Asia Digital Library Workshop*, August 6-7 1998, 156-167.

Bian, G.W. and Chen, H.H. (1998b) "Integrating Query Translation and Document Translation in a Cross-Language Information Retrieval System." *Machine Translation and Information Soup*,, Lecture Notes in Computer Science, No. 1529, Spring-Verlag, 250-265.

Bian, G.W. and Chen, H.H. (forthcoming) "Cross Language Information Access to Multilingual Collections on the Internet." *Journal of American Society for Information Science*,,*Special Issue on Digital Libraries*.

Breen, J. (1998) EDICT, School of Computer Science & Software Engineering, Monash University, Australia, 1998.
URL: http://ftp.monash.edu.au/pub/nihongo/edict.gz

Chen, H.H., *et al.* (1998) "Proper Name Translation in Cross-Language Information Retrieval." *Proceedings of 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics*,, 1998, 232-236.

Chen, H.H., *et al.* (1999) "Resovling Translation Ambiguity and Target Polysemy in Cross-Language Information Retrieval." *Proceedings of 37th Annual Meeting of the Association for Computational Linguistics*,, 1999, 215-222.

Davis, M.W. (1997) "New Experiments in Cross-Language Text Retrieval at NMSU's Computing Research Lab." *Proceedings of TREC 5*,, 39-1~39-19.

Davis, M.W. and Dunning, T. (1996) "A TREC Evaluation of Query Translation Methods for Multi-lingual Text Retrieval." *Proceedings of TREC-4*,, 1996.

Hayashi, Y., Kikui, G. and Susaki, S. (1997) "TITAN: A Cross-linguistic Search Engine for the WWW." *Working Notes of AAAI-97 Spring Symposiums on Cross-Language Text and Speech Retrieval*,, 58-65.

Hull, D.A. and Grefenstette, G. (1996) "Querying Across Languages: A Dictionary-based Approach to Multilingual Information Retrieval." *Proceedings of the 19th ACM SIGIR*,, 49-57.

Kowk, K.L. (1997) "Evaluation of an English-Chinese Cross-Lingual Retrieval Experiment." *Working Notes of AAAI-97 Spring Symposiums on Cross-Language Text and Speech Retrieval*,, 110-114.

Kurohashi, S. and Nagao, M. (1999) JUMAN Version 3.61, Nagao Laboratory, Kyoto University, 1999.
URL: http://pine.kuee.kyoto-u.ac.jp/nl-resource/juman-e.html

Oard, D.W. (1997) "Alternative Approaches for Cross-Language Text Retrieval." *Working Notes of AAAI-97 Spring Symposiums on Cross-Language Text and Speech Retrieval*,, 131-139.

Oard, D.W. and Dorr, B.J. (1996) *A Survey of Multilingual Text Retrieval.* Technical Report UMIACS-TR-96-19, University of Maryland, Institute for Advanced Computer Studies. http://www.ee.umd.edu/medlab/filter/papers/mlir.ps.

Salton, G. and Buckley, C. (1988) "Term Weighting Approaches in Automatic Text Retrieval." *Information Processing and Management*,, Vol. 5, No. 24, 513-523.

Sheridan, P. and Ballerini, J.P. (1996) "Experiments in Multilingual Information Retrieval Using the SPIDER System." *Proceedings of the 19th ACM SIGIR*,, 58-65.