NTCIR-4 CLIR Experiments at Oki

Tetsuji Nakagawa and **Mihoko Kitamura**
nakagawa378@oki.com, kitamura655@oki.com
Corporate Research and Development Center
Oki Electric Industry Co., Ltd.
2−5−7 Honmachi, Chuo-ku, Osaka 541-0053, Japan

Abstract

We participated in SLIR, BLIR(PLIR) and MLIR subtasks at the NTCIR-4 CLIR task. Our IR system can handle queries and documents in Chinese, English and Japanese. The system utilizes multiple language resources (bilingual dictionaries, parallel corpora and machine translation systems) for query translation. We adopted the pivot language approach for C-J and J-C search using English as a pivot language. We submitted formal runs for 12 subtasks, and confirmed that the combination of language resources makes performance of BLIR high and that the pivot language approach is promising.

Keywords: *language model, query translation, multilingual information retrieval, pivot language.*

1 Introduction

We developed a cross-lingual IR system and participated in the NTCIR CLIR task for the first time. Our goal is to develop a flexible CLIR system which can handle many languages. The system can conduct C-C, E-E, J-J SLIR, C-E, C-J, E-C, E-J, J-C, J-E BLIR and C-CEJ, E-CEJ, J-CEJ MLIR. We use bilingual dictionaries, parallel corpora and machine translation systems for query translation. The pivot language method is used for C-J and J-C BLIR because of unavailability of language resources.

This paper is organized as follows: Section 2 describes our IR system. Section 3 discusses experimental results and Section 4 concludes.

2 System Description

The system uses word-based indexing for Chinese, English and Japanese. Language models are used for document scoring, and the pseudo-relevance feedback is used for query expansion. In bilingual IR, we use the query translation approach, and the cross-lingual pseudo-relevance feedback method is used for query translation. In multilingual IR, each result of SLIR and BLIR is merged using the normalized-score method. We explain these methods in the following subsections. More detailed information is described in Appendix A.

2.1 Keyword Extraction

There are mainly two approaches for keyword extraction from queries and documents written in Chinese or Japanese — the n-gram-based approach and the word-based approach. The n-gram-based approach uses character n-grams for indexing and needs no word segmenters, but the size of the index is large. Our system adopts the word-based approach which uses words for indexing. The statistical Chinese word segmenter we developed [9] and the Japanese morphological analyzer ChaSen [8] are used respectively for Chinese and Japanese keyword extraction. Stop words are not removed for these languages. In English keyword extraction, the Porter stemmer is used and stop words (429 words) are removed.

2.2 Document Scoring

The language models [11] are used for document scoring. Given a query q and a document d, the method uses a joint probability that q and d are generated as the relevance between q and d, and the retrieval status value RSV is calculated as follows:

$$RSV \quad = \quad \log P(d) + \log P(q|d). \qquad (1)$$

We use the following equation ($score_4$ model described in [4]) for the calculation of the probability:

$$RSV = \log \sum_s tf(s,d) + \sum_{t \in q \cap d} tf(t,q) \ \log \left\{ 1 + \frac{\lambda \, tf(t,d) \sum_s df(s)}{(1-\lambda) \, df(t) \sum_s tf(s,d)} \right\}, \qquad (2)$$

where $tf(t,q)$ is the frequency of the term t in the query q, $tf(t,d)$ is the frequency of the term t in the

document d, $df(t)$ is the number of documents containing the term t, $q \cap d$ is a set of terms which appear in both q and d, and λ is a smoothing parameter. The calculation can be done using an inverted file, and we use Generic Engine for Transposable Association (GETA) [5] for indexing and scoring. We set the value of the smoothing parameter λ to 0.25.

2.3 Query Expansion

We use the pseudo-relevance feedback (PRF) method to expand queries. Given a query, the method retrieves the top N documents $d_1, \cdots, d_N$. Each term t in the documents are ranked by a certain score $L(t)$ and the top M terms are added to the initial query. We use the ratio method [10] for the scoring, and the score is defined as follows:

$$L(t) = \sum_{j=1}^{N} \log\left(\frac{P(t|d_j)}{P(t)}\right), \\ = \sum_{j=1}^{N} \log\left(\frac{tf(t,d_j)}{\sum_s tf(s,d_j)} \frac{\sum_s \sum_d tf(s,d)}{\sum_d tf(t,d)}\right). \quad (3)$$

We set the both values of N and M to 10.

2.4 Cross-Lingual IR

We use the query translation approach for Cross-lingual IR. In the query translation approach, a given query is first translated to the language in which documents to be retrieved are written, and then monolingual IR is performed.

We prepared the following language resources for query translation:

Bilingual Dictionary

- *EDICT* — Japanese-English dictionary (106,843 words) [1]
- *CEDICT* — Chinese-English dictionary (26,404 words) [3]

Parallel Corpus

- *Japanese-English News Article Alignment Data* — Japanese and English news articles with sentence-level alignment (30,000 pairs of sentences) [13]

Machine Translation System

- *YakushiteNet* — English-Japanese and Japanese-English Machine Translation System [6]

The machine translation system can be directly used for query translation, but the bilingual dictionaries and the parallel corpus cannot. We use the cross-lingual PRF (CLPRF) method [2, 12] for query translation with the bilingual dictionaries and the parallel corpus. The CLPRF method is similar to the PRF method, but it uses parallel data. Given a query, the top N' documents are retrieved from the parallel data written in a source language, then M' terms are extracted from the corresponding parallel data written in a target language and used as a translated query. The method has two advantages. First, the method can be applied to both bilingual dictionaries and parallel corpora. Second, the method can handle both directions of translation regardless of the direction of the bilingual dictionary which is used; e.g., a Japanese-English dictionary can be used for both Japanese-to-English and English-to-Japanese translation.

Since we could not find any language resources for the language pair of Chinese and Japanese, we use the pivot language method using English as a pivot language in C-J and J-C search. In the method, queries written in Chinese (Japanese) are translated to English queries first, and then the English queries are translated to Japanese (Chinese) queries.

Figure 1 shows the usage of the language resources in each query translation. We use monolingual PRF for queries before and after the query translation (see Appendix A for detailed description).

2.5 Multilingual IR

Several methods for MLIR are studied [7] and they are classified into two approaches; the centralized approach and the distributed approach. In the centralized approach, one unified index is created for documents in different languages, and retrieval is performed at one time. In the distributed approach, documents written in different languages are indexed and retrieved independently for each language, and the retrieved results are merged later. There are some merging methods for the distributed approach. The round-robin method interleaves the retrieved documents of different languages by assuming that the significance of the ranking in each language is equal. The raw-score method merges the retrieved documents using raw values of RSV. The normalized-score method merges the retrieved documents using normalized values of RSV, and the normalized value of RSV for a language l's ith ranked document, RSV_i^l, is calculated as follows:

$$NormalizedRSV_i^l = \frac{RSV_i^l - \min_j\{RSV_j^l\}}{\max_j\{RSV_j^l\} - \min_j\{RSV_j^l\}}. \quad (4)$$

In preliminary experiments, the normalized-score method showed high performance and we use it in the system.

3 Experiments

In this section, we show the results of preliminary experiments and the NTCIR-4 formal run.

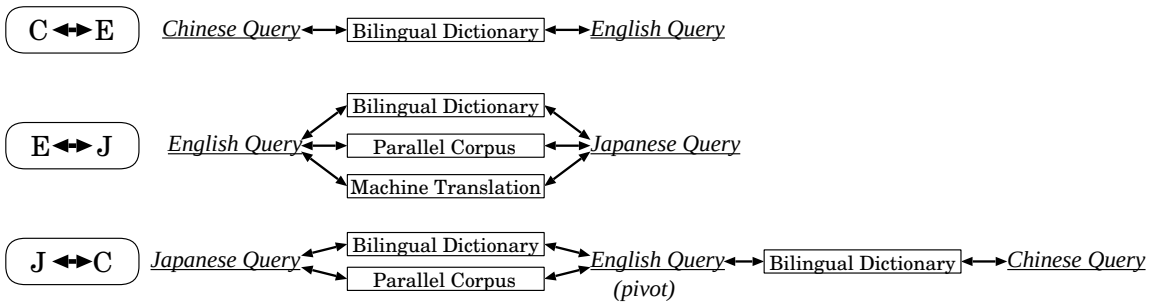

Figure 1. Query Translation in BLIR

3.1 Preliminary and Post-Submission Experiments

We conducted some preliminary experiments to tune parameters using NTCIR-3 CLIR data sets. We also conducted the same experiments using NTCIR-4 data sets. In the experiments, the DESC field of topics is used as a query and the mean average precision (MAP) with the Relaxed relevance criterion is used for evaluation.

3.1.1 Parameters of PRF

When PRF is used, we must determine the value N(the number of documents to be retrieved) and M(the number of terms to be extracted). We conducted experiments for different values of N and M. Table 1 shows the results. The MAP values for NTCIR-3 data and NTCIR-4 data are shown in the left and the right of the slashes respectively. We set the both values of N and M to 10 in the formal run because the parameter setting showed the highest performance in E-E search on the NTCIR-3 data. The values worked well in C-C search on the NTCIR-4 data. However, the performance was sensitive to the parameters and some room is left for improvement.

3.1.2 Resources for BLIR

To examine a contribution of each translation resource to performance of BLIR, we conducted E-J and J-E search experiments using different combinations of the resources. Table 2 shows the results. High performance was obtained by using the parallel corpus or the machine translation. The bilingual dictionary is a Japanese-English dictionary, but the E-J search using the dictionary showed enough performance compared to J-E search. Using the all resources, the highest performance was obtained.

3.1.3 Methods for MLIR

We conducted MLIR experiments to examine the performance for different merging strategies. Table 3 shows the results. On the NTCIR-3 data, high performance was obtained by using the normalized-score method. Since performance of the round-robin method was relatively low and the performance of the centralized method varied for different subtasks, we employed the normalized-score method. The method was also effective on the NTCIR-4 data.

3.1.4 Indexing Methods and PRF

To examine performance of different indexing methods and an effectiveness of PRF, we conducted some SLIR experiments on the NTCIR-4 data. We examined three indexing methods; the word-based indexing method, the word-based indexing method with stop word removal, and the n-gram-based indexing method using character unigrams and (overlapping) bigrams. Table 4 shows the results. PRF improved the performance in all the cases. In the formal run, we used the word-based indexing method for Chinese, English and Japanese. We removed stop words for only English, because when used with PRF, stop word removal for Chinese and Japanese decreased performance in our preliminary experiments. However, the performance of J-J search was low in the NTCIR-4 data when stop words are not removed.

3.2 Results of Formal Run

The MAP values of formal runs are shown in Table 5 and Figure 2. Our system had high performance in the E-J BLIR and three MLIR subtasks. The MAP values for various queries and documents are summarized in Table 6 with the rates when the MAP values of SLIR are regarded as 1.0. The performance of E-J and J-E subtasks is satisfactory, which is greater than 80% of the SLIR's performance. However, the performance of C-E and E-C subtasks is low. The reason seems to be the shortage of language resources. C-E and E-C subtasks are performed using only the small bilingual dictionary (CEDICT), and the results of post-submission experiments of E-J and J-E subtasks using only the dictionary (EDICT) was also low (see Table

2). The C-J and J-C subtasks are performed using the pivot language method and low performance was obtained, but this seems to be mainly caused by the above issue and not by the use of the pivot language. The performance of the J-C (J→E→C) subtask is comparable to that of the E-C subtask, therefore we may obtain higher performance with the pivot language if we improve C-E and E-C search.

The MAP values for different runs are shown in Figure 3. The NARR-field has positive effect in many cases except the C-E and C-J subtasks. The NARR-field has positive effect in the C-C subtask and C-J search is pivot-based search utilizing C-E search, therefore the C-E search may have a problem but we need more analysis.

4 Conclusion and Future Work

We developed the CLIR system which handles Chinese, English and Japanese, and participated in the SLIR, BLIR(PLIR) and MLIR subtasks. The system utilizes the bilingual dictionaries, the parallel corpus and the machine translation system for BLIR, and also uses the pivot language method in C-J and J-C search. The performance of E-J and J-E search was high and the pivot language method is promising. However, the performance of C-E and E-C search was low apparently because of the shortage of language resources, and our future work will include automatic acquisition and utilization of translation knowledge for BLIR to overcome the problem.

Acknowledgements

We used CEDICT [3], ChaSen [8], EDICT [1], GETA [5] and Japanese-English News Article Alignment Data [13] in the research. We sincerely thank all the people who made these software and data.

This work was supported by a grant from the National Institute of Information and Communications Technology (NICT) of Japan.

References

[1] J. Breen. The EDICT Project, 2003. `http://www.csse.monash.edu.au/~jwb/edict.html`.

[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 Fifteenth International Joint Conference on Artificial Intelligence*, pages 708–715, 1997.

[3] P. Denisowski. CEDICT: Chinese-English Dictionary, 2003. `http://www.mandarintools.com/cedict.html`.

[4] D. Hiemstra. *Using Language Models for Information Retrieval*. Ph.D. Thesis, Centre for Telematics and Information Technology, University of Twente, 2001.

[5] IPA. Generic Engine for Transposable Association (GETA), 2003. `http://geta.ex.nii.ac.jp/e/`.

[6] M. Kitamura and T. Murata. Practical Machine Translation System allowing Complex Patterns. In *Proceedings of the Ninth Machine Translation Summit*, pages 232–239, 2003. `http://www.yakushite.net/`.

[7] W.-C. Lin and H.-H. Chen. NTU at NTCIR3 MLIR Task. In *Working Notes of the Third NTCIR Workshop Meeting, Part II: Cross Lingual Information Retrieval Task*, pages 101–106, 2002.

[8] Y. Matsumoto, A. Kitauchi, T. Yamashita, Y. Hirano, H. Matsuda, K. Takaoka, and M. Asahara. *Morphological Analysis System ChaSen version 2.2.8 Manual*. Nara Institute of Science and Technology, 2001.

[9] T. Nakagawa. Chinese and Japanese Word Segmentation Using Word-Level and Character-Level Information. In *Proceedings of the 20th International Conference on Computational Linguistics*, 2004. (to appear).

[10] J. M. Ponte. *A Language Modeling Approach to Information Retrieval*. Ph.D. Thesis, Graduate School of the University of Massachusetts Amherst, 1998.

[11] J. M. Ponte and W. B. Croft. A Language Modeling Approach to Information Retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 275–281, 1998.

[12] M. Rogati and Y. Yang. Cross-Lingual Pseudo-Relevance Feedback Using a Comparable Corpus. In *CLEF 2001*, pages 151–157, 2001.

[13] M. Utiyama and H. Isahara. Reliable Measures for Aligning Japanese-English News Articles and Sentences. In *Proceedings of the 41th Annual Meeting of the Association for Computational Linguistics*, pages 72–79, 2003. `http://www2.crl.go.jp/jt/a132/members/mutiyama/jea/`.

Subtask	$M \backslash N$	5	10	20
C-C	5	0.3112 / 0.2533	0.3284 / 0.2671	0.3380 / 0.2865
	10	0.3333 / 0.2617	0.3519 / 0.2854	0.3655 / 0.2845
	20	0.3233 / 0.2532	0.3479 / 0.2665	0.3479 / 0.2824
E-E	5	0.4551 / 0.3822	0.4593 / 0.4005	0.4577 / 0.3890
	10	0.4703 / 0.3983	0.4948 / 0.4145	0.4740 / 0.4148
	20	0.4370 / 0.4111	0.4689 / 0.4245	0.4426 / 0.4185
J-J	5	0.4067 / 0.4091	0.3902 / 0.4175	0.3889 / 0.4219
	10	0.3937 / 0.4162	0.3800 / 0.4295	0.3893 / 0.4400
	20	0.3973 / 0.4090	0.3908 / 0.4373	0.3988 / 0.4466

Table 1. MAP for Different Parameters of PRF (D-run, Relax, NTCIR-3 / NTCIR-4)

Resources	E-J	J-E
Bilingual Dictionary	0.1429 / 0.1524	0.1476 / 0.1901
Parallel Corpus	0.2740 / 0.3033	0.3489 / 0.2815
Machine Translation	0.2103 / 0.2968	0.3050 / 0.3203
Bilingual Dictionary + Parallel Corpus	– / 0.3251	– / 0.3163
Parallel Corpus + Machine Translation	– / 0.3509	– / 0.3406
Machine Translation + Bilingual Dictionary	– / 0.2844	– / 0.3031
Bilingual Dictionary + Parallel Corpus + Machine Translation	0.3141 / 0.3566	0.3655 / 0.3594

Table 2. MAP for Different Resources in BLIR (D-run, Relax, NTCIR-3 / NTCIR-4)

Method	C-CEJ	E-CEJ	J-CEJ
Centralized	0.1256 / 0.1185	0.1801 / 0.1986	0.1067 / 0.1737
Round-Robin	0.1401 / 0.1088	0.1390 / 0.1845	0.1376 / 0.1839
Raw-Score	0.1638 / 0.1057	0.1685 / 0.1841	0.1695 / 0.1965
Normalized-Score	0.1617 / 0.1175	0.1765 / 0.2093	0.1759 / 0.2093

Table 3. MAP for Different Methods in MLIR (D-run, Relax, NTCIR-3 / NTCIR-4)

	C-C	E-E	J-J
Word-based indexing	0.1969	0.3253	0.3513
Word-based indexing + PRF	0.2854	0.4165	0.4295
Word-based indexing + Stop Word Removal	0.2055	0.3351	0.3761
Word-based indexing + Stop Word Removal + PRF	0.2800	0.4145	0.4534
Character 1,2-gram indexing	0.2075	–	0.3647
Character 1,2-gram indexing + PRF	0.2533	–	0.4182

Table 4. MAP for Different Indexing Methods and PRF (D-run, Relax, NTCIR-4)

	D-run		DN-run		T-run		TC-run		TDNC-run	
	Relax	Rigid	Relax	Rigid	Relax	Rigid	Relax	Rigid	Relax	Rigid
C-C	0.2854	0.2274	0.3037	0.2476	0.2761	0.2312	0.2915	0.2425	0.3050	0.2556
E-C	0.0611	0.0481	0.0651	0.0532	0.0704	0.0505	0.0731	0.0562	0.0716	0.0567
J-C	0.0582	0.0404	0.0405	0.0281	0.0596	0.0386	0.0703	0.0589	0.0509	0.0342
C-E	0.1716	0.1265	0.0616	0.0372	0.1900	0.1502	0.1476	0.1105	0.0818	0.0486
E-E	0.4145	0.3286	0.4514	0.3679	0.4300	0.3361	0.4331	0.3466	0.4803	0.3881
J-E	0.3594	0.2813	0.4058	0.3200	0.3785	0.2934	0.3885	0.3034	0.3983	0.3139
C-J	0.1561	0.1088	0.0727	0.0565	0.1448	0.1016	0.1599	0.1100	0.0950	0.0689
E-J	0.3566	0.2674	0.4043	0.3099	0.3525	0.2735	0.3599	0.2735	0.4053	0.3210
J-J	0.4295	0.3082	0.4563	0.3343	0.4256	0.3162	0.4223	0.3131	0.4729	0.3499
C-CEJ	0.1175	0.0923	0.0634	0.0522	0.1165	0.0947	0.1078	0.0837	0.0773	0.0614
E-CEJ	0.2093	0.1588	0.2543	0.1986	0.2173	0.1704	0.2145	0.1640	0.2584	0.2043
J-CEJ	0.2093	0.1566	0.2287	0.1677	0.2127	0.1579	0.2189	0.1639	0.2383	0.1748

Table 5. MAP of Formal Runs

Query	Document			
	C	E	J	CEJ
C	0.2854 [1.00]	0.1716 [0.41]	0.1561 [0.36]	0.1175
E	0.0611 [0.21]	0.4145 [1.00]	0.3566 [0.83]	0.2093
J	0.0582 [0.20]	0.3594 [0.87]	0.4295 [1.00]	0.2093

Table 6. MAP of Formal Runs [Rate for SLIR] (D-run, Relax)

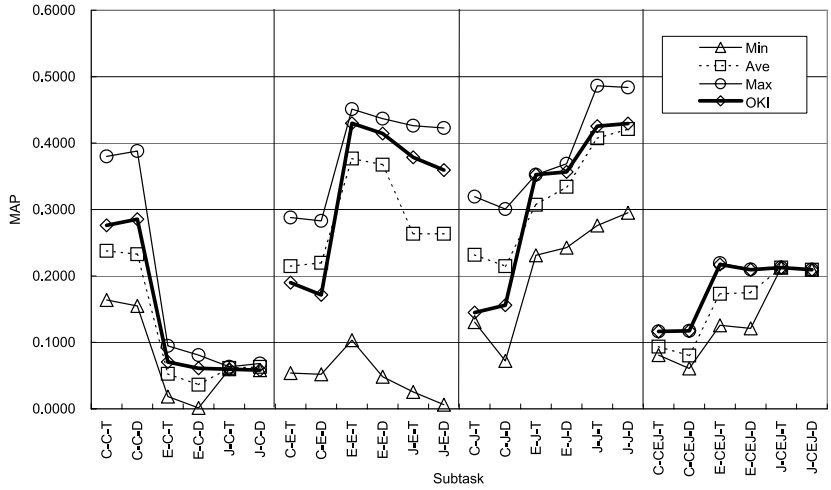

Figure 2. MAP of Formal Runs (D-run, Relax)

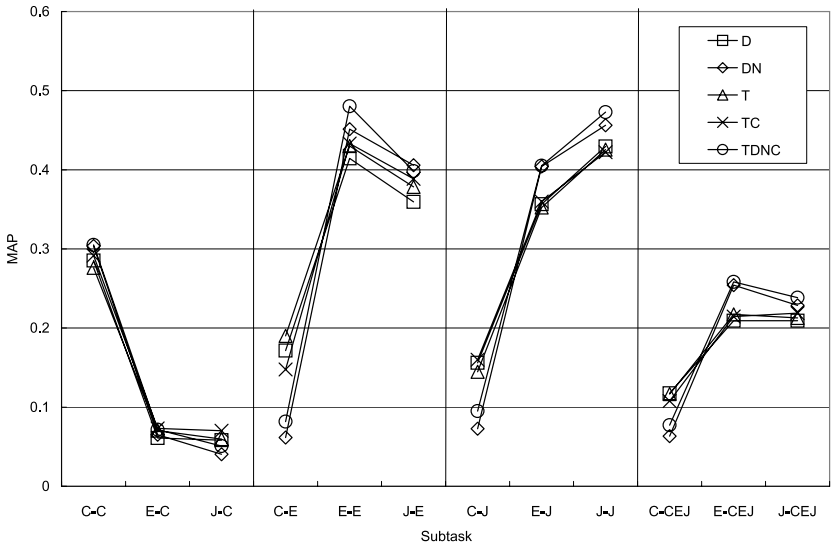

Figure 3. MAP for Different Runs (Relax)

A System Architecture

Our IR system uses the methods explained in Section 2, and this appendix describes in detail how the components are combined.

We define the following symbols and functions:

W:
A set of words

X:
A set of documents

$\mathcal{R}(D \subseteq X, q \subseteq W \times \mathbf{N}) \subseteq X \times \mathbf{R}$:
This function performs retrieval for the query q from the documents D, and returns top 1,000 documents with their scores.

$\mathcal{X}_{N,M}(D \subseteq X, q \subseteq W \times \mathbf{N}) \subseteq W \times \mathbf{N}$:
This function performs PRF for the query q using the documents D, and returns the expanded query.

$\mathcal{C}_{N',M'}(D^t \subseteq X, D^s \subseteq X, q \subseteq W \times \mathbf{N}) \subseteq W \times \mathbf{N}$:
This function performs CLPRF for the query q using the document D^s in a source language and the document D^t in a target language, and returns a translated query.

$\mathcal{M}(R_1, R_2, R_3) \subseteq X \times \mathbf{R}, (R_1, R_2, R_3 \subseteq X \times \mathbf{R})$:
This function merges the documents R_1, R_2, R_3 which are attached scores using the normalized-score method and returns top 1,000 documents.

$q_1 + q_2 \subseteq W \times \mathbf{N}, \ \ (q_1, q_2 \subseteq W \times \mathbf{N})$:
This function merges the query q_1 and q_2.

$q^C, q^E, q^J \subseteq W \times \mathbf{N}$:
The Chinese, English and Japanese queries represented as bags of words

$q^{E \to J}, q^{J \to E} \subseteq W \times \mathbf{N}$:
The Japanese query translated from the English query and the English query translated from the Japanese query by the machine translation system YakushiteNet

$D^C, D^E, D^J \subseteq X$:
The Chinese, English and Japanese documents to be retrieved

$R^x \subseteq X \times \mathbf{R}$:
The result of x subtask

$CEDICT^C, CEDICT^E \subseteq X$:
Chinese and English parts of the Chinese-English bilingual dictionary CEDICT

$EDICT^E, EDICT^J \subseteq X$:
English and Japanese parts of the Japanese-English bilingual dictionary EDICT

$NEWS^E, NEWS^J \subseteq X$:
English and Japanese parts of the parallel corpus

Note that PRF is a special case of CLPRF, i.e. $\mathcal{X}_{N,M}(D, q) = \mathcal{C}_{N,M}(D, D, q) + q$. We also define the following auxiliary functions:

$$
\begin{aligned}
\mathcal{F}^{C-E}(q) &\equiv \mathcal{C}_{30,30}(CEDICT^E, CEDICT^C, \mathcal{X}_{10,10}(D^C, q)),\\
\mathcal{F}^{E-C}(q) &\equiv \mathcal{C}_{30,30}(CEDICT^C, CEDICT^E, \mathcal{X}_{10,10}(D^E, q)),\\
\mathcal{F}^{E-J}(q) &\equiv \mathcal{C}_{30,30}(EDICT^J, EDICT^E, \mathcal{X}_{10,10}(D^E, q)) +\\
&\quad\ \mathcal{C}_{10,10}(NEWS^J, NEWS^E, \mathcal{X}_{10,10}(D^E, q)),\\
\mathcal{F}^{J-E}(q) &\equiv \mathcal{C}_{30,30}(EDICT^E, EDICT^J, \mathcal{X}_{10,10}(D^J, q)) +\\
&\quad\ \mathcal{C}_{10,10}(NEWS^E, NEWS^J, \mathcal{X}_{10,10}(D^J, q)).
\end{aligned}
$$

The result of each subtask is obtained as follows:

$$
\begin{aligned}
R^{C-C} &= \mathcal{R}(D^C, \mathcal{X}_{10,10}(D^C, q^C)),\\
R^{E-E} &= \mathcal{R}(D^E, \mathcal{X}_{10,10}(D^E, q^E)),\\
R^{J-J} &= \mathcal{R}(D^J, \mathcal{X}_{10,10}(D^J, q^J)),\\
R^{C-E} &= \mathcal{R}(D^E, \mathcal{X}_{10,10}(D^E, \mathcal{F}^{C-E}(q^C))),\\
R^{C-J} &= \mathcal{R}(D^J, \mathcal{X}_{10,10}(D^J, \mathcal{F}^{E-J}(\mathcal{F}^{C-E}(q^C)))),\\
R^{E-C} &= \mathcal{R}(D^C, \mathcal{X}_{10,10}(D^C, \mathcal{F}^{E-C}(q^E))),\\
R^{E-J} &= \mathcal{R}(D^J, \mathcal{X}_{10,10}(D^J, \mathcal{F}^{E-J}(q^E) + q^{E \to J})),\\
R^{J-C} &= \mathcal{R}(D^C, \mathcal{X}_{10,10}(D^C, \mathcal{F}^{E-C}(\mathcal{F}^{J-E}(q^J)))),\\
R^{J-E} &= \mathcal{R}(D^E, \mathcal{X}_{10,10}(D^E, \mathcal{F}^{J-E}(q^J) + q^{J \to E})),\\
R^{C-CEJ} &= \mathcal{M}(R^{C-C}, R^{C-E}, R^{C-J}),\\
R^{E-CEJ} &= \mathcal{M}(R^{E-C}, R^{E-E}, R^{E-J}),\\
R^{J-CEJ} &= \mathcal{M}(R^{J-C}, R^{J-E}, R^{J-J}).
\end{aligned}
$$