Cross-Lingual Information Retrieval (CLIR) Task at the NTCIR Workshop 3

Kazuaki Kishida
Surugadai University
Hanno 357-8555, Japan
kishida@surugadai.ac.jp

Kuang-hua Chen
National Taiwan University
Taipei 10617, Taiwan
khchen@ntu.edu.tw

Sukhoon Lee
Chungnam National University
Taejon 305-764 Korea
shlee@stat.cnu.ac.kr

Hsin-Hsi Chen
National Taiwan University
Taipei 10617, Taiwan
hh_chen@ccms.ntu.edu.tw

Noriko Kando
National Institute of Informatics (NII)
Tokyo 101-8430, Japan
kando@nii.ac.jp

Kazuko Kuriyama
Shirayuri College
Chofu 182-8525, Japan
kuriyama@shirayuri.ac.jp

Sung Hyon Myaeng
Information and Communications University
Daejeon 305-600, Korea
myaeng@icu.ac.kr

Koji Eguchi
National Institute of Informatics (NII)
Tokyo 101-8430, Japan
eguchi@nii.ac.jp

1. Introduction

The CLIR task at the NTCIR-3 Workshop aims to promote research efforts on cross-lingual text retrieval among three East Asian languages (Chinese, Japanese and Korean) and English. This is achieved through TREC-style IR experiments, in which large-scale document collections and search topics were prepared by organizers, and the search results submitted by participants were evaluated using the so-called pooling method. The third workshop started from October 2001, and the final workshop was held in October 2002 in Tokyo. In total, over 20 teams from eight countries (including Canada and USA) submitted final search results, and shared and discussed findings of the experiments. The overview of the task [Chen et al. 2002] has further details.

2. Design and Test Collection

The task organizers prepared a multilingual document collection, which included news articles written in Mandarin Chinese with Traditional Chinese Characters (C), Japanese (J), Korean (K) and English (E) languages. Unfortunately, while the document sets for CJE (Chinese, Japanese, and English) comprised records of news articles published in 1998–99, the publication year of the Korean articles in the collection was 1994. Therefore, we divided the collection into CJE (1998-99) and Korean (1994) sections. A total of 624,686 (CJE) and 66,146 (Korean), documents were used.

For the document collection, three types of search runs were recommended for execution: (1) SLIR (single language IR); (2) BLIR (bilingual CLIR); and (3) MLIR (multilingual CLIR). In the case of SLIR, the language of the search topics is identical to that of the documents (i.e., this is not a cross-lingual task). BLIR denotes that a document set in a single language is searched

using topics in a different language (e.g., using Chinese topics for Japanese documents), and the MLIR is a search task where the target collection consists of documents in two or more languages (e.g., searching CJE collection for Chinese topics).

The task organizers prepared 50 topics for the 1998–99 CJE collection and 30 topics for the 1994 Korean collection. Both sets include topics in all four languages (CJKE). The topics were delivered to participants on December 2001, and the participants were requested to execute runs for a month. It should be noted that these are "formal" runs, and trials (or "dry" runs) were previously executed between October and November 2001.

3. Results

In total, 189 runs were submitted, of which 110 were runs for SLIR, 50 were for BLIR and 29 were for MLIR. While SLIR runs were executed by almost all research groups participating in the CLIR, only seven groups submitted MLIR results. Many participants focused on the search tasks from English to Asian languages (e.g., E–C, E–J) for BLIR runs. This is partly because each participant has an indexing system for their own language and bilingual resources for English are more available. It seems that research efforts on CLIR for East Asian languages have not yet been attempted for such large-scale IR experiments before the NTCIR-3 CLIR task. In this sense, the CLIR task gave a good opportunity for developing methods and algorithms for CLIR related to East Asian languages. As a next step, MLIR or other patterns of BLIR are expected to attract more research groups.

A typical CLIR method employed in the NTCIR-3 is query translation using bilingual dictionaries or machine translation systems. Many participants tried enhancing search performance with automatic query expansion techniques (e.g., pseudo-relevance feedback, or PRF). Some groups attempted to incorporate more complicated CLIR techniques into their systems, such as translation disambiguation, corpus-based translation, or merging of ranked output lists for MLIR.

Microsoft Research Asia [He and Gao 2003] made use of an original technique for translation disambiguation, the "decaying co-occurrence model," and of a phrase translation model for E–C CLIR. Another interesting approach is corpus-based translation by UC Berkeley [Chen and Gey 2003], in which some translations are specified using results from web search engines. The UC Berkeley group also attempted a kind of cognate matching between Chinese and Japanese through conversion of Kanji character codes and found that hybrid of query translation and cognate matching worked well. National Taiwan University [Lin and Chen 2003] explored a unique MLIR strategy for merging ranked lists based on translation performance.

There are many other remarkable outcomes. For example, Thomson Legal and Regulatory [Moulinier et al. 2003] and Hummingbird [Tomlinson 2003] compared performance between various indexing techniques for CJK, and Hong Kong Polytechnic University [Luk et al. 2003] examined the comparative effectiveness of some indexing methods and retrieval models (such as vector, 2-Poisson, logistic regression and Pircs) on Chinese information retrieval. In addition, Toshiba [Sakai et al. 2003] intensively investigated the performance of variations of PRF, and the Communication Research Laboratory [Murata et al. 2003] attempted to incorporate information about keyword location into the calculation of document scores.

4. About the CLIR task at NTCIR-4 Workshop

The next workshop, NTCIR-4, contains a CLIR task again, in which a set of Korean news articles published in 1998–99 is added, allowing the participants to conduct experiments on a multilingual CJKE document set. Furthermore, the CJE document sets (1998–99) are augmented so that the size of the document sets in each language are well balanced (In the NTCIR-3 collection, the English part is relatively small). As a result, each language part of the NTCIR-4 collection consists of two or more sources. For example, the Japanese part includes articles from Mainichi Newspapers and Yomiuri Newspapers, and the entire document collection, including the Korean 1998–99 set, contains more than 1,579,000 records in total.

From a technical viewpoint, the task organizers are recommending that participants explore the "pivot language approach," in which an intermediary language (e.g., English) is used for translating search topics. At the NTCIR-3 task, only a small number of runs were submitted for CLIR between CJK languages (C–J, C–K and J–K), probably due to insufficient language resources. However, it is easier to obtain resources between English and each of the three languages (E–C, E–J and E–K). Therefore, we can execute the CLIR task between CJK languages without direct resources by using transitive translation of topics, for example, translating Chinese topics into English (C to E) and the English translation into Japanese (E to J).

From the experiences at NTCIR-3, the task organizers consider that one of the key elements of CLIR between CJKE is the proper noun. If the proper noun can be correctly translated, good search performance becomes easier. Inversely, untranslatable proper nouns cause serious problems, because cognate matching may have almost no effect between CJKE. The translatability is related to the regional locality of topics. For example, it would be difficult to translate a name of an internationally unknown Japanese person into Chinese using machine translation systems. Therefore, in the NTCIR-4 CLIR task, the set of topics was carefully created, considering the balance between locality and internationality, and the balance between proper nouns and abstract nouns). Further information on the NTCIR-4 CLIR task is available at http://research.nii.ac.jp/ntcir/ntcir-ws4/clir/index.html.

5. References

Chen, A. and Gey, F. C., "Experiments on cross-language and patent retrieval at NTCIR-3 workshop". In *NTCIR Workshop 3: Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Question Answering and Summarization*, Tokyo Japan, Oct. 2001–Oct. 2002. NII, Tokyo (2003) (http://research.nii.ac.jp/ntcir/ workshop/OnlineProceedings3/NTCIR3-CLIR-ChenA.pdf)

Chen, K. H., Chen, H. H., Kando, N., Kuriyama, K., Lee, S., Myaeng, S. H., Kishida, K., Eguchi, K. and Kim, H., "Overview of CLIR Task at the third NTCIR workshop". In *NTCIR Workshop 3: Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Question Answering and Summarization*, Tokyo Japan, Oct. 2001–Oct. 2002. NII, Tokyo (2003) (http://research.nii.ac.jp/ntcir/ workshop/OnlineProceedings3/NTCIR3-OV-CLIR-ChenK.pdf)

He, H. and Gao, J., "NTCIR-3 CLIR experiments at MSRA". In *NTCIR Workshop 3: Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Question Answering and Summarization*, Tokyo Japan, Oct. 2001–Oct. 2002. NII, Tokyo (2003) (http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings3/NTCIR3-CLIR-HeH.pdf)

Lin, W. C. and Chen, H. H., "Description of NTU at NTCIR3: Multilingual information retrieval". In *NTCIR Workshop 3: Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Question Answering and Summarization*, Tokyo Japan, Oct. 2001–Oct. 2002. NII, Tokyo (2003) (http://research.nii.ac.jp/ntcir/ workshop/OnlineProceedings3/NTCIR3-CLIR-LinW.pdf)

Luk, R. W. P., Wong, K. F. and Kwok, K. L., "Different retrieval models and hybrid term indexing". In *NTCIR Workshop 3: Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Question Answering and Summarization*, Tokyo Japan, Oct. 2001–Oct. 2002. NII, Tokyo (2003) (http://research.nii.ac.jp/ntcir/ workshop/OnlineProceedings3/NTCIR3-CLIR-LukR.pdf)

Moulinier, I., Molina-Salgado, H. Jackson, P., "Thomson Legal and Regulatory at NTCIR-3: Japanese, Chinese and English Retrieval Experiments". In *NTCIR Workshop 3: Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Question Answering and Summarization*, Tokyo Japan, Oct. 2001–Oct. 2002. NII, Tokyo (2003) (http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings3/NTCIR3-CLIR-MoulinierI.pdf)

Murata, M, Ma, Q, and Isahara, H., "Applying multiple characteristics and techniques to obtain high levels of performance in information retrieval". In *NTCIR Workshop 3: Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Question Answering and Summarization*, Tokyo Japan, Oct. 2001–Oct. 2002. NII, Tokyo (2003) (http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings3/NTCIR3-CLIR-MurataM.pdf)

Sakai, T., Koyama, M., Suzuki, M. and Manabe, T., "Toshiba KIDS at NTCIR-3: Japanese and English-Japanese IR". In *NTCIR Workshop 3: Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Question Answering and Summarization*, Tokyo Japan, Oct. 2001–Oct. 2002. NII, Tokyo (2003) (http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings3/NTCIR3-CLIR-SakaiT.pdf)

Tomlinson, S., "Asian language parsing evaluated by Hummingbird SearchServer™ at NTCIR-3". In *NTCIR Workshop 3: Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Question Answering and Summarization*, Tokyo Japan, Oct. 2001–Oct. 2002. NII, Tokyo (2003) (http://research.nii.ac.jp/ntcir/ workshop/OnlineProceedings3/NTCIR3-CLIR-TomlinsonS.pdf)