OASIS at NTCIR-6: On-line Query Translation for Chinese-Japanese Cross-Lingual Information Retrieval

Vitaly Klyuev
University of Aizu
Tsuruga, Ikki-Machi, Aizu-Wakamatsu, Fukushima, 965-8580, Japan
vkluev@u-aizu.ac.jp

Abstract

This paper reports results of Chinese – Japanese CLIR experiments using on-line query translation techniques. Approaches to employ English as a pilot language and to utilize several on-line translation systems are introduced. They were tested on NTCIR – 3, 4, 5, and 6 collections. Proposed procedures can be helpful under certain circumstances.

Keywords: *Cross-lingual information retrieval, Query translation.*

1 Introduction

Cross-lingual information retrieval (CLIR) deals with searching for information written in a language different from the language of the user's query. Several techniques and heuristics were proposed over past years to manage a key problem related to this type of retrieval: converting the CLIR task into the monolingual retrieval task [1, 3, 5, 6, 7]. The main directions can be characterized as follows. Texts are still considered as "bags of words". The logical outcome from this: Statistical methods dominate in the area. A large portion of research to develop new methods is oriented towards testing different heuristics. The common method is translation using dictionaries. The easiest way is to translate queries. The usual auxiliary tools are bilingual corpora. The typical method is the statistical approach utilizing training corpora to adjust the parameters of retrieval systems. Created techniques and methods work under certain circumstances. The most promising solution from our point of view is to use on-line translation systems.

Internationalization of the Internet has triggered development of translation systems freely available on the net. The quality of these systems has improved dramatically over past years. They are adequate for translating technical descriptions of goods, manuals (such as computer manuals) and instructions (such as instructions to install software). The language used in these texts is simple and the technologies for automatic translation work well when applied to this area. [4].

In the case of cross-lingual information retrieval, queries are very small (2 to 3 terms on average) and because of this a polysemy problem becomes a crucial issue.

On-line query translation was applied by many researchers. Authors of one such study [1] investigated the features of the Systran on-line translation system for the Chinese - English language pair when queries are presented in Chinese, and documents are given in English. Their test collection includes 28,133 documents. Each document is about 50 words. They applied the statistical translation model to manage the ambiguity problem for the queries submitted. The key idea of their approach is to create a bilingual corpus automatically. Firstly, they translated the document set into Chinese and then back into English. After that, they calculated the translation probabilities between Chinese and English words. Using these probabilities, they translated Chinese queries into English and gained an improvement in the retrieval: Average precision increased from 0.245 to 0.293. This is a resource-intensive approach. It can be applied to the ad-hoc task. In the case of Web retrieval, its utilization seems to be difficult.

The aim of this study is to investigate the efficiency of on-line translation systems in managing the polysemy problem for the query terms, when systems automatically translate them and to test techniques which do not require utilizing training data and adjusting the parameters of the retrieval system.

2 On-line translation techniques

In our experiments at the NTCIR 6 Workshop, we investigated how on-line translation systems are

Table 1. Characteristics of the official runs (Stage 1)

Run	Description	Comments
OASIS-C-J-T-01	Web based query translation: Chinese-Japanese using service [11]	Baseline to make a comparison
OASIS-C-J-T-02	semi automatic Web translation: Chinese into English and then into Japanese using service [11]	English as the pilot language
OASIS-C-J-T-03	Web based query translation: Chinese-Japanese using services [11] and [12], merging results to expand queries	Merging results of two different on-line translation systems
OASIS-C-J-D-04	Web based query translation: Chinese-Japanese using services [11] and [12], merging results to expand queries	Merging results of two different on-line translation systems

applicable as auxiliary tools when translating from Chinese into Japanese. We tested two approaches to translate queries. One is to use different on-line systems and to merge the translation results. The second is to utilize English as a pilot language and make translation from Chinese into Japanese as follows: Chinese – English –Japanese.

Our aim was to investigate the following questions: How efficient is a translation system in managing the polysemy problem for the query terms, when the system automatically translates them? Can merging results of translations by different on-line systems improve the retrieval accuracy?

On-line translation models for the Chinese – Japanese pair can be characterized as follows:

- Direct translation model: Japanese -> Chinese and Chinese -> Japanese
- Model using a pilot language (English): Japanese -> English -> Chinese and Chinese -> English -> Japanese.

According to our view, the model utilizing a pilot language has several advantages compared to the direct translation model.

- The number of available dictionaries for language pairs of English – Japanese and English - Chinese is much large than for the language pair of Chinese – Japanese.
- The most advanced translation methods are firstly implemented for the pairs of English – other language because these pairs are more demanding by customers.
- Using English as the pilot language provides extra chances to expand queries in a more accurate way.

A pilot language may affect the semantics of the query: Meaning can be changed. This is a disadvantage of this approach. On the other hand, queries submitted to the general purpose search engines are very short. They usually consist of 2.5 terms on average. An outcome from this is: A search engine has always to guess somehow the query meaning.

The advantage of the direct translation model is in the accurate term translation.

In any case, it is not clear how to manage the polysemy problem for the query terms, when the system automatically translates them.

After testing on-line dictionaries of Japanese and English synonyms [9, 10], we decided not to apply them because the number of chouses is very large for every entry; and it is difficult to apply them without human inspection.

OASIS participated in Stage 1 and Stage 2 of the CLIR task. All test collections and test queries provided by organizers were applied. Their description is presented in [8].

We utilized the following translation systems: WorldLingo [11], and Excite [12]. Our strategy was as follows:

- Apply them to translate queries directly from Chinese into Japanese,
- Utilize the model of Chinese – English – Japanese translation, where English is the pilot language,
- Merge the translation results produced by different systems for the possible expansion of queries.

There are many tools (search engines) in the public domain [13]. They implement any model: vector space, probabilistic, and boolean. Some of them are reliable and work stable. There is no need for researchers in information retrieval to develop them. The researcher can concentrate on designing, implementing and testing heuristics. Anyway, we use an improved version of the OASIS search engine. It employs the vector space model.

Nowadays, powerful morphological analyzers become standard de facto to index text in Asian languages. We utilized Mecab [2] as a segmentation tool. We used information about the part of speech of the words generated by the morphological analyzer: Nouns and verbs were filtered in the “D” runs.

3 Stage 1: Results of experiments

Four official runs were submitted for Stage 1. Their characteristics are presented in Table 1.

Table 2. Results of the Stage 1 runs ("Relaxed" relevance judgment)

Runs	R-precision	Precision at 5 docs	Precision at 10 docs	Precision at 30 docs	Comments
OASIS-C-J-T-02	0.11	0.18	0.15	0.13	Calculated for 43 "successful" queries
OASIS-C-J-T-03	0.13	0.20	0.18	0.16	Calculated for the full set of 50 queries
OASIS-C-J-D-04	0.10	0.16	0.14	0.12	

Results of the OASIS-C-J-T-01 and OASIS-C-J-T-02 runs are very similar. English as the pilot language did not help much. The search engine failed to retrieve documents in response to queries: 15, 17, 20, 37, 41, 53, and 58. Table 2 presents the main statistics.

Merging results of translation from different on-line translation systems to expand queries improved the accuracy of retrieval (see Table 2, row OASIS-C-J-T-03). The translation results were simply combined. An assumption, that terms at the beginning of the query are more important, was used when applying the merging operation. This assumption is common for the general purpose search engines. We have to note that this merging operation is not commutative. An illustration of this point is given in Section 5. The key outcome from this strategy is: Currently being in use on-line translation systems significantly differ from each other and combining results makes sense.

In the case of "Relaxed" relevance judgment, produced results are slightly better (see Table 3, column OASIS-C-J-T-03). The system retrieved all relevant documents for query 03. The number of top ranked documents which are relevant to the query is very important in practice. In our case, the number of queries with precision more than 0.4 at 5 and 10 documents is equal to 15 and to 9 respectively. The number of queries with zero precision at 10 documents is equal to 20. The total number of queries is 50.

Our "D" run OASIS-C-J-D-04 used information from the topic description for each topic to generate queries. Topic descriptions consist of ordinary Chinese sentences.

When we set up our tests, we took into account that the aim of on-line translation systems is to translate sentences accurately. Translation of sentences has to simplify the managing the polysemy problem because the longer the text is, the clearer its semantics becomes. The latter statement is correct when people communicate with each other. The difficult problem is to select key terms as a query to retrieve relevant information. We utilized a very simple translation procedure. Two on-line translation services [11, 12] were applied. Results of direct translation from Chinese into Japanese were passed to Mecab. Only nouns and verb were selected from Mecab's outcome. Outputs from both services were merged. This merging is also not commutative because importance of a term depends on its place in the query string. The maximum number of terms was 18 (this is a restriction of our search engine).

According to the data distributed by organizers, accuracy of all systems decreased on average when they performed retrieval using descriptions of topics to form queries (see Table 4). Our system demonstrated the same tendency. Its performance with queries generated from the description part was relatively worse compared to the "T" runs. "Rigid" and "Relaxed" judgments of the retrieval results for "T" and "D" runs showed that results were more accurate when the "Relaxed" case was applied.

Table 3 provides comparison between runs OASIS-C-J-T-04 and OASIS-C-J-D-04.

Figure 1 gives an illustration of how our solution relates to other approaches tested by NTCIR participants: The performance of our system is less than average.

4 Stage 2: Results of experiments

The same approaches were utilized at the Stage 2 tests. At the time we were running our tests (September 2006), the WorldLingo service [11] did not manage with translation of the following topics: NTCIR-4: 004, title; NTCIR-5: 027, title, description; 028 description; 030 title, description. Table 5 characterizes the runs. Table 6 presents the main results for runs produced the better retrieval.

Analyzing the data presented in Table 6, we can conclude that same techniques produced different accuracy on different collections. Using English as the pilot language when applying the WorldLingo service does not help to improve retrieval results (see rows 4 to 6). Direct translation of queries for the NTCIR-4 and NTCIR-5 collection sets brought about the better accuracy compared to other approaches (see rows 2, 3, 5, 6, 8, 9, 11, and 12). The most difficult for retrieval was the NTCIR-4 collection (see rows 2, 5, 8, and 11). The easiest was the NTCIR-3 collection. Tests using the NTCIR-3 collection performed better results when

Table 3. Results of the OASIS-C-J-T-03 and OASIS-C-J-D-04 runs (Stage 1)

Query	Relaxed: Relevant	OASIS-C-J-T-03				OASIS-C-J-D-04			
		Relaxed: Rel-retrieved	Precision			Relaxed: Rel-retrieved	Precision		
			At 5	At 10	At 30		At 5	At 10	At 30
03	60	60	0.8	0.7	0.56	59	0.8	0.5	0.67
14	20	18	0.0	0.1	0.07	16	0.0	0.1	0.1
15	30	27	0.4	0.3	0.16	30	0.4	0.3	0.13
16	47	45	0.2	0.2	0.23	44	0.2	0.3	0.17
17	46	35	0.0	0.2	0.26	32	0.0	0.1	0.2
18	100	28	0.0	0.0	0.03	29	0.0	0.0	0.0
19	173	37	0.0	0.1	0.1	34	0.0	0.0	0.0
20	51	3	0.0	0.0	0.0	2	0.0	0.0	0.0
21	26	23	0.4	0.4	0.27	23	0.4	0.3	0.26
23	64	49	0.0	0.1	0.13	44	0.0	0.0	0.03
24	36	19	0.2	0.1	0.13	16	0.0	0.0	0.13
26	36	24	0.0	0.0	0.07	23	0.0	0.0	0.07
27	27	18	0.2	0.1	0.03	16	0.0	0.1	0.07
30	239	38	0.0	0.0	0.0	19	0.0	0.0	0.0
33	99	10	0.0	0.0	0.0	13	0.0	0.0	0.0
36	127	101	0.2	0.1	0.37	79	0.2	0.2	0.33
37	146	39	0.0	0.0	0.0	21	0.0	0.0	0.0
39	59	12	0.0	0.0	0.0	9	0.0	0.0	0.0
41	216	117	0.8	0.6	0.46	111	0.8	0.6	0.43
42	24	7	0.0	0.0	0.0	6	0.0	0.0	0.0
43	48	14	0.4	0.3	0.1	9	0.0	0.1	0.07
44	33	25	0.4	0.2	0.07	24	0.4	0.2	0.07
45	233	92	0.0	0.1	0.1	21	0.0	0.0	0.03
46	94	79	0.6	0.6	0.33	49	0.6	0.4	0.23
47	154	56	0.0	0.0	0.0	69	0.0	0.1	0.1
48	122	81	0.0	0.0	0.13	65	0.0	0.0	0.07
50	174	17	0.0	0.0	0.0	1	0.0	0.0	0.0
53	104	58	0.6	0.4	0.4	56	0.8	0.5	0.36
58	129	83	0.8	0.5	0.4	58	0.2	0.3	0.23
59	311	205	0.8	0.9	0.77	191	0.8	0.9	0.7
60	266	65	0.2	0.4	0.23	20	0.0	0.0	0.07
64	105	65	0.4	0.3	0.4	51	0.4	0.3	0.27
65	29	5	0.0	0.0	0.0	11	0.0	0.0	0.03
70	4	1	0.0	0.0	0.03	1	0.0	0.0	0.0
74	198	176	0.6	0.3	0.4	57	0.0	0.0	0.03
75	57	45	0.0	0.0	0.0	40	0.2	0.1	0.7
77	14	4	0.0	0.0	0.0	4	0.0	0.0	0.0
79	112	76	0.2	0.3	0.3	73	0.2	0.3	0.2
80	122	93	0.6	0.6	0.6	93	0.6	0.6	0.6
83	45	29	0.4	0.2	0.1	28	0.4	0.2	0.1
95	63	1	0.0	0.0	0.0	6	0.0	0.0	0.0
96	180	31	0.0	0.0	0.07	17	0.0	0.0	0.03
97	71	21	0.0	0.0	0.0	6	0.0	0.0	0.0
99	128	57	0.4	0.3	0.16	48	0.4	0.3	0.17
100	19	11	0.0	0.0	0.0	5	0.0	0.0	0.0
102	29	16	0.0	0.0	0.03	15	0.0	0.0	0.03
103	106	16	0.2	0.1	0.07	46	0.0	0.1	0.1
105	104	40	0.0	0.1	0.2	24	0.0	0.0	0.03
106	65	40	0.2	0.1	0.03	0	0.0	0.0	0.0
110	19	7	0.0	0.1	0.03	7	0.0	0.0	0.03
Average	4764	2219	0.2	0.17	0.15	1721	0.16	0.14	0.12

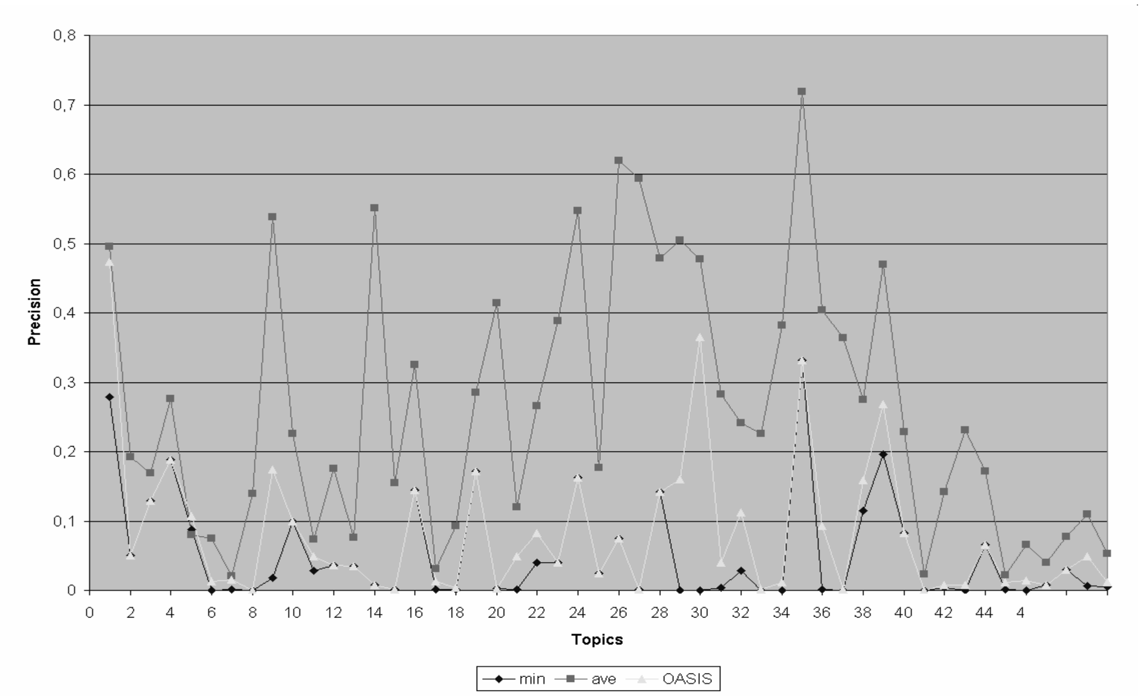

Figure 1. Compared precision Stage 1: MAP_C-J-T-Relaxed – OASIS-C-J-T-03

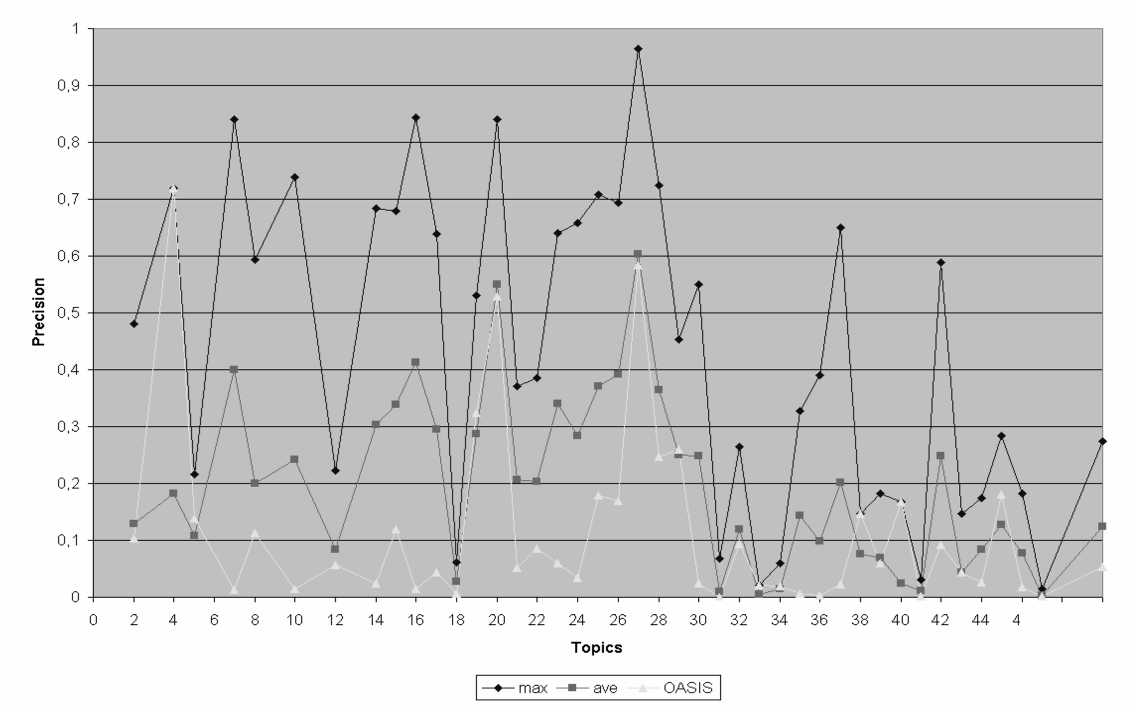

Figure 2. Compared precision Stage 2: MAP_C-J-D-Rigid – OASIS-C-J-D-N3

Table 4. Performance of T and D runs ("Relaxed" relevance judgment)

All participants	Precision	
	Average	Max
T runs	0.27	0.38
D runs	0.25	0.37

"Rigid' relevance judgment was applied. "Relaxed' relevance judgment generated relatively higher results for all approaches for other collections. The outcome of "D" runs was less accurate except only for the NTCIR-3 collection (see rows 1, 4, 7, and 10). Removing hiragana characters from the queries did not affect much the quality of retrieval.

Table 5. Characteristics of the official runs (Stage 2)

Runs	Description	Comments
OASIS-C-J-T-04-N3, OASIS-C-J-T-04-N4, OASIS-C-J-T-04-N5	Web based query translation: Chinese-Japanese using service [12]	Hiragana characters were not deleted
OASIS-C-J-T-03-N3, OASIS-C-J-T-03-N4, OASIS-C-J-T-03-N5	semi automatic Web translation: Chinese into English and then into Japanese using service [11]	Hiragana characters were not deleted
OASIS-C-J-T-02-N3, OASIS-C-J-T-02-N4, OASIS-C-J-T-02-N5	Web based query translation: Chinese-Japanese using services [11] and [12], merging results to expand queries	Hiragana characters were deleted
OASIS-C-J-D-01-N3, OASIS-C-J-D-01-N4, OASIS-C-J-D-01-N5	Web based query translation: Chinese-Japanese using services [11] and [12], merging results to expand queries	Hiragana characters were deleted

Table 6. Results of the Stage 2 runs

Num	Run	R-Precision	Precision at 5 docs	Precision at 10 docs	Precision at 30 docs	Relevance judgment
1	OASIS-C-J-T-04-N3	0.1286	0.1952	0.1619	0.1190	Rigid
2	OASIS-C-J-T-04-N4	0.0656	0.1164	0.0964	0.0885	Relaxed
3	OASIS-C-J-T-04-N5	0.1150	0.1915	0.1660	0.1319	Relaxed
4	OASIS-C-J-T-03-N3	0.0995	0.1381	0.1190	0.0817	Rigid
5	OASIS-C-J-T-03-N4	0.0391	0.0545	0.0491	0.0467	Relaxed
6	OASIS-C-J-T-03-N5	0.0649	0.1106	0.0936	0.0844	Relaxed
7	OASIS-C-J-T-02-N3	0.1387	0.1952	0.1619	0.1254	Rigid
8	OASIS-C-J-T-02-N4	0.0628	0.0982	0.0873	0.0806	Relaxed
9	OASIS-C-J-T-02-N5	0.1012	0.1745	0.1532	0.1163	Relaxed
10	OASIS-C-J-D-01-N3	0.1336	0.1952	0.1643	0.1222	Rigid
11	OASIS-C-J-D-01-N4	0.0479	0.0982	0.0764	0.0648	Relaxed
12	OASIS-C-J-D-01-N5	0.0879	0.1234	0.1340	0.1149	Relaxed

Table 7. Results of additional experiments (Stage 1)

Num	Run	R-Precision	Precision at 5 docs	Precision at 10 docs	Precision at 30 docs	Comments
1	Babel Fish	0.0955	0.1600	0.1300	0.1113	Service [14]
2	DictDotCom	0.1112	0.1760	0.1620	0.1333	Service [15]
3	Google	0.1163	0.1840	0.1680	0.1427	Service [16]
4	Google – Babel Fish	0.1087	0.1840	0.1640	0.1347	Services [14, 16]
5	Babel Fish – Google	0.1068	0.1800	0.1620	0.1333	Services [14, 16]
6	DictDotCom – Babel Fish	0.1008	0.1720	0.1480	0.1220	Services [14, 15]
7	Google – DictDotCom	0.1139	0.1680	0.1620	0.1407	Services [15, 16]

Figure 2 gives a comparison between the tested approach of direct query translation (run OASIS-C-J-D-01-N3) and average results generated by other participants on the NTCIR-3 collection. Our run produced the best retrieval (among all participants) for topics: 4, 38, and 40. It performed the results at the average level for topics 5, 19, 20, 27, 29, 31, 34, 39, 41, 43, 45, and 47. Its total performance was less than average.

5 Additional experiments

To investigate the impact of English as the pilot language on the retrieval performance, we conducted additional experiments using the set of the Stage 1 test collections. We applied the procedure described in Section 3 utilizing the following on-line translation services: Babel Fish, Dictionary.com, and Google [14, 15, 16].

There is only one possible way to convert Chinese queries into Japanese with help of these tools: The first translation has to be done into English. Main results of our runs in the case of "Relaxed" relevance judgment are presented in Table 7. They were more accurate compared to the "Rigid" case. For runs 4 to 7, queries were translated utilizing two systems. After that, the merging operation was applied to expand them. Before retrieval, hiragana characters were removed from the queries. Retrieval results were more accurate when the Google service [16] was applied for translation (see row 3, Table 7). Runs 4 and 5 illustrate our statement (see Section 3) that our merging operation is not commutative: In run 4, keywords generated by Google were followed by query terms produced by the Babel Fish service. In run 5, keywords from the Babel Fish service were at the head of the query string.

From our results (run OASIS-C-T-J-03, Table 2 and all runs, Table 7), we can see that English as the pilot language did not impact on retrieval positively. Using services for direct translation (Chinese into Japanese) and merging their results to expand queries is the more efficient strategy compared to utilizing services with English as the pilot language.

Our query expansion technique affected retrieval positively in the case of services [14, 15] and negatively in the case of service [16].

6 Conclusions

The efficiency of on-line translation systems were tested when they were applied to translate queries from Chinese into Japanese for the CLIR task. Techniques which do not require utilizing training data and adjusting the parameters of the retrieval system were applied. Results showed that outcomes of translation systems are different because the systems use various approaches, different dictionaries, etc.

Using several systems and combining their outcomes may be helpful to expand queries and improve the quality of retrieval for the CLIR task. This strategy can be used as part of a set of measures to improve cross-lingual retrieval. Our experiments to employ English as the pilot language did not produce the significant changes in the retrieval.

References

[1] Rong Jin and Joyce Y. Chai. Study of Cross Lingual Information Retrieval Using On-line Translation Systems, in Proc. *of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Salvador, Brazil, 2005.

[2] MeCab: Yet Another Part-of-Speech and Morphological Analyzer. Available at: http://mecab.sourceforge.net/ [accessed in March 2007].

[3] Ying Zhang and Phill Vines. Using the Web for Translation Disambiguation: RMIT University at NTCIR-5 Chinese-English CLIR, in the Proc. *of the Fifth NTCIR Workshop 5 Meeting on evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access*, Tokyo, Japan, 2005.

[4] V. Klyuev. Modern Language Information Technologies: How they Help Japanese Students, in the Proc. *of the 6th IEEE International Conference on Advanced Learning Technologies*, Kerkrade, the Netherlands, 2006.

[5] Jianqiang Wang and Douglas W. Oard. Combining Bidirectional Translation and Synonymy for Cross-Language Information Retrieval, in the Proc. *of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Seattle, Washington, USA, 2006.

[6] Jianfeng Gao and Jian-Yun Nie, A Study of Statistical Models for Query Translation: Finding a Good Unit of Translation, in the Proc. *of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Seattle, Washington, USA, 2006.

[7] Kazuaki Kishida, Kuang-hua chen, et al., Overview of CLIR Task at the Fifth NTCIR Workshop, in the Proc. *of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access*, Tokyo, Japan, 2005.

[8] Kazuaki Kishida, Kuang-hua Chen, et al., Overview of CLIR Task at the Sixth NTCIR Workshop, in the Proc. *of the Sixth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access*, Tokyo, Japan, 2007.

[9] On-line Dictionary of Japanese Synonyms, available at http://www.gengokk.co.jp/thesaurus/ [Accessed in March 2007].

[10] On-line dictionary of English Synonyms, available at http://vancouver-webpages.com/synonyms.html [Accessed in March 2007].

[11] WorldLingo on-line translation system, available at http://www.worldlingo.com/ [Accessed in March 2007].

[12] Excite on-line translation system, available at www.excite.co.jp [Accessed in March 2007].

[13] Open source Search engines, available at http://www.searchtools.com/tools/tools-opensource.html [Accessed in March 2007].

[14] Babel Fish on-line translation system, available at http://babelfish.altavista.com/ [Accessed in April 2007].

[15] Dictionary.com Translator, available at http://dictionary.reference.com/translate/ [Accessed in April 2007].

[16] Google on-line translation system, available at http://www.google.com/language_tools [Accessed in April 2007]