Chapter 2
Experiments on Cross-Language Information Retrieval Using Comparable Corpora of Chinese, Japanese, and Korean Languages

Kazuaki Kishida and Kuang-hua Chen

Abstract This paper describes research activities for exploring techniques of cross-language information retrieval (CLIR) during the NACSIS Test Collection for Information Retrieval/NII Testbeds and Community for Information access Research (NTCIR)-1 to NTCIR-6 evaluation cycles, which mainly focused on Chinese, Japanese, and Korean (CJK) languages. First, general procedures and techniques of CLIR are briefly reviewed. Second, document collections that were used for the research tasks and test collection construction for retrieval experiments are explained. Specifically, CLIR tasks from NTCIR-3 to NTCIR-6 utilized multilingual corpora consisting of newspaper articles that were published in Taiwan, Japan, and Korea during the same time periods. A set of articles can be considered a "pseudo" comparable corpus because many events or affairs are commonly covered across languages in the articles. Such comparable corpora are helpful for comparing the performance of CLIR between pairs of CJK and English. This comparison leads to deeper insights into CLIR techniques. NTCIR CLIR tasks have been built on the basis of test collections that incorporate such comparable corpora. We summarize the technical advances observed in these CLIR tasks at the end of the paper.

2.1 Introduction

A "comparable corpus" can be defined as multiple sets of documents, each in different languages, which approximately describe the same things or events. Unlike a parallel corpus, explicit alignments of words, sentences, paragraphs, or documents are not necessarily contained in the comparable corpus. In this sense, pairs of scientific abstracts written in Japanese and English that were used for retrieval experiments

K. Kishida (✉)
Keio University, Mita 2-15-45, Minato-ku, Tokyo 108-8345, Japan
e-mail: kz_kishida@keio.jp

K. Chen
National Taiwan University, No.1, Sec.4, Roosevelt Rd., Taipei 10617, Taiwan
e-mail: khchen@ntu.edu.tw

T. Sakai et al. (eds.), *Evaluating Information Retrieval and Access Tasks*,
The Information Retrieval Series 43,
https://doi.org/10.1007/978-981-15-5554-1_2

during the first and second *NACSIS Test Collection for Information Retrieval/NII Testbeds and Community for Information access Research (NTCIR)* evaluation cycles (i.e., NTCIR-1 and -2) as test documents can be considered document-linked comparable corpora.

Scholarly journals or conference proceedings published in Japan often ask authors to attach an English title and abstract to their Japanese paper to promote scientific communication. Such a set of Japanese and English titles and abstracts is a parallel corpus, in which explicit alignments of titles or abstracts may be included if the authors attempted to write the English title or abstract such that they were equivalent to those in Japanese. Even though all the authors did not necessarily do so, the set can be regarded as a comparable corpus at least.

In NTCIR-1, a corpus of such titles and abstracts was used for experiments of *cross-language information retrieval (CLIR)* in which English (E) documents were searched for Japanese (J) queries (i.e., a J to E bilingual search). Note that even if only a monolingual corpus in English is available, J to E bilingual searching can be tested by creating Japanese queries as search topics. However, Japanese and English comparable (or parallel) corpora allow us to compare results of J to E and E to J searching in a controlled setting, as the two target document sets in Japanese and English are topically similar. This type of comparison would play an important role in developing more sophisticated CLIR techniques. Actually, in NTCIR-2, a research task of E to J searching was added.

This policy of designing CLIR experiments based on comparable corpora had been maintained for NTCIR-3 to -6, in which CLIR between Chinese (C), Japanese (J), Korean (K), and English (E) was explored as one of the research tasks. More specifically, as target documents, NTCIR CLIR tasks used newspaper articles published in Taiwan, Japan, and Korea during the same time periods, which can be considered to be topically sufficiently comparable because they include many descriptions of common events and affairs occurring globally or locally in regions of East Asia. Actually, a comparison between pairs of CJKE languages based on such document sets largely contributed to the development of CLIR techniques between the CJKE languages even though the sets of the CJKE newspaper articles were "more loosely" comparable corpora than the sets of Japanese and English titles and abstracts in NTCIR-1 and -2.

This paper mainly describes research efforts of CLIR tasks from NTCIR-3 to -6. Specifically, construction of test collections based on so-called "pseudo comparable corpora" (i.e., time- and region-aligned newspaper article sets) and CLIR techniques that were explored by research groups participating in the NTCIR CLIR tasks are the focus. In addition, CLIR experiments in NTCIR-1 and -2 are briefly mentioned before reviewing the NTCIR CLIR tasks. The NTCIR-3 CLIR task started on September in 2001 and the NTCIR-6 CLIR task ended on May 2007. Therefore, readers can understand the technical development of CLIR among CJKE during the time period from a historical perspective.

2.2 Outline of Cross-Language Information Retrieval (CLIR)

Before describing research efforts of NTCIR CLIR tasks, this section gives a concise, general overview of CLIR operations. Grefenstette (1998), Oard and Diekema (1998), Nie (2010), and Peters et al. (2012) provide more in-depth coverage of CLIR. Note that this section is based on a review article (Kishida 2005), which includes an exhaustive reference list on CLIR techniques that this section describes.

2.2.1 CLIR Types and Techniques

Some form of CLIR is required when a search query and target documents are written in different languages. If only a single language is used in documents then the task is termed *bilingual information retrieval (BLIR)*. An example is J to E searching, in which only English documents are involved. In the case of *multilingual information retrieval (MLIR)*, the target set consists of documents in two or more languages. In the NTCIR CLIR tasks, the most difficult challenge was to search a set of documents in four languages (CJKE). Note that if a query is written in C then standard monolingual information retrieval (i.e., C to C searching) may be included as a part of MLIR on the CJKE documents. Monolingual IR was specifically referred to as single language IR (SLIR) in the NTCIR CLIR tasks. Therefore, NTCIR CLIR tasks had three subtasks: SLIR, BLIR and MLIR.

Generally, research efforts of CLIR can be traced back to a work by Gerald Salton in 1970 (Kishida 2005). Many researchers had attempted to develop CLIR techniques, particularly since the 1990s following popularization of the Internet. At that time, the main research task was to explore cross-lingual techniques for conventional ad hoc IR, which was also focused on by NTCIR CLIR tasks. However, it is possible to apply cross-lingual techniques to other applications related to ad hoc IR.

An important operation for CLIR is to translate a query and/or individual documents. If the query is perfectly translated into a language of the target documents via *machine translation (MT)* software then CLIR transforms back to normal monolingual IR. However, the translation is often incomplete because the queries are generally short and ambiguous (Oard and Diekema 1998). For example, when a query including only two single words "mercury earth" is entered into a search engine, the "mercury" in the source language has to be correctly translated into an equivalent that corresponds to a planet in the target language, not the chemical substance, in most cases. Sense disambiguation is often difficult because the queries may not contain sufficient contextual information for determining the correct meaning of each query term. To maintain the accuracy of the translation, it may be better to translate documents that are typically longer than the queries although document translation is more time-consuming in comparison to query translation. Another difficulty using

document translation is that index files of the IR system increase in size because translations have to be added as index terms.

Therefore, CLIR techniques typically consist of two main modules: (1) translation and (2) monolingual IR. Their effectiveness has an influence on the overall CLIR performance in the case that translation and monolingual searching independently work to generate a final search output, which would be a typical architecture of CLIR systems. However, there are IR models more sophisticatedly incorporating both modules. For example, *language modeling (LM)* can elegantly implement a CLIR operation by combining two conditional probabilities $p(s|t)$ and $p(t|d)$ during a process of computing document scores for ranked output where s and t denote a query term and a term in document d, respectively (Xu et al. 2001). Particularly, $p(s|t)$ is termed as translation probability.

2.2.2 Word Sense Disambiguation for CLIR

As previously exemplified by an instance of "mercury," *word sense disambiguation (WSD)* is important in CLIR. Typical methods for WSD in CLIR utilize (1) *part-of-speech (POS)* tags, (2) term co-occurrence statistics in the target document set, and (3) *pseudo relevance feedback (PRF)* techniques.

When POS tags are used, target terms having the same POS tags as the source term are selected from a set of candidates as final query terms. The candidate target terms can be easily obtained from a machine-readable bilingual dictionary.

In the case of utilizing term co-occurrence statistics in the target document set, the operation is more complicated. It is assumed that two translations t_1 and t_2 are extracted from a bilingual dictionary for a query term and that other translations u_1 and u_2 are similarly obtained for another term in the same query. If t_1 and u_1 are semantically correct translations in the context of the given query then it is expected that t_1 and u_1 co-occur more frequently in the target corpus than a pair of t_1 and u_2 and that of t_2 and u_1. Therefore, the co-occurrence frequencies aid in selecting final query terms in the target language, which is a basic assumption of the disambiguation method. When a large number of terms are included in an original source query, too many translations may be extracted from the dictionary. Because selection of final query terms is computationally expensive in such cases, some special techniques for solving the problem have been explored thus far (Kishida 2005).

Whereas the co-occurrence frequencies have to be computed before actual searching, such a type of preparatory work is not required for applying disambiguation techniques based on PRF. Instead, the searching operation is repeated during the process, which may be time-consuming in a real situation. That is, first, the target document collection is searched for a set of all translations that were obtained from a dictionary, and thereafter, final query terms are selected from the set of top-ranked documents (e.g., from the top 30 documents). Searching for the selected query terms is again repeated to obtain a final result.

Originally, the PRF attempts to expand a given query by adding some "significant" terms in the top-ranked documents to the query under an assumption that they also indicate an information need represented by the original query. The newly added terms may mainly contribute to enhancing the recall ratio. In the context of disambiguation for CLIR, it is expected that documents including a semantically correct combination of target terms (e.g., t_1 and u_1 in the aforementioned example) are at a higher position in a ranked list of the first search (Ballesteros and Croft 1997). As a result, terms that co-occur in the top-ranked documents tend to be selected, which has the same effect as using term co-occurrence statistics. Thus, the selection based on the top-ranked documents works incidentally as a system for disambiguation. Note that the co-occurrences in the top-ranked documents are limited to a local context of the original query, unlike term co-occurrence statistics in the entire document set. Final query terms are typically selected according to term weights that are calculated using a formula of standard PRF techniques (Kishida 2005).

2.2.3 Language Resources for CLIR

As mentioned previously, a typical language resource for implementing CLIR is a machine-readable bilingual dictionary or MT software. When both the dictionary and the software are not available for a given pair of source and target languages, it is possible to apply a pivot language approach. For example, even if a resource between Japanese and Swedish (S) is not found, J to English and English to S resources allow us to execute J to S bilingual searching, where English is a pivot language. More specifically, by translating each Japanese query term into English equivalents and converting them again to Swedish terms, a final Swedish query can be obtained. Thus, the resulting Swedish query can be used for retrieval of the Swedish documents. Because English is an international language, many language resources related to English are actually available.

In addition, parallel corpora play an important role in CLIR. Without a dictionary or MT software, CLIR can be executed by searching a parallel corpus for a query written in the source language. That is, because textual data that were found via searching have another part in the target language, it is possible to extract final query terms in the target language from the data. Additionally, a parallel corpus consisting of sentence alignments can be used for estimating translation probabilities, in which the well-known IBM Model 1 for statistical MT has often been applied. The list of the translation probabilities works as a bilingual dictionary, and is indispensable for LM-based CLIR (see Sect. 2.2.1).

Of course, standard language processing tools such as a POS tagger (or a morphological analyzer), a stemmer, and a named entity recognizer are also employed in CLIR.

2.3 Test Collections for CLIR from NTCIR-1 to NTCIR-6

A main contribution of the NTCIR CLIR tasks is to examine whether or not the CLIR techniques that were reviewed in the previous section can be applied to the CJK languages and to enhance the techniques by tailoring them to situations in which CJK languages are used. When the NTCIR CLIR task started, the Chinese language had been already explored in the *Text REtrieval Conference (TREC)* (Voorhees and Harman 2005). In contrast, a systematic and large-scale CLIR experiment related to the Japanese and Korean languages would be considered as an original NTCIR contribution. This section provides a simple overview of test collections on which various trial-and-error attempts were made in NTCIR from the very beginning.

2.3.1 Japanese-English Comparable Corpora in NTCIR-1 and NTCIR-2

As previously mentioned, a set of Japanese and English titles and abstracts in conference proceedings that were published by Japanese academic societies was a source of documents in NTCIR-1. More specifically, in total, 339,483 bibliographic records of conference papers were collected. Because the set included three types of records having (1) only Japanese abstracts, (2) only English abstracts, and (3) both Japanese and English abstracts, the set of Japanese documents (J collection) and the set of English documents (E collection) were constructed as a subset of the whole set (JE collection). Research groups participating in NTCIR-1 were able to use the three sets during IR experiments (Kando et al. 1999).

All search requests for the experiments (i.e., search topics) were written in Japanese (30 topics for training and 53 topics for evaluation). Therefore, it was possible for the participants to examine only J to E bilingual searching as CLIR experiments. The NTCIR-1 conference was held in September of 1999, which would be the first opportunity for discussing internationally CLIR issues related to Japanese language.

In NTCIR-2, by adding bibliographic records of some scientific reports published in Japan, the document sets were substantially extended. The English and Japanese versions of 49 search topics were prepared by the task organizers (Kando et al. 2001). The test collection allowed the participants to experiment in E to J and J to E bilingual searching and in J to JE and E to JE multilingual searching.

2.3.2 *Chinese-Japanese-Korean (CJK) Corpora from NTCIR-3 to NTCIR-6*

Based on the knowledge that was obtained by the efforts of NTCIR-1 and -2, more sophisticated CLIR experiments involving C, J, K, and E were started as an independent task beginning with NTCIR-3. In the CLIR tasks from NTCIR-3 to -6, newspaper articles that were collected from various news agencies in East Asia were employed as target documents. Each record of the articles included its headline and full text. Table 2.1 summarizes the document sets; the number of documents is indicated in Table 2.2. Note that the CLIR task of NTCIR-6 had two stages (i.e., stages 1 and 2). The purpose of stage 2 was to obtain a more reliable measurement of search performance. Newspaper articles published in 1998 and 1999 were basically used for experiments in NTCIR-3 and -4 whereas newspaper articles for NTCIR-5 and stage 1 in NTCIR-6 were from 2000 and 2001.

For some reason, only the Korean document set in NTCIR-3 consisted of newspaper articles in 1994. However, from NTCIR-4, newspaper articles matching time periods (i.e., 1998–99 and 2000–01) were provided as CJKE document sets for experiments (English documents were out of scope in NTCIR-6). As previously discussed, the sets can be considered as types of comparable corpora because the newspaper articles in the sets were commonly concerned with worldwide or East Asian events and affairs of the time, allowing a CLIR performance comparison between the pairs of CJKE languages partly because documents in the individual languages are topically homogeneous to some extent. Notably, the Chinese documents were represented by only traditional Chinese characters, not simplified ones.

A newspaper article is typically written for general audiences; its text is relatively plain and shorter in comparison to that of scientific or technical papers. There is no explicit structure in the text of newspaper articles except for a headline and paragraphs, which is different from XML documents having a more complex structure. Additionally, newspaper article records in NTCIR CLIR tasks did not include any

Table 2.1 Document sets used by NTCIR CLIR tasks[a]

	Period of tasks	Date of newspaper articles	Set for MLIR
NTCIR-3	2001–02	C, J, E: 1998–99, K:1994	CJ, CE, JE, CJE
NTCIR-4	2003–04	C, J, K, E: 1998–99	CJE, CJKE
NTCIR-5	2004–05	C, J, K, E: 2000–01	CJKE
NTCIR-6	2006–07		
Stage1		C,J,K:2000-01	CJK
Stage2		NTCIR-3, –4, –5 test collections[b]	

[a]Search topics in C, J, K, and E were created for the document sets
[b]In stage 2 of NTCIR-6, a cross-collection analysis was attempted

Table 2.2 Number of records in document sets in the NTCIR-3 to -6 CLIR tasks

Language	No. of records	Usage (denoted by the mark x)				
		–3	–4	–5	–6	
					Stage 1	Stage 2
1994						
Korean	Korea Economic Daily: 66,146	x				x
1998–99						
Chinese	UDN[a]+others: 381,375	x	x			x
Japanese	Mainich: 220,078	x	x			x
	Yomiuri: 373,558		x			x
Korean	Hankookilbo+Chosunilbo: 254,438		x			x
English	Mainichi Daily+EIRB[b]: 22,927	x	x			
	Xinhua+others: 324,449		x			
2000–01						
Chinese	UDN[a]+others: 901,446			x	x	x
Japanese	Mainichi+Yomiuri: 858,400			x	x	x
Korean	Hankookilbo+Chosunilbo: 220,374			x	x	x
English	Xinhua+others: 259,050			x		

[a]UDN: United Daily News
[b]EIRB: Taiwan News and China Times English News

topic keywords such as descriptors that are often assigned to bibliographic records of scientific papers. Today various types of documents are exploited for current research on IR or related areas, but the test collections using such newspaper articles still provide IR researchers a sound experimental setting for examination of fundamental techniques that underlie more complicated searches.

2.3.3 CJKE Test Collection Construction

Test collections incorporating the CJKE documents were constructed according to a traditional pooling method explored by TREC. In general, a test collection consists of three components: a document set, topic set (set of search requests), and answer set (result of relevance judgments). By employing the answer set, metrics for evaluating IR results such as precision or recall can be computed. When calculating the recall, it is required to determine all relevant documents included in the document set, which is typically impossible for large-scale document sets. Therefore, the pooling method was developed for using such large sets. Figure 2.1 shows an operational model for IR evaluation based on the pooling method.

First, a document set such as that shown in Table 2.1 is sent to participants in the tasks for implementing into their own IR systems. Then, task organizers deliver a topic set to participants and ask them to submit search results by the designated day.

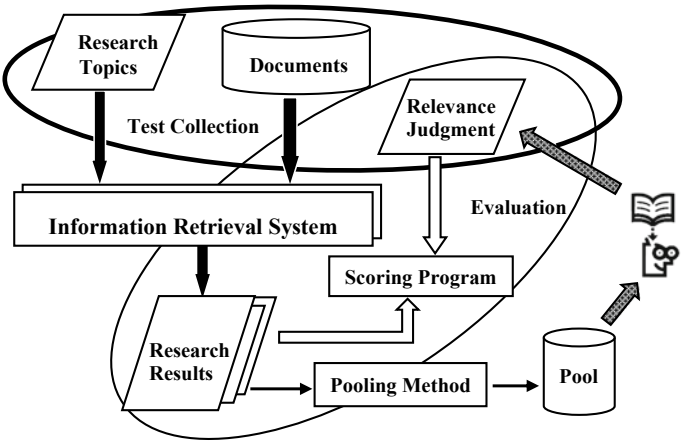

Fig. 2.1 Construction of test collection and evaluation

Under management of the task organizers, the degree of relevance is judged for each pair of a topic and a document included in the search results, by which an answer set is obtained. Finally, the search performance of the participating IR system is scored based on the answer set. By checking the scores, the advantages or disadvantages of IR theories or techniques are clarified. Because the relevance judgment is completed for pooled documents that are extracted from the search results that participants submitted, and not for the entire set of documents, this procedure for creating the answer set is termed the pooling method, which is an efficient means for constructing a large-scale test collection. Strictly speaking, scores of some evaluation metrics obtained from this procedure are only approximations because the entire set is not examined. However, a comparison of search effectiveness between the IR systems or models within the test collection is sufficiently feasible.

The organizers of the NTCIR CLIR tasks consisted of IR researchers in Taiwan, Japan, and Korea who collaboratively worked in designing the research tasks, creating the topics, managing the relevance judgment process, and evaluating the participating IR systems. The authors of this paper were members of the organizer group.

In our experience, it was difficult to create topics that were effective for measuring CLIR performance between the CJKE languages compared to a case of simple monolingual IR. The typical procedure for topic creation in the NTCIR CLIR tasks was as follows:

1. Topic candidates were created in Taiwan, Japan, and Korea, respectively, and were translated into English.
2. The English candidates were again translated into Chinese, Japanese, and Korean as necessary, and the task organizers preliminarily examined whether or not relevant documents were sufficiently included in the C, J, K, and E document sets.
3. Final topics were selected based on the preliminary examination result.

This complicated procedure was adopted for using topics commonly on all document sets in the CJKE languages, by which comparisons of search performance between bilingual searches of CJKE (e.g., between C to J and J to C searches that are related to processing of Chinese and Japanese texts) became easier.

Search topics created for NTCIR CLIR tasks can be approximately classified into two types: 1) event-based topics and 2) general concept-based topics. Event-based topics typically contain one or more proper nouns of a social event, geographic location, or person. An example is "Find reports on the G8 Okinawa Summit 2000" (ID 005 in NTCIR-5). If a CLIR system cannot find any corresponding translation of the proper noun during its process then the search performance is expected to be low. This is generally termed an *out-of-vocabulary (OOV)* problem.

Meanwhile, it may be relatively easier to find translations of a general concept, but CLIR systems often need to disambiguate translation candidates for the concept. For example, a correct translation in the context of the search topic often has to be selected from many terms listed in a bilingual dictionary (see Sect. 2.2.2). An instance of the general concept-based topics is "Find documents describing disasters thought to be caused by abnormal weather" (ID 044 in NTCIR-5). Even though "weather" has a relatively definite meaning, many translations are actually enumerated in an E to J dictionary and selection of final translations substantially affects CLIR performance. The task organizers considered a careful balance of the two topic types for allowing researchers to develop more effective systems. During the topic creation process, approximately 50 topics were included in each of the test collections for NTCIR-3 to -6, respectively.

Needless to say, jobs for pooling documents also are not easy. If the pool (i.e., a document set to be checked during a process of relevance judgment) is too large then it is impossible for an assessor to maintain consistent judgment for all documents. For avoiding this problem, the document pool size has to be appropriately adjusted when extracting top-ranked documents from the search results of each participant. This is a special matter of so-called pooling depth (Kuriyama et al. 2002).

Also, a system of relevance judgments developed by *National Institute of Informatics (NII)*, of which name was NACSIS at the time, was used for providing the assessors with a comfortable human-machine interface for the judgment task, which contributed to enhancing consistency and reliability of the judgment results. A windows-based assessment system created in Taiwan for the special purpose is explained by Chen (2002).

2.3.4 IR System Evaluation

During the process of relevance judgment, the assessors evaluated each document using a four-point scale: 1) highly relevant, 2) relevant, 3) partially relevant, and 4) irrelevant. The IR research field has a long history of studying relevance concepts and operational assessment of them. A multi-grade assessment based on the four-

point scale was adopted in ad hoc IR tasks beginning with NTCIR-1 after carefully examining discussions in the literature of relevance.

However, evaluation metrics based on a multi-grade assessment such as the *Normalized Discounted Cumulative Gain (nDCG)* were not yet popular at the time of NTCIR-1 to -6 and the main indicator for evaluating IR systems was *Mean Average Precision (MAP)*.[1] For calculating the average precision, the four-point scale has to be reduced to a binary scale. When "highly relevant" and "relevant" were considered to be relevant and the others to be irrelevant, it was specifically termed "rigid" relevance in the NTCIR CLIR tasks. If "partially relevant" was included in the relevant category then "relaxed" relevance was used. Therefore, in NTCIR CLIR tasks, two MAP scores were typically computed for a single search result based on rigid and relaxed relevance, respectively. Sakai (2020) summarizes evaluation metrics and methods in the overall NTCIR project.

2.4 CLIR Techniques in NTCIR

This section briefly summarizes typical techniques used in CLIR tasks from NTCIR-3 to -6. For knowing details of the techniques and systems, overviews of each task that were published at NTCIR conferences are helpful (Chen et al. 2002; Kishida et al. 2004, 2005, 2007). Lists of research groups participating in each task are also included in the overviews.

2.4.1 Monolingual Information Retrieval Techniques

IR systems of groups participating in NTCIR CLIR tasks typically have two independent components for 1) monolingual IR and 2) translation as explained in Sect. 2.2.1. Because computer processing of Chinese, Japanese, and Korean textual data had not yet been sufficiently developed at the time, NTCIR CLIR tasks also contributed to obtaining useful knowledge regarding CJK text processing for monolingual IR (or single language IR: SLIR). The resulting SLIR performance improvement can be considered as an achievement in the NTCIR CLIR tasks.

Particularly, sentences or phrases in the CJK texts have no explicit word boundary, which is a characteristic that is different from that of English texts (note that Korean texts include white spaces as a delimiter between phrasal units). To construct index files in SLIR systems for these languages, either

1. Word-based indexing, or
2. Overlapping character bigrams (i.e., n-grams when $n = 2$)

[1] Only in NTCIR-6, nDCG was used for evaluating CLIR performance as a trial.

were typically used in the NTCIR CLIR tasks. For word-based indexing, some groups employed morphological analyzers, whereas index terms were identified from texts by simply matching with entries of machine-readable dictionaries in some other systems.

Extracting character bigrams from texts are a characteristic during the indexing process of East Asian languages. Assume that a Japanese sentence is "ABCDE" where A, B, C, D, and E are a Japanese character, respectively. In the case of overlapping character bigrams, "AB," "BC," "CD," and "DE" are automatically selected as index terms. This was known as an effective method for processing texts that were represented by ideograms. Although character unigrams (i.e., n-grams when $n = 1$) are extracted from the target text in the current Internet search engines or some online public access catalog (OPAC) systems, $n = 2$ was used in NTCIR CLIR tasks.

By utilizing an index file constructed according to an indexing method, documents have to be ranked by the degree of relevance to each search query. The relevance degree is operationally estimated in the system based on a retrieval model. In NTCIR CLIR tasks, participant groups typically adopted some standard and well-known models such as the vector space mode (VSM), Okapi BM25, LM, INQUERY, PIRCS, or logistic regression model. In addition, *query expansion (QE)* by PRF or techniques using external resources (e.g., statistical thesauri based on term co-occurrence statistics or web pages) were incorporated for enhanced search performance. The retrieval models and QE techniques were originally developed in the USA or Europe mainly for English IR. NTCIR CLIR tasks provided good opportunities for systematically confirming their effectiveness for IR of CJK languages.

2.4.2 Bilingual Information Retrieval (BLIR) Techniques

Section 2.2 reviewed typical CLIR techniques, which were also utilized in NTCIR CLIR tasks. Dictionaries and MT software that were employed by participants in the NTCIR-4 CLIR task were extensively enumerated in Kishida et al. (2004).

Specifically, important problems to be solved for translation in CLIR among CJKE were as follows.

1. Query translation versus document translation: Most participating groups adopted a means of translating search topics (queries), whereas some explored "pseudo" document translation in which terms in target documents were simply replaced with equivalents in another language using a bilingual dictionary (i.e., not MT). Additionally, search performance may be improved by combining search results from both the query and document translations because it is possible that the probability of successful matching of terms between a topic and a relevant document increases. This technique was attempted by one group.
2. Pivot language approach: English was typically used as a pivot language for CLIR among CJK, whereas one group attempted bilingual searching via Japanese.

Selection of the pivot language depends on translation resources such as MT software.

3. OOV problem: As previously mentioned, when a term representing an important concept in a search topic is not included in the dictionaries for translation, search performance largely degrades. Some search topics in the NTCIR CLIR tasks contained names related to current events or affairs and they were not often covered if using only a standard bilingual dictionary (see Sect. 2.3.3). For solving this problem, some groups attempted to extract translations from web pages for the unknown term.
4. Automatic transliteration: In general, when a word in a foreign language is imported, transliteration is often used without semantically representing the word in its own language. For example, an English word "hotel" is transliterated into three *Katakana* characters corresponding phonetically to "ho," "te," and "ru" in Japanese. Although popular *Katakana* words are listed in standard bilingual dictionaries, an OOV problem occurs if this is not the case. At this time, an English word may be automatically converted into a *Katakana* word (and vice versa) via heuristic rules phonetically measuring the similarity between them (Fujii and Ishikawa 2001). This type of automatic transliteration was explored in the NTCIR CLIR tasks.
5. Conversion of *Kanji* character codes: An idea similar to automatic transliteration is automatic conversion of *Kanji* characters between Chinese and Japanese. In the NTCIR CLIR tasks, one group attempted to convert traditional Chinese characters encoded by the BIG5 character code into Japanese characters represented by Extended Unix Code-Japanese (EUC-JP).
6. Term disambiguation (or WSD): A typical method for term disambiguation was to use statistical information of term co-occurrences in the set of target documents. In addition, many CLIR systems incorporated a PRF process, which had an effect of increasing the rank of documents that included a combination of correct translations (see Sect. 2.2.2). Both methods do not require any external resource. In contrast, some external resources such as web pages or parallel corpora were also applied for term disambiguation by some groups. For example, one system attempted to select final query terms based on web pages that were extracted from a web category to which the search topic corresponded.

As a technique for improving CLIR performance, pre-translation PRF was explored in the NTCIR CLIR tasks. That is, if a corpus in the source language of an original query is available as an external resource, then a PRF operation on the external resource may result in a set of more useful query terms, termed the pre-translation PRF. After obtaining a "richer" representation of the original query in the source language using it, standard CLIR is executed based on the modified query. More common was to include PRF in the form of post-translation PRF after the retrieval process proper. A combination of pre- and post-translation PRF was used in some groups in the NTCIR CLIR tasks.

In addition, participants in the NTCIR CLIR tasks attempted to address other various challenges such as document re-ranking, QE via a statistical thesaurus, trans-

Table 2.3 Best MAP scores of SLIR and BLIR in NTCIR-6: Rigid relevance, DESC field[a]

	Documents (X)		
Search topics	Chinese	Japanese	Korean
Monolingual (baseline)	0.313 (100%)	0.325 (100%)	0.454 (100%)
BLIR			
Chinese (C to X search)	–	0.312 (95.8%)	N/A
Japanese (J to X search)	0.078 (24.7%)	–	0.287 (63.2%)
Korea (K to X search)	0.102 (32.6%)	0.267 (82.1%)	–
English (E to X search)	0.191 (61.0%)	0.307 (94.4%)	0.292 (64.3%)

[a] A short sentence describing each search topic was included in a <DESC> element of an XML file of the topics. The sentence was used as a query for executing searches in this table

lation probability estimation, the use of an ontology to enhance the effectiveness of mono- or cross-lingual IR. Although similar research efforts had already been completed in TREC or CLEF (Cross-Language Evaluation Forum, at that time), a special aspect of the NTCIR CLIR tasks was the larger differences in language types between English and CJK. For example, nobody would deny that "linguistic distance" between English and Japanese is greater than that between English and Swedish. The special characteristics of CJK as languages may have contributed to unique modification or refinement of CLIR techniques (e.g., automatic transliteration).

It is difficult to concisely present an overview of search performance attained by CLIR systems participating in CLIR tasks from NTCIR-3 to -6. Only the best performance of the SLIR and BLIR subtasks in NTCIR-6 is shown in Table 2.3 (Kishida et al. 2007), which provides the best MAP scores based on "rigid" relevance by each language combination. When comparing the MAP scores between monolingual searching (SLIR) and BLIR, it appears that BLIR to Japanese documents was more successful than to other languages because the percentages were 95.8% for C, 82.1% for K and 94.4% for E search topics. However, the percentage highly depends on the system performance of the research group participating in the task at the time; thus, Table 2.3 does not indicate any research finding based on a scientific examination. This table is only an example for superficially understanding an aspect of NTCIR CLIR tasks. Readers that are interested in the search runs of Table 2.3 can refer to Kishida et al. (2007) for more detail.

2.4.3 Multilingual Information Retrieval (MLIR) Techniques

Two types of MLIR strategies are most commonly used:

- (A) All documents and the query are translated into a single language (e.g., English), and then monolingual IR is executed thereafter and
- (B) BLIR is repeated for all pairs of document language and query language, and then all search results are finally merged into a single ranked document list.

Fewer research groups participated in MLIR subtasks compared to those in SLIR and BLIR subtasks, and most adopted the type B strategy. In the strategy, an important choice is how search results (actually, individual ranked lists by language pairs) are merged, which can be considered as a type of data fusion problem. The merging operation is also important for applications other than MLIR.

Typical merging methods in NTCIR CLIR tasks are as follows.

1. Round-robin merging: Documents are repeatedly selected from the top of each ranked list in a sequence.
2. Raw score merging: All documents are merged and re-ranked according to document scores calculated by an IR model.
3. Normalized score merging: Document scores that are calculated by an IR model are normalized before the documents are merged and re-ranked.

When applying these methods, there are some difficulties. For example, if the number of relevant documents included in the C, J, K, and E components is significantly different, then the difference makes the MLIR more difficult. In this situation, an "absolute" relevance probability that is effective over all languages may have to be estimated for each document to achieve better performance. Braschler (2004) discusses the other difficulties of MLIR. Actually, MAP scores of MLIR were typically lower than those of SLIR and BLIR in the NTCIR CLIR tasks.

2.5 Concluding Remarks

Research activity for exploring the cross-lingual ad hoc IR of newspaper articles in the NTCIR project ended at the CLIR task in NTCIR-6, for which the conference was held in May of 2007. Thereafter, during the 2010s, the Internet search engine performance remarkably improved, more easily allowing one to search Chinese, Japanese, and Korean documents in situations of monolingual IR. In addition, several excellent tools or resources for language processing have become available. Specifically, new technologies such as statistical machine translation or neural machine translation have drastically enhanced the effectiveness of MT.

The current state of monolingual IR and language processing has largely changed from the time of the NTCIR CLIR tasks. Experimental findings that were obtained from the tasks have contributed to such technological advances and aided researchers

in developing a more sophisticated CLIR system based on the current technologies of monolingual IR and language processing. In addition, the authors believe that experience of constructing test collections that consist of comparable corpora in NTCIR CLIR tasks is useful for further development of IR theories and techniques in multilingual environments.

Acknowledgements Many researchers in Taiwan, Japan, and Korea worked collaboratively as organizers in managing NTCIR CLIR tasks as follows: Hsin-Hsi Chen, Koji Eguchi, Noriko Kando, Kazuko Kuriyama, Hyeon Kim, Sukhoon Lee, and Sung Hyon Myaeng (as well as the authors of this paper). Additionally, in NTCIR-1 and -2, Toshihiko Nozue, Souichiro Hidaka, Hiroyuki Kato, and Masaharu Yoshioka also joined to organize the IR tasks. This paper attempted to summarize some aspects of valuable activities and efforts by the organizers and by all participants in the research tasks.

References

Ballesteros L, Croft WB (1997) Phrasal translation and query expansion techniques for cross-language information retrieval. In: Proceedings of the 20th annual international ACM SIGIR conference on research and development in information retrieval, pp 84–91

Braschler M (2004) Combination approaches for multilingual text retrieval. Inf Retr 7(1/2):183–204

Chen KH (2002) Evaluating Chinese text retrieval with multilingual queries. Knowl Organ 29(3/4):156–170

Chen KH, Chen HH, Kando N, Kuriyama K, Lee S, Myaeng SH, Kishida K, Eguchi K, Kim H (2002) Overview of CLIR task at the third NTCIR workshop. In: Proceedings of the Third NTCIR workshop on research in information retrieval, automatic text summarization and question answering

Fujii A, Ishikawa T (2001) Japanese/English cross-language information retrieval: Exploration of query translation and transliteration. Comput Human 35(4):389–420

Grefenstette G (1998) Cross-language information retrieval. Springer, Berlin

Kando N, Kuriyama K, Nozue T, Eguchi K, Kato H, Hidaka S (1999) Overview of IR tasks. In: Proceedings of the First NTCIR workshop on research in Japanese text retrieval and term recognition, pp 11–44

Kando N, Kuriyama K, Yoshioka M (2001) Overview of Japanese and English information retrieval tasks (JEIR) at the second NTCIR workshop. In: Proceedings of the second NTCIR workshop on research in Chinese and Japanese text retrieval and text summarization

Kishida K (2005) Technical issues of cross-language information retrieval: a review. Inf Process Manag 41(3):433–455

Kishida K, Chen KH, Lee S, Kuriyama K, Kando N, Chen HH, Myaeng SH, Eguchi K (2004) Overview of CLIR task at the fourth NTCIR workshop. In: Proceedings of the fourth NTCIR workshop on research in information access technologies: information retrieval, question answering and summarization

Kishida K, Chen KH, Lee S, Kuriyama K, Kando N, Chen HH, Myaeng SH (2005) Overview of CLIR task at the fifth NTCIR workshop. In: Proceedings of the fifth NTCIR workshop meeting on evaluation of information access technologies: information retrieval, question answering and cross-lingual information access

Kishida K, Chen KH, Lee S, Kuriyama K, Kando N, Chen HH (2007) Overview of CLIR task at the sixth NTCIR workshop. In: Proceedings of the 6th NTCIR workshop meeting on evaluation of information access technologies: information retrieval, question answering and cross-lingual information access

Kuriyama K, Kando N, Nozue T, Eguchi K (2002) Pooling for a large-scale test collection: an analysis of the search results from the first NTCIR workshop. Inf Retr 5(1):41–59
Nie JY (2010) Cross-language information retrieval. Morgan & Claypool Publishers
Oard DW, Diekema AR (1998) Cross-language information retrieval. Ann Rev Inf Sci Technol 33:223–256
Peters C, Braschler M, Clough P (eds) (2012) Multilingual information retrieval. Springer, Berlin
Sakai T (2020) Graded relevance. In: Sakai T, Oard DW, Kando N (eds) Evaluating information retrieval and access tasks. Springer, Singapore. The Information Retrieval Series, (in this book)
Voorhees E, Harman DK (eds) (2005) TREC: experiment and evaluation in information retrieval. MIT Press
Xu J, Weischedel R, Nguyen C (2001) Evaluating a probabilistic model for cross-lingual information retrieval. In: Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval, pp 105–110