Justsystem-Clairvoyance CLIR Experiments at NTCIR-4 Workshop

Yan Qu[†], Gregory Grefenstette[†], David A. Hull[†], David A. Evans[†]
Motoko Ishikawa[‡], Setsuko Nara[‡], Toshiya Ueda[‡], Daisuke Noda[‡]
Kousaku Arita[‡], Yuki Funakoshi[‡], Hiroshi Matsuda[‡]
[†]Clairvoyance Corporation
5001 Baum Boulevard, Suite 700
Pittsburgh, PA 15213-1854, USA
{yqu,grefen,hull,dae}@clairvoyancecorp.com
[‡]Justsystem Corporation
108-4 Hiraishi Wakamatsu, Kawauchi-cho
Tokushima-shi Tokushima-ken 771-0189 Japan
{Motoko_Ishikawa, Setsuko_Nara,Toshiya_Ueda, Daisuke_Noda}@justsystem.co.jp
{Kousaku_Arita,Yuki_Funakoshi,Hiroshi_Matsuda}@justsystem.co.jp

Abstract

At the NTCIR-4 workshop, Justsystem Corporation and Clairvoyance Corporation collaborated in participating in the Cross-Language Retrieval Task (CLIR). We submitted results to the sub-tracks of SLIR and BLIR. For the SLIR track, we submitted Chinese, English, and Japanese monolingual runs. For the BLIR track, we submitted Japanese-English and Chinese-English runs. The major goal of our participation is to evaluate performance and robustness of our recently developed commercial-grade CLIR systems for English, Japanese, and Chinese.

Keywords: *Cross-lingual information retrieval; Evaluation; Retrieval experiments*

1. Introduction

At the NTCIR-4 workshop, Justsystem Corporation (JSC) in Japan and Clairvoyance Corporation (CC) in the USA collaborated in participating in the Cross-Language Retrieval Task (CLIR). A major goal of our participation is to evaluate performance and robustness of our recently developed commercial-grade CLIR systems for English, Japanese, and Chinese. We compared three systems under development or upgrade. The CLIR track has four sub-tracks: single language IR (SLIR), bilingual CLIR (BLIR), bilingual IR via pivot languages (PLIR), and multilingual CLIR (MLIR). We submitted results to the sub-tracks of SLIR and BLIR. For the SLIR track, we submitted Chinese, English, and Japanese monolingual runs. For the BLIR track, we submitted Japanese-English and Chinese-English runs. For each language pair, we submitted two runs based on the *title* (denoted by T) field, two runs based on the *desc* (D) field, and one run based on the *desc* and *narrative* (DN) field. For all the runs, we report the average precisions and overall recalls using the set of *rigid* relevant documents.

2. System Description

Justsystem Corporation and Clairvoyance Corporation share a common system framework for information retrieval and management, which serves as the foundation of the commercial CLARIT APIs from Clairvoyance for the English language [1] and the commercial ConceptBase product in Japan for the English, Japanese, and Chinese languages. Major functionalities include natural language processing, ad-hoc retrieval, feedback, visualization, etc. Recently, we have added cross-language text retrieval (CLIR) capability into the framework. Both the monolingual systems and the CLIR systems are highly parameterized to allow for system experimentation and optimization.

In an effort to test the performance of all of our available text retrieval tools, we used two different indexing systems and three different retrieval systems in these experiments. CLARIT, a commercial information management toolkit developed at Clairvoyance, served as the indexing and retrieval system for two English runs. ConceptBase Java (CBJ), a commercial text retrieval system developed at Justsystem, served as the indexing system for all the Japanese and Chinese runs, as well as the remaining English runs. In addition, we tested a research text retrieval toolkit called CLJ that runs on top of either the CLARIT or the CBJ indexing engine, and serves as a development environment for our latest research text retrieval algorithms. At NTCIR-4, all CLJ runs are based on the CBJ indexing engine. These three different systems are referred to as CLARIT, CBJ, and CLJ in the remainder of this report.

We have experimented with different features of the monolingual and cross-language systems to identify the significant contributors to an effective

IR system. In the following, we first present components and features shared by both the monolingual and bilingual retrieval system: indexing, retrieval, pseudo-relevance feedback, and multi-word term down-weighting. Then we present strategies related to bilingual retrieval, including query translation disambiguation and translation structuring.

2.1 Indexing and Retrieval

Both CLARIT and CBJ use NLP for tokenization, storing individual words, full noun phrases, and attested sub-phrases as index terms. An attested sub-phrase is a constituent of a longer noun phrase that also appears independently as a full noun phrase elsewhere in the document collection. CLARIT uses a lexicon-based tokenizer and finite state machine based grammar for English processing. CBJ uses a statistical part-of-speech tagger for tokenization and finite state machine based grammar for processing English, Japanese and Chinese. Indexing involves statistical analysis of a text corpus and construction of an inverted index, with each index entry specifying the index word and a list of texts. Both systems allow the index to be built upon full documents or variable-length subdocuments. We used subdocuments as the basis for indexing and document scoring in our experiments. Sub-documents range in size from 8 to 20 sentences and average about 12 sentences in length.

Retrieval is based on the vector space retrieval model. Various similarity measures are supported in the model. For CBJ and CLARIT in NTCIR-4, we used the dot product function for computing similarities between a query and a document:

$$sim(Q,D) = \sum_{t \in Q \cap D} W_Q(t) \bullet W_D(t) \qquad (1)$$

where $W_Q(t)$ is the weight associated with the query term t and $W_D(t)$ is the weight associated with the term t in the document D. The two weights were computed as follows:

$$W_D(t) = TF_D(t) \bullet IDF(t) \qquad (2)$$

$$W_Q(t) = C(t) \bullet TF_Q(t) \bullet IDF(t) \qquad (3)$$

where IDF and TF are standard inverse document frequency and term frequency statistics, respectively. *IDF(t)* was computed with the target corpus for retrieval. The coefficient *C(t)* is an "importance coefficient", which can be modified either manually by the user or automatically by the system (e.g., updated during feedback).

CLJ uses the same inner produce of the query term weights $W_Q(t)$ and the document term weights $W_D(t)$ as shown in formula (1) to compute the similarity score between query Q and document D. The query term weights are computed with formula (3) again with the coefficient *C(t)* for assigning differential weights to terms.

The document term weights are standard BM25 [7], as shown in formula (4), in which k_1 is the term frequency smoothing parameter, *b* is the document length smoothing parameter, *d* is the document length, and **Δ** is the average document length in the collection.

$$W_D(t) = \frac{(k_1+1) * TF_D(t)}{k_1[(1-b)+b*(d/\Delta)]+TF_D(t)} \qquad (4)$$

2.2 Query Expansion

Query expansion through (pseudo-) relevance feedback has proved to be effective for improving IR performance. We used pseudo-relevance feedback for augmenting the queries. After retrieving some documents for a given topic from the target corpus, we took a set of top ranked documents, regarding them as relevant documents to the query, and extracted terms from the these documents. We use two formulae – Prob2 and Rocchio – for extracting and ranking terms for expansion.

$$\mathrm{Pr}ob2(t) = \log(R_t + 1) \times \left(\log(\frac{N-R+2}{N_t - R_t + 1} - 1) - \log(\frac{R+1}{R} - 1) \right) \qquad (5)$$

where *N* is the number of sub-documents in the reference corpus, N_t is the number of sub-documents that contain the term *t* in the corpus, *R* is the number of sub-documents in the top *n* documents, and R_t is the number of sub-documents that contain the term *t* in the top *n* documents. The *k* terms with the highest score according to this measure are selected and merged with original query to create the final expanded query.

Another formula for extracting terms is a modified version of the Rocchio formula to rank terms in a given set of documents:

$$Rocchio\ (t) = IDF\ (t) \bullet \frac{\sum_{D \in DocSet} TF_D(t)}{NumDoc} \qquad (6)$$

where IDF(t) is the Inverse Document Frequency of term *t* in reference database, NumDoc the number of documents in the given set of documents, and $\mathrm{TF_D}$(t) the term frequency score for term *t* in document D.

Once terms for expansion were extracted and ranked, they were combined with the original terms in a query to form an expanded query.

$$Q_{new} = k \times Q_{orig} + Q_{\exp} \qquad (7)$$

in which Q_{new} , Q_{orig} , Q_{orig} stand for the new expanded query, the original query, and terms extracted for expansion, respectively. Weighting options for Q_{exp} include:

- Constant: all expansion terms take the same weight (e.g., *W(t)* = 1)
- Normalized: the expansion terms take the Rocchio or Prob2 scores normalized by their appropriate max scores, e.g., with the Rocchio formula,

$$W(t) = \frac{W_{Rocchio}(t)}{\max(W_{Rocchio}(t))} \tag{8}$$

- Scaled: both Q_{orig} and Q_{exp} are normalized with the sum of the term scores. For Q_{orig} , the original weights are used. For Q_{exp}, the Rocchio or Prob2 scores are used. E.g., with the Prob2 formula,

$$W(t) = \frac{W_{\mathrm{Pr}\,ob2}(t)}{\sum W_{\mathrm{Pr}\,ob2}(t)} \tag{9}$$

The following is a complete description of the term scoring and expansion steps using CLJ as an example:

- Retrieve the top **10** documents using the original query **Q_{orig}** and weight function **BM25(k₁=1.2, b=0.2)** and phrase weight = **0.1**
- Select the top **35** terms based on the Prob2 formula after throwing out all terms with frequency one in the top 10 documents to create **$\mathbf{Q_{exp}}$**. The original query terms are included in the selection process.
- Merge the queries as follows: **$\mathbf{Q_{new} = 0.25 \times Q_{orig} + Q_{exp}}$** [Note that **$\mathbf{Q_{orig}}$** and **$\mathbf{Q_{exp}}$** are normalized so that their weights sum to 1.0 prior to merging.]
- Retrieve the final document set using **$\mathbf{Q_{new}}$** and weight function **BM25($\mathbf{k_1}$=1.2, b=0.3)** and phrase weight = **0.2**

For CLJ, all parameters were optimized on the NTCIR-3 query set for the three languages as a whole. We did not try to optimize on each language independently to reduce over-fitting. In practice, we have found the optimal parameters to be very similar for all languages. For CBJ and CLARIT, we optimized on the NTCIR-3 query sets for individual language pairs. The settings will be reported in sections with the corresponding evaluation runs.

2.3 Multi-Word Term Down-weighting

Fujita (1999) observed that down-weighting of phrasal terms helped with retrieval performance for the NTCIR-1 tasks [2]. We confirmed this observation in our training experiments with NTCIR-3 data [3]. We applied down-weighting phrasal terms in all three retrieval systems with the use of the coefficient *C(t)*. For NTCIR-4, we applied a weight of 1.0 to all words and 0.1 or 0.2 to all multi-word phrases.

2.4 Cross-Lingual Retrieval Strategies

For bilingual CLIR, we adopted query translation as the means for bridging the language gap between the query language (e.g., Japanese) and the document language (e.g., English).

In addition to the language independent features such as pseudo-relevance feedback and multi-word term down-weighting, we have implemented several cross-language specific strategies: (1) translation term filtering, (2) translation disambiguation, (3) down-weighting of multi-word terms, (4) suppression of translations with low distribution, and (5) translation structuring. We describe these in detail below.

First, we filter out one-character source language terms or one-character subterms, as we have observed in our training experiments that these terms often result from wrong segmentation or that these terms are too ambiguous.

Then, the remaining terms are translated via look-up in the bilingual dictionaries. Multiple translations of a source term are disambiguated through the parallel corpora. We use the aligned Yomiuri Japanese–English parallel corpora [9]. The process is as follows:

1) Get translations for each term in the source language query;
2) For a source term that has multiple possible translations, search for sentences in the source language corpus that contain the source term in the query;
3) Obtain corresponding sentences in the target language corpus;
4) Keep translations of the source terms that are present in the obtained target language sentences;
5) Collate the kept translations into a target language query.

Third, we apply several down-weighting techniques. The weight of multi-word terms is multiplied by 0.2; weights for translations for parts of phrases are multiplied by 0.8.

Fourth, we suppress translations of terms and multi-word terms with low distributions. The intuition is that if a translation has very low distribution, then it is unlikely to be a good translation.

Finally, multiple translations are balanced in retrieval [6]. To compute similarity scores between a document *D* in the target language and multiple translations t_i of a source term *s*, we use only one translation that produces the highest similarity score (MAX), which is computed as the highest similarity among all similarities between *D* and each target language term *t* translated from *s*.

3. CLIR Retrieval Track

We have participated in two sub-tracks of the CLIR track: single language IR (SLIR) and bilingual IR (BLIR). For details on the CLIR track and its sub-tracks, the topic sets, the document collections, and evaluation of the tracks, the reader is referred to the overview of the CLIR track [4]. For SLIR, we submitted runs for Japanese, English, and Chinese monolingual retrieval. For BLIR, we submitted runs for Japanese-English and Chinese-English retrieval. We report the results of these runs based on evaluation against "rigid" relevant documents.

3.1 Single Language IR track

This section describes the parameters used for the monolingual runs at NTCIR-4. First, the documents and topics were parsed into linguistically meaningful units: NPs, Adj, Adv, and Verbs, which were then used as indexing terms for building monolingual database. For Japanese and Chinese, we used CBJ for such processing. For English, we used either CLARIT or CBJ (used by CLJ). Surface variants were normalized to their root forms. Stop word lists were constructed for the three languages to filter out general stop words and query-dependent words such as 記述/description and 内容/information.

For Chinese processing, the part-of-speech tagger in CBJ was originally developed for simplified Chinese. We conducted character-based substitution between simplified Chinese characters and traditional Chinese characters to make the module process traditional Chinese characters. The simple conversion was prone to error because of the ambiguity in converting traditional Chinese characters to simplified Chinese characters.

3.1.1 Japanese Retrieval

For Japanese retrieval, we used CBJ to process the documents and topics. For retrieval, we compared CBJ and CLJ. CBJ used formulae (2) and (3) term weighting and the Rocchio method was used for extracting terms. We used the top 30 terms from the top 30 documents for query expansion. Formula (8) was used for feedback term weighting. The CLJ system used Prob2 for extracting 35 query expansion terms from the top 10 documents and formula (9) for merging expansion terms with the original query terms.

The official results from NTCIR-4 were presented in Table 1. CLJ outperformed CBJ overall.

From the description of the term weighting algorithms in section 2.2, we see that CLJ has a lot of parameters. This is not a serious drawback as long as an appropriate set of default parameters is available and the system is relatively robust to minor changes in the parameter settings. In order to measure the robustness of the system, we performed a sensitivity analysis on the monolingual Japanese title run from CLJ, based on the rigid relevance judgments. For each of the system components, we compute average precision over a range of parameter values. To conserve space, we report only the total range of results, rather than the complete performance table. The first row of Table 2 shows the pre-expansion performance. All other results are computed after query expansion.

Run	Feature	Avg prec	Recall
J-J-T-cbj	Rocchio Formula (8)	0.2686	4487/7137
J-J-T-clj	Prob2 Formula (9)	0.389	5868/7137
J-J-D-cbj	Rocchio Formula (8)	0.2622	4417/7137
J-J-D-clj	Prob2 Formula (9)	0.3747	5684/7137

Table 1: Japanese Retrieval, Rigid

We can see that performance varies by no more than about 10% for each set of parameters, indicating that the system is relatively stable (Table 2). Even more encouraging, we find that optimizing over the NTCIR-3 collection was extremely effective, putting us at or near the top of the range in every case. The only way we could have improved our performance would have been to expand with 30 terms instead of 35, giving us a meager gain of 0.002.

Parameters	Submission	Range
BM25 (k_1, b)	0.311	0.306-0.311
Phrase weight (0.0-1.0)	0.389	0.361-0.389
# docs (5-20) # terms (20-40)	0.389	0.361-0.391
Query weight (0.0-1.0)	0.389	0.366-0.389

Table 2: Japanese Retrieval Parameter Calibration, Rigid

3.1.2 English Retrieval

run	feature	Avg prec	Recall
E-E-T-clarit	Rocchio Formula (8)	0.3145	4403/5866
E-E-T-clj	Prob2 Formula (9)	0.3412	4259/5866
E-E-D-clarit	Rocchio Formula (8)	0.307	4380/5866
E-E-D-clj	Prob2 Formula (9)	0.3382	4500/5866

Table 3: English Retrieval, Rigid

For English retrieval, we compared CLARIT and CLJ, with CLJ taking the index from CBJ. CLARIT used formulae (2) and (3) term weighting and the Rocchio method was used for extracting terms. We used the top 30 terms from the top 20 documents for query expansion. Formula (8) was used for feedback term weighting. The CLJ system used Prob2 for extracting query expansion terms and formula (9) for merging expansion terms with the original query terms.

The official results from NTCIR-4 were presented in Table 3. The CLJ based runs performed overall better compared with the CLARIT based runs.

3.1.3 Chinese Retrieval

For Chinese retrieval, we compared CBJ and CLJ. In CBJ runs, we used tf*idf score for term weighting and the Rocchio method for query expansion. We used the top 30 terms from the top 20 documents for query expansion. In CLJ runs, Prob2 was used for extracting expansion terms and scaled term weighting was used for merging feedback terms with the original query terms.

run	feature	Avg prec	Recall
C-C-T-cbj	Rocchio Formula (8)	0.1327	874/1318
C-C-T-clj	Prob2 Formula (9)	0.1899	1017/1318
C-C-D-cbj	Rocchio Formula (8)	0.1384	809/1318
C-C-D-clj	Prob2 Formula (9)	0.1886	1062/1318

Table 4: Chinese Retrieval, Rigid

The official results from NTCIR-4 were presented in Table 4. Again, CLJ based runs had higher scores than the CBJ based runs. However, the results of all our runs were low compared with those of many groups in the NTCIR-4 submission. Preliminary analysis suggests that missing lexical terms in the parsing dictionary and wrong conversion between simplified Chinese and traditional Chinese are the main causes for extracting wrong terms from the topics for indexing and retrieval.

For example, in topic 22 (合法經營，起亞汽車，意見) the word “起亞” (Kia) is not registered in the CBJ traditional Chinese dictionary used for tokenization. Consequently, it was interpreted as verb “起” (rise, occur) and noun “亞” (Asia). Another example of missing lexical entries is topic 39 (外勞，驅逐，人權), in which “外勞” (foreign worker) is parsed as person name “外” (foreign) and verb “勞” (work). As a result, the system had low score for these topics.

In topic 26 (中國，反應，台灣，外交關係), the CBJ traditional Chinese dictionary has "關系", but doesn't have the correct word "關係". "關係 (relation)" was parsed as noun "關 (checkpoint, custom)" and unknown word "係". This is due to the error in character convert between Simplified Chinese and Traditional Chinese characters.

The above errors suggest that we need to develop a better conversion algorithm between simplified Chinese characters and traditional Chinese characters, and that lexicon-free approaches, such as n-gram based indexing should be incorporated into the indexing and retrieval processes.

3.2 Bilingual CLIR track

For bilingual CLIR, we adopted query translation as the means for bridging the language gap between the query language and the document language. For Japanese-English retrieval, first, the Japanese topics were parsed into words and phrases with Japanese NLP module in CBJ. Then the terms were translated into English. For Chinese-English retrieval, we used a part-of-speech tagger to get the terms, without phrase construction, and then translate the Chinese terms into English.

For both types of runs, the English document collection was indexed as described in section 3.1.2. Once queries were translated from the source language to the target language English, English documents were retrieved the same way as in English monolingual retrieval as described in section 3.1.2.

3.2.1 Japanese-English Retrieval

The Japanese-English translation lexicon was a combination of several lexicons: the EDR Japanese-English bilingual dictionary[1], the EDICT and ENAMDICT dictionaries[2], a commercial front-end input lexicon ATOK developed by Justsystem[3], a lexicon extracted from the Yomiuri parallel corpus [9] via a translation pair extraction tool, and a list of famous Chinese person names collected from the WWW.

We compared CBJ and CLJ for Japanese-English retrieval. In addition to a very comprehensive translation lexicon and language independent features such as pseudo-relevance feedback, CBJ employs CLIR specific techniques as described in section 2.4 for obtaining and balancing the translations. For post-translation feedback, the Rocchio method was used for extracting the top 20 terms from the top 20 subdocuments with the desc topics and top 30 terms from the top 10 documents with the title topics. Feedback terms were merged with original query terms based on formula (8).

Run	feature	Avg prec	Recall
J-E-T-cbj	Rocchio Formula (8)	0.2131	3688/5866
J-E-T-clj	Prob2 Formula (9)	0.2125	2965/5866
J-E-D-cbj	Rocchio Formula (8)	0.262	3885/5866
J-E-D-clj	Prob2 Formula (9)	0.2427	3733/5866

Table 5: Japanese-English Retrieval, Rigid

[1] http://www.iijnet.or.jp/edr/E05JEBIL.txt

[2] http://www.csse.monash.edu.au/~jwb/edict.html

[3] www.atok.com

For CLJ runs, we took the translated expanded query from CBJ and conducted another round of query expansion. Here, Prob2 was used for extracting feedback terms, which were then merged with the CBJ-query terms based on formula (9).

Table 5 shows the results with both systems for title and description queries. The results show that with additional round of feedback on top of CBJ output, CLJ was not able to improve the retrieval performance further.

3.2.2 Chinese-English Retrieval

The Chinese-English translation lexicon was based on CEDICT version 3[1], which has a total of more than 51,400 entries, expanded with a lexicon of technical terms of about 1400 entries collected from the WWW, and a list of names of about 1000 famous person names. The list of famous persons was constructed by converting the famous person name lexicon in the Japanese-English lexicons described in the previous section. We call this lexicon expanded CEDICT or expCEDICT. For the Chinese-English retrieval task, we further expanded the expCEDICT lexicon by adding translations of multi-word term translations. We used CLARIT for English retrieval for all the experiments reported here.

Our approach to translating multi-word terms for Japanese and Chinese is based on previous work for European languages [5]. The method, similar to [8], involves generating possible candidate translations using a bilingual dictionary and then attesting the candidates, ranking them by their frequency in a reference corpus. The steps are as follows:

- Extract all multiword terms using NLP modules from corpus
- Find those terms unknown to CEDICT
- Among those find those terms whose parts are known to CEDICT
- Generate English translation candidates for these phrases by translating their subparts translation candidates in NTCIR-3

With the NTCIR-3 Chinese evaluation corpus, we extracted and validated 236,652 multi-word Chinese terms and their corresponding translations.

We used the Chinese-English retrieval track as a small-scale experiment on the effectiveness of the additional translations of the multi-word terms. For this experiment, we used the Rocchio method for post-translation query expansion, by extracting the top 30 terms from the top 20 documents. Multi-word terms were down-weighted to 0.2. We did not use translation disambiguation for choosing the best translations, as we did not have time to adapt our existing disambiguation module to deal with multi-word terms.

Table 6 shows the retrieval results for both the title and description topics. The results showed that by expanding the base lexicon with automatically extracted translations of phrases, retrieval performance can be improved slightly, but the improvement is not significant. Description based runs had lower precision scores than the title based runs. This is probably due to the increasing noise in translation when more terms were translated, which suggests that translation disambiguation should be incorporated into the process.

Run	feature	Avg prec	Recall
C-E-T-1	expCedict	0.1627	3041/5866
C-E-T-2	expCedict MWE	0.166	3378/5866
C-E-D-1	expCEDICT	0.1552	3103/5866
C-E-D-2	expCedict MWE	0.1557	3184/5866

Table 6: Chinese-English Retrieval, Rigid

4. Post-NTCIR-4 Analysis

After NTCIR-4 workshop, we have classified the errors in our NTCIR-4 Japanese-English retrieval submission into types as shown in Table 7. Table 8 presents the distributions of the error types for both our T-run (title) and D-run (description).

As one can see from Table 8, disambiguation (E2.1) was the major cause of error, followed by pseudo-relevance feedback (E2.2). As mentioned in section 2.4, the disambiguation module currently implemented in our system simply checks whether a translation appears in any corresponding sentences in parallel corpora, without checking the semantic validity of the translations in relation with other terms or translations, or checking the strengths of associations between the query terms and their translations. As a result, the disambiguation process was not always effective in filtering out incorrect translations. PRF in the CBJ-based cross-language retrieval system used normalized merging strategies (section 2.2) and lacked effective tuning of weights of original query terms and expansion terms. As a result, the greater the number of expansion terms, the bigger their effect on document scoring. We have observed in our monolingual experiments that the scaled merging strategy is overall more robust, which we should consider implementing in our next version of CLIR system.

As part of our system analysis, we examined the contributions of the different strategies implemented in our CLIR system as previously described in section 2.4. The examined strategies included multi-word term down-weighting and pseudo-relevance feedback, which are general strategies for information retrieval, and disambiguation, translation structuring, and suppression of low distribution terms, which are CLIR-specific strategies.

[1] http://www.ldc.upenn.edu/doc/LDC2002L27/readme.txt

As an example, Table 9 presents the contribution of the individual strategies when used alone with the bilingual baseline for the T-runs based on the Rigid evaluation. We compare the results against both the English monolingual baseline runs (with an average precision of 0.3412 based on Rigid evaluation) and the bilingual baseline runs (designated as Baseline).

Compared with the baselines, the T-run with combined strategies achieved 62.5% of monolingual baseline and 141.2% of bilingual baseline based on Rigid evaluation. We make two observations about the results. First, when the strategies are combined, we observe that the system performance achieved by the integrated combination of strategies performed better than simply summing up the contributions of the individual strategies. For instance, summing up the contribution of each component in Table 9 produced a cumulative improvement of 0.0347, while the integrated system produced an improvement of 0.0622. Second, while the strategies "disambiguation", "RPF", and "suppress low distribution terms" improved the MAP scores, the strategies "multi-word term down-weighting" and "query structuring" caused performance degradation. Yet the integrated system recovered those losses and actually achieved a performance improvement. This suggests that the interaction of the strategies is synergetic, overcoming any negative effects of the individual strategies.

Our follow-up experiments show that multi-word term down-weighting favorably interacts with other strategies, resulting in general performance improvements. Query structuring, however, causes complications when combined with other strategies, generally reducing performance. In our future work, we plan to explore whether other approaches of query structuring such as OR, or AVG (average) would improve its contribution.

In contrast, with the longer D-run topics, we observed that all the investigated strategies improved performance over the baselines, including the two strategies "multi-word term down-weighting" and "query structuring" that degraded performance for the T-runs when individually applied. In fact, the best two contributors included pseudo-relevance feedback and query structuring.

5. Summary

In NTCIR-4, we conducted monolingual and bilingual experiments to compare three retrieval systems under development at Justsystem Corporation in Japan and Clairvoyance Corporation in the USA. With the experiments at NTCIR-4, we evaluated the commercial and research versions of our retrieval systems. The CLIR experiments have shown promise for some of the newly developed techniques, such as scaled merging during feedback.

After NTCIR-4, we conducted more Japanese-to-English retrieval experiments to evaluate the contributions of individual strategies. With the shorter Title topics, PRF and disambiguation have been major performance enhancers, but query structuring and multi-word term down-weighting have been shown to affect retrieval performance negatively. Examination of multi-word term down-weighting with other strategies showed positive interaction between multi-word term down-weighting and other strategies, suggesting synergy in combinations of strategies. However, query structuring with other system component did not show any positive interaction. Query structuring affects how weights are computed for multi-word terms and their subterms, and how multiple translations are weighted. Given the positive effect of query structuring reported in previous work [6] and its positive effect on longer topics, in our future work, we will explore more other methods of query structuring.

References

[1] Evans, D.A., and Lefferts, R.G.: CLARIT–TREC Experiments. Information Processing and Management. 31(3) (1995) 385–395

[2] Fujita, S. 1999. Notes on Phrasal Indexing: JSCB Evaluation Experiments at NTCIR AD HOC" In Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition.

[3] Kishida, K., Chen, K. Lee, S., Kuriyama, K., Kando, N., Chen, H., Myaeng, S. H., Eguchi, K. 2004. Overview of CLIR Task at the Forth NTCIR Workshop. In this volume.

[4] Oyama, K, Ishida, E., and Kando, N. (eds) 2003. NTCIR Workshop 3: Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering.

[5] Grefenstette, G. 1999. The WWW as a Resource for Example-based MT Tasks. In ASLIB'99 Translating and the Computer 21

[6] Pirkola, A., Puolamaki, D., Jarvelin, K. 2003. Applying query structuring in cross-language retrieval. Information Processing and Management, Vol 39(3) 391–402

[7] Robertson, S.E. and Walker, S. 1994. Some Simple Effective Approximations to the 2-Poisson model for Probabilistic Weighted Retrieval. In Proceedings of the ACM SIGIR Conference, pp. 232–241.

[8] Sato, S. and Nagao, M. 1990. Towards memory-based translation. In H. Karlgren, editor, In Proceedings of COLING'90, pages 247–252, Helsinki

[9] Utiyama, M., Isahara, H. Alignment of Japanese-English News Articles and Sentences. In IPSJ SIGNotes Fundamental Infology Abstract No. 068–003

Error Type	**Subtypes**	**Description**
Failure to obtain correct translations (E1)	Missing translations (E1.1)	Query contains word(s) that do not have translations from the bilingual dictionaries.
	Incorrect translations (E1.2)	Translation(s) in the bilingual dictionaries are not correct.
	Interfering stop words (E1.3)	Failure to generation multi-word terms results from elimination of stop words as parts of multi-word terms.
	Interfering one-character suppression (E1.4)	Suppression of one-character subterm(s) results in missing translations of some important subterms, such as one-character person names.
	Incomplete phrasal types (E1.5)	Phrase type is limited to NPs during indexing.
Suboptimal tuning of system (E2)	Disambiguation (E2.1)	Disambiguation does not work well.
	PRF (E2.2)	Pseudo-relevance feedback (PFR) does not work well.
	Term weighting (E2.3)	Important terms (e.g., proper nouns) have high distributions so their weights become low.
Misc others		Other errors such as morphological analysis error, not enough context for disambiguation.

Table 7: Error types in Japanese-English retrieval runs

Run ID	**Error Types**								
	Failure to obtain correct translations					Suboptimal tuning of system			Misc others
	E1.1	E1.2	E1.3	E1.4	E1.5	E2.1	E2.2	E2.3	
J-E-T-cbj J-E-D-cbj	2 1	1 4	0 1	1 1	6 0	16 17	15 10	2 4	2 2
Total	3	5	1	2	6	33	25	6	4

Table 8: Distribution of error types in Japanese-English retrieval runs (T-run and D-run)

Condition	MAP	Diff.	Mono	Recall	Diff.
Baseline	0.1509	-	44.2%	3108/5866	-
+disambiguation	0.1715	+13.7%	50.3%	3280/5866	+172
+multi-word term down-weighting	0.1468	-2.7%	43.0%	2918/5866	-190
+PRF	0.1708	+13.2%	50.1%	3246/5866	+138
+query structuring	0.1328	-12.0%	38.9%	2975/5866	-133
+suppress low distribution terms	0.1673	+10.9%	49.0%	3184/5866	+76
+all components (J-E-T-cbj)	0.2131	+41.2%	62.5%	3688/5866	+580

Table 9: Japanese-English Retrieval, T-run, Rigid evaluation