Notes on the Limits of CLIR Effectiveness
NTCIR-2 Evaluation Experiments at Justsystem

Sumio FUJITA
JUSTSYSTEM Corporation
Brains park, Tokushima, 771-0189 JAPAN
Sumio_Fujita@justsystem.co.jp

Abstract

NTCIR-2 evaluation experiments at the Justsystem site are described with a focus on comparative study of CLIR effectiveness with monolingual retrieval effectiveness of the same retrieval engine.

Experiments on the effects of phrasal translation, indexing of translated phrasal terms, pre-translation feedback and parallel documents feedback in diverse retrieval settings, are reported.

The results show that pre-translation pseudo-relevance feedback is always effective.

We confirmed that phrasal translation is effective and crucial even the retrieval system does not use phrasal indexing units.

Keywords: *CLIR, query vector translation, pre-translation feedback, parallel documents feedback, phrasal translation, pseudo-relevance feedback, vector space model.*

1. Introduction

In the cross-lingual information retrieval (CLIR) runs, our query vector translation approach is always symmetrical for J-E retrieval and E-J retrieval utilizing J-E and E-J bilingual dictionaries respectively that were built automatically from a parallel corpus of the ntc1 data set.

Pre-translation query expansion is done in JSCB5-JSCB8 and post-translation query expansion is done in all cross-lingual runs. Another query expansion technique was tried utilizing parallel part of ntc1 data set but this does not improve the result so much, mainly because of lack of enough coverage for subject topics in parallel corpus.

After the queries in source language are thus translated and expanded, the same retrieval process as monolingual baseline runs is executed against the target language text database.

Here, one question arises namely, given monolingual retrieval procedures against both source language retrieval and target language retrieval, what is the practical (not theoretical) limit of CLIR effectiveness above which it might be very difficult to improve the performance.

2. Limits of CLIR Effectiveness

An interesting point of ntc2 test set is that both parallel text collections and parallel topic sets are provided as well as relevance judgement, so that the cross-language retrieval performance is compared with both source and target language monolingual baseline. This can be an ideal test set for evaluating some CLIR related techniques and their limits in laboratory experiments.

2.1 Monolingual baseline

Monolingual retrieval performance against the target language is generally acknowledged as the practical limit. Normally the cases where CLIR effectiveness surpasses the monolingual one are either that the monolingual run is not very effective [11] or some test set specific knowledge that furnishes the system with more information about relevance than monolingual topic description, are available.

If the query translation procedure is so good that it outputs translated queries comparable to the original topic description of the target language side, CLIR without pre-translation feedback might be equivalent to monolingual retrieval in the target language although such good translation is not realistic in automatic query construction.

If pre-translation feedback or parallel collection based method is utilized, further improvement might be possible.

Our questions to pose are:

1) How good the query translation output should be, in order that the CLIR effectiveness is comparable to the monolingual baseline.

2) If the query translation output is so good that the CLIR effectiveness is comparable to the monolingual baseline, is further improvement possible by pre-translation query expansion or other techniques? and what are the conditions in which it is possible?

Another practical question that may summarize the following two is:

3) What is the robust retrieval processing that works pretty well even the query translation output is not so good? Since perfect automatic query translation is not realistic, how can a retrieval process supplement the information lost by the translation? How can a retrieval process neutralize effects of noises introduced by the translation?

2.2 Measures of Query Translation Quality in view of "Relevance"

In order to answer the previous questions, we need to measure the quality of translated query vectors somehow with regards to their containing information about "relevance" but not to comprehensibility by human readers.

In order to measure the information about relevance given by the occurrence of each single term, log(p(occ|rel)/p(occ)), that is equivalent to the mutual information MI(occ;rel) is utilized [8][6].

$$\log\left(\frac{p(occ \mid rel)}{p(occ)}\right) = \log\left(\frac{p(occ, rel)}{p(occ)\,p(rel)}\right) \quad (1)$$

Another measure we tried is the probabilistic term weighting proposed by Robertson and Sparck-Jones [9].

$$w(t) = \log\left(\frac{P(occ \mid rel)(1 - P(occ \mid \overline{rel}))}{P(occ \mid \overline{rel})(1 - P(occ \mid rel))}\right) \quad (2)$$

For the whole query, a simple sum of each term weight was utilized.

3. System Description

For the NTCIR-2 CLIR experiments, we utilized the engine of Justsystem ConceptBase Search™ version 2.0 as the monolingual base system except that Japanese morphological analysis module of version 1.2, that is the same as our NTCIR-1 system, is utilized so that the retrieval performance is comparable with our NTCIR-1 system [4]. A dual Pentium III™ server (670MHz) running Windows NT™ server 4.0 with 1024MB memory and 136GB hard disk is used for experiments.

The document collections are indexed wholly automatically, and converted to inverted index files of terms.

3.1 CLIR Evaluation System

As shown in Figure 1, our query translation CLIR approach is symmetrical for J-E retrieval and E-J retrieval utilizing J-E and E-J bilingual dictionaries respectively that were built automatically from the parallel keyword fields of the ntc1-je1 data set.

Japanese and English monolingual information retrieval engines input the topic description in each language, parse it and send query vectors to the query translation module. Each query translation module translates a query vector into the target language by referring to a bilingual dictionary and sends the translated query vector to the target language IR engine.

The query vector is expanded before / after translation by a pseudo relevance feedback procedure when applied.

3.2 Query Translation

The bilingual dictionaries are built from ntc1-je1 parallel test collection extracting keywords from KYWD and KYWE fields as described in [2].

This field provides many phrasal keywords as well as single words and they are similarly registered as entry words. Thus extracted parallel keyword lists are organized in both J-E dictionary and E-J dictionary of 1,439,992 entries.

Query vectors are translated by referring to these dictionaries. The most frequent keyword pair is used when more than two translation pairs are found.

3.3 Term Extraction

Queries and documents in target databases are analyzed by the same module that decomposes an input text stream into a word stream and parses it using simple linguistic rules, in order to compose possible noun phrases. Extracted units are single word nouns as well as simple linguistic noun phrases that consist of a sequence of nouns or nouns preceded by adjectives.

3.4 Vector Space Retrieval

Each document is represented as a vector of weighted terms by tf*idf in inverted index files and the query is converted in similar ways.

Similarity between vectors representing a query and documents are computed using the dot-product measure, and documents are ranked according to decreasing order of RSV. In post submission experiments, OKAPI BM25 function is utilized as TF part of weighting function [10].

3.5 Phrasal Indexing and Weighting

Our approach consists of utilizing noun phrases extracted by linguistic processing as supplementary indexing terms in addition to single word terms contained in phrases. Phrases and constituent single terms are treated in the same way, both as independent terms, where the frequency of each term is counted independently based on its occurrences.

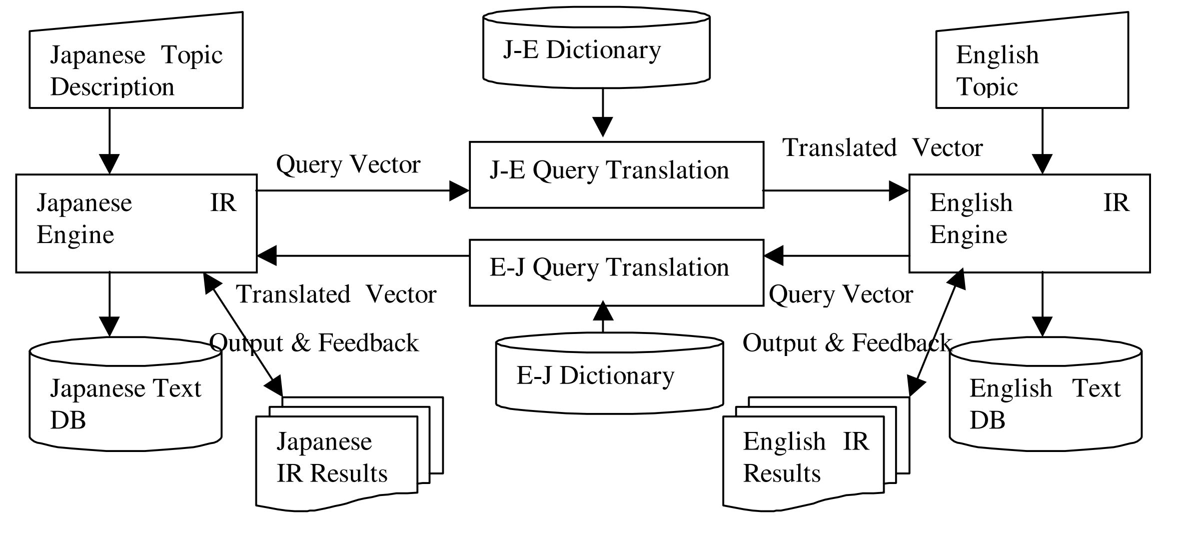

Figure 1: J-E & E-J CLIR Evaluation System

3.6 Pre-translation and Post-translation Pseudo-Relevance Feedback

Automatic feedback strategy using pseudo-relevant documents is adopted for automatic query expansion.

The system submits the first query generated automatically from topic descriptions against the source or target language database, and considers the top n documents from relevant ranking list as relevant.

The term selection module extracts salient terms from these pseudo-relevant documents and adds them to the query vector.

Then the expanded query vector is submitted against the target database again and the final relevance ranking is obtained.

In CLIR runs, such a feedback is applied before query translation executing a pilot search against the source language database as well as after the query translation.

The whole retrieval procedure is as follows:

1) Automatic initial query construction from the source language topic description
2) 1st pilot search submitted against the source language database
3) Term extraction from pseudo-relevant documents and feedback
4) Query vector translation using a bilingual dictionary
5) 2nd pilot search submitted against the target language database
6) Term extraction from pseudo-relevant documents and feedback
7) Final search against the target language database to obtain the final results

3.7 Term Selection

Each term in example documents are scored by some term frequency and document frequency based heuristics measures described in [3].

The terms thus scored are sorted in decreasing order of each score and cut off at a threshold determined empirically.

In effect, the following parameters in feedback procedures should be decided:

1) How many documents to be used for feedback?

2) Where to cut off ranked terms?

3) How to weight these additional terms?

These parameters are carefully adjusted using ntc-1 queries (topic 31-83), ntc-1 data set and their relevance judgement provided by NACSIS.

3.8 Parallel Corpus Usage

Because some part of target data sets are parallel in the document level, we first tried to utilize these parallel documents as the resource of CLIR.

The source language query is submitted against source language database and obtained the top n documents. Their counterparts in the target language database are utilized for term extraction in order to construct the target language query vector.

This strategy worked perfectly well for ntc-1 test set and the mean average precision of J-E retrieval of ntc-1 set reaches 36.80% (long query, rel1 judgement, 39 topics of ntc1 test set) without using any query translation method. This seems to be better than any NTCIR-1 CLIR systems.

Run tag	Lang	Qlen	PFB	Avg. Prec	R-Prec
JSCB1	JJ	S	Yes	0.3404	0.3555
JSCB2	JJ	L	Yes	0.4092	0.4159
JSCB3	EE	S	Yes	0.3193	0.3104
JSCB4	EE	L	Yes	0.4023	0.3926
JJS-BASE	JJ	S	Yes	0.3611	0.3783
JJL-BASE	JJ	L	Yes	0.4020	0.4196
EES-BASE	EE	S	Yes	0.3637	0.3510
EEL-BASE	EE	L	Yes	0.4137	0.3993

Table 1: Performance of official runs and other runs in monolingual baseline runs (S&A judgement)

Among 332,918 Japanese documents in the ntc-1 collection, only 181,485 are known to have their translation in the English collection. The fact that such small coverage (only 55% of the whole collection are parallel) of parallel corpus made this performance gave us a wrong impression such that the method is generally applicable.

But this does not work at all for ntc-2 test collections where the portion of parallel documents is much smaller (25%).

4. Experiments

The experiments are designed to evaluate the limits of the effectiveness of query translation based CLIR given a monolingual IR system with regards to different query types.

Short runs utilize only "description" fields of topic description and long runs utilize all the fields except "field" fields.

4.1 Monolingual Baseline Runs

As the baseline for CLIR runs, we submitted Japanese (short JSCB1, long JSCB2) and English (short JSCB3, long JSCB4) monolingual runs both with short queries and long queries. Japanese runs utilize essentially the same system as our runs in NTCIR-1 but the parameters were re-tuned with ntc1 test set topic 31-83. The very similar approach was utilized as well for English monolingual retrieval.

Pseudo relevance feedback is applied both in short and long queries.

Table 1 shows the results of official runs and other post submission runs.

Since the performance seems to be not very well in comparison with our NTCIR-1 results, we tried to re-run JJ experiments with different parameter sets. First, we tried BM25 TF that we found to be more effective than SMART TF with the collection consisting of documents of diverse length, in our TREC-9 main web experiments [7].

After trying some parameters of BM25 function, 37.64% of avg.prec. in short query run and 43.10% in long query run are achieved.

Second, We tried Japanese part of E-J CLIR system parameter setting and its monolingual performance was 36.11% of avg. prec. as shown as JJS-BASE in Table 1. EES-BASE is its counterparts of E-E monolingual retrieval. JJL-BASE and EEL-BASE are the same thing but with long queries. These CLIR parameter settings perform much better than our official monolingual runs especially with short queries.

It seems that our official monolingual runs are over-tuned with ntc1 test set and parameters are over-fitted to the training test set and did not perform well with ntc2 test set.

In the following CLIR experiments, JJS-BASE, JJL-BASE, EES-BASE and EEL-BASE are utilized as monolingual baseline.

4.2 J-E and E-J CLIR Runs

Table 2 shows our CLIR official runs and other runs.

"PTFB" refers to "Pre-translation feedback" and "Par" refers to "parallel corpus usage" that consists

Lang (run tag)	Q-len	PTFB	Par	AvgPrec	R-Prec
JE(JSCB5)	S	Yes	Yes	0.3026	0.3176
JE	S	Yes	No	0.3065	0.3138
JE (JSCB9)	S	No	Yes	0.2597	0.2922
JE (JES-BASE)	S	No	No	0.2634	0.2854
JE(JSCB6)	L	Yes	Yes	0.3885	0.3867
JE	L	Yes	No	0.3786	0.3789
JE(JSCB10)	L	No	Yes	0.3642	0.3792
JE(JEL-BASE)	L	No	No	0.3519	0.3680
EJ(JSCB7)	S	Yes	Yes	0.2651	0.2874
EJ	S	Yes	No	0.2543	0.2764
EJ(JSCB11)	S	No	Yes	0.2297	0.2518
EJ(EJS-BASE)	S	No	No	0.2151	0.2356
EJ(JSCB8)	L	Yes	Yes	0.3234	0.3464
EJ	L	Yes	No	0.3044	0.3268
EJ(JSCB12)	L	No	Yes	0.3025	0.3304
EJ(EJL-BASE)	L	No	No	0.2805	0.3098

Table 2: Performance of official runs and other runs in CLIR experiments (S&A judgement)

of utilizing target language counterparts of pseudo relevant documents found in source language collections for term extraction. Post-translation feedback was applied all runs described in the table.

Pre-translation feedback seems to be always effective and improvement of as big as 16.5% is observed.

Parallel corpus usage makes some small improvement except in J-E short retrieval where some degradation was observed.

Table 3 shows average number of terms in original queries or translated queries. Translated query vectors are sometimes longer than original queries because translated phrasal terms are decomposed into single word terms again in the target language side.

Table 4 shows examples of original query vectors and translated query vectors.

The translation processes introduced many noisy terms or wrongly translated terms but they are mainly very frequent single word terms and their IDF values are low.

Instead phrasal translation provides high IDF and possibly useful terms.

Once phrasal terms in the source language query are translated, translated terms, which are normally phrases as well, can be used either as supplemental phrasal indexing units or only decomposed single words.

Table 5 compares the following different treatment of phrasal translation with monolingual baseline of target language side and with CLIR baseline runs listed in Table 2.

1)MBASE: Monolingual BASEline

Monolingual runs corresponding to the target language of CLIR runs but pre-translation feedback and parallel corpus usage are applied where designated. These runs are considered as CLIR runs where original query translation is ideal.

2)CBASE: CLIR BASEline

CLIR runs including official runs.

3)DSW: Decomposed Single Words

Phrases and sub-phrases in source language query vectors are translated but translated terms (normally phrases) are decomposed into single words and only these single words are added in the target vector.

4)NO-SUBPTR: NO SUB Phrase TRanslation

Phrases in source language query vectors are translated except sub-phrases that are embedded in detected phrases.

5)NO-PTR: NO Phrase Translation

Phrases in source language query vectors are not translated but only single word terms are translated.

In the runs DSW, NO-SUBPTR and NO-PTR, phrasal terms added by feedback procedures are processed as usual and only phrasal terms extracted from source language topic description are eliminated.

Run	JJ Short	JJ Long	EE Short	EE Long	JE Short	JE Long	EJ Short	EJ Long
words	5.49	32.61	5.45	31.33	6.90	34.22	8.24	46.14
phrases	1.82	12.73	1.20	12.65	1.39	9.65	2.33	13.06
total	7.31	45.35	6.65	43.98	8.29	43.88	10.57	59.20

Table 3: Average number of terms in the initial queries or translated queries

Lang	Term vectors
JJ	ｂ 型 肝炎 ワクチン,遺伝子 工学 的 手法,ワクチン,肝炎,工学,ｂ,遺伝子,手法,開発
JE	engineer method, vaccine, hepatitis, engineer, b, gene, method, development,
EE	hepatitis b vaccine, genetic engineer technique, hepatitis b, vaccine, genetic, engineer, technique, development
EJ	ｂ 型 肝炎 ワクチン,遺伝子 工学,工学 的 予測 手法, 予測,手法,遺伝子,手 技,手,工学,ｂ 型 肝炎,肝炎,ｂ, ワクチン,ｇ ａ,遺伝, 技術者,技術者 教育,教育,発生,開発

Table 4: Examples of original term vectors and translated vectors (topic 101)

Lang (run tag)	Q-len	PTFB	Par	MBASE	CBASE	DSW	NO-SUBPTR	NO-PTR
JE(JSCB5)	S	Yes	Yes	0.3381	0.3026	0.3012	0.2929	0.2571
JE	S	Yes	No	0.3511	0.3065	0.3072	0.2921	0.2546
JE (JSCB9)	S	No	Yes	0.3525	0.2597	0.2640	0.2433	0.1918
JE (JES-BASE)	S	No	No	0.3637	0.2634	0.2662	0.2420	0.1750
JE(JSCB6)	L	Yes	Yes	0.4077	0.3885	0.3821	0.3684	0.3021
JE	L	Yes	No	0.4087	0.3786	0.3777	0.3600	0.3010
JE(JSCB10)	L	No	Yes	0.4136	0.3642	0.3563	0.3540	0.2763
JE(JEL-BASE)	L	No	No	0.4137	0.3519	0.3456	0.3427	0.2618
EJ(JSCB7)	S	Yes	Yes	0.3707	0.2651	0.2679	0.2631	0.2384
EJ	S	Yes	No	0.3654	0.2543	0.2599	0.2380	0.2166
EJ(JSCB11)	S	No	Yes	0.3645	0.2297	0.2255	0.2176	0.1767
EJ(EJS-BASE)	S	No	No	0.3611	0.2151	0.2008	0.1939	0.1408
EJ(JSCB8)	L	Yes	Yes	0.4267	0.3234	0.3213	0.3136	0.2533
EJ	L	Yes	No	0.4185	0.3044	0.3151	0.2879	0.2355
EJ(JSCB12)	L	No	Yes	0.4152	0.3025	0.2858	0.2895	0.2324
EJ(EJL-BASE)	L	No	No	0.4020	0.2805	0.2655	0.2615	0.2060
Average				0.3631	0.2818	0.2789	0.2683	0.2188
%Change from CBASE				+28.9%	100%	-1.0%	-4.8%	-22.4%
%Change from MBASE				100%	-22.4%	-23.2%	-26.1%	-39.7%

Table 5: Mean average precision of official runs and other runs in CLIR experiments (S&A judgement)

Even translated phrasal terms themselves are not utilized, their constituent single word terms work well and retrieval effectiveness does not change so much.

But if the phrasal terms or their sub-phrases extracted from the original topic description are not used as translation units, retrieval effectiveness is degraded so much. Even their constituent single words are translated as word unit, it does not help since word by word translation is not able to properly translate multi-word concepts.

From these runs, we can confirm that phrases are normally very good translation units although phrases themselves are not as good as single words as indexing units.

Ballesteros et al. suggested that phrasal translation can greatly improve effectiveness but improvements are more sensitive to the quality of the translations than single words. They observed that one poor translation can counteract any improvement gained by the correct translation of several phrases [1].

Our observation is that the sensitivity to translation quality of phrases is a natural consequence of the fact that phrases are normally over weighted. Even in monolingual retrieval cases, some bad phrasal terms probably harm the performance more than their constituent single words [5]. If the translated phrases are decomposed into their constituent single words and down-weighted properly, they should behave just like translated single words as word-by-word basis.

It seems that the CBASE runs are 22.4% worse than MBASE runs where original query translation is assumed to be ideal. Although feedback strategies always helps and make 15% of improvement at best, a perfect query translation is more desirable for better retrieval effectiveness.

It is also worth noting that in E-J retrieval pre-translation feedback makes some small improvement where the query translation is ideal (this is J-J monolingual retrieval with E collection feedback), while it is not the case in J-E retrieval.

If we compare CLIR performance with monolingual excluding any feedback influences, JES-BASE is –27.6% from EES-BASE, JEL-BASE –14.9% from EEL-BASE, EJS-BASE –40.4% from JJS-BASE and EJL-BASE –30.2% from JJL-BASE respectively.

These results suggest that J-E CLIR is better than E-J CLIR and long query is better than short query where no source language side feedback is applied. The reason why the long query is better than short in CLIR seem to be the same as the monolingual retrieval case, where longer queries have normally more information about relevance that can neutralize the effects of noisy information.

For the inequality between degradation of J-E and E-J CLIR, we can not help assuming that E-J query translation is much worse than J-E translation, which is not difficult to imagine from the example of

Table 4 where E-J translation introduces many noisy terms while J-E does not, even though our dictionaries are symmetrical.

4.3 Quality of Translated Query Vectors

Now we return to the first question posed in the chapter 2 of this paper.

As our CLIR systems are symmetrical as seen in the Figure 1, target language parts of CLIR runs EJS-BASE, EJL-BASE, JES-BASE and JEL-BASE are equivalent to monolingual runs JJS-BASE, JJL-BASE,EES-BASE and EEL-BASE respectively, except the query vectors are translated by a translation procedure in CLIR runs while directly constructed from the topic description in monolingual runs. Information needs are almost the same even the written language, length of the description and creating process of search requests are different according to each run contexts. They are addressing the same abountness even the query vectors are different. For each aboutness, two types (S&A and S&A&B) of relevance judgement against two test collections (Japanese/English) are provided.

The problem is how to measure the information about relevance that different query vectors of 8 runs may provide to the retrieval system. As described in the chapter 2, mutual information and probabilistic term weighting are utilized and shown in Table 6 and Table 7. Each query vector is measured either by sum of mutual information between occurrence of each term and relevance or by sum of probabilistic term weighting of each term.

Pearson's correlation coefficients between these measures and search effectiveness of topic-by-topic basis are 0.36 to 0.57 in short query runs and 0.16 to

Run tag (REL1)	JJS-BASE	JJL-BASE	EES-BASE	EEL-BASE	JES-BASE	JEL-BASE	EJS-BASE	EJL-BASE
CORR(MI,AP)	0.5035	0.3694	0.4865	0.1687	0.4371	0.3040	0.3669	0.3002
AVG(SUM(MI))	18.72	122.98	24.24	103.00	22.93	83.08	23.17	83.19
CORR(PW,AP)	0.5703	0.4104	0.4981	0.1950	0.4625	0.3273	0.4061	0.3353
AVG(SUM(PW))	33.84	137.19	28.02	111.93	26.85	90.97	28.52	94.29
Mean Avg. Prec.	0.3611	0.402	0.3637	0.4137	0.2634	0.3519	0.2151	0.2805

Table 6: Correlation between measure of query vector quality and retrieval effectiveness (S&A rel1 relevance judgement)

Run tag (REL2)	JJS-BASE	JJL-BASE	EES-BASE	EEL-BASE	JES-BASE	JEL-BASE	EJS-BASE	EJL-BASE
CORR(MI,AP)	0.4475	0.3347	0.4271	0.2433	0.5551	0.3735	0.3856	0.3490
AVG(SUM(MI))	27.51	125.61	23.73	106.82	23.06	84.84	23.32	85.19
CORR(PW,AP)	0.5138	0.3741	0.4588	0.2665	0.5859	0.4111	0.4271	0.3838
AVG(SUM(PW))	33.53	139.85	27.34	116.17	26.90	92.92	28.30	96.04
Mean Avg. Prec.	0.3553	0.4353	0.3214	0.4009	0.2387	0.3391	0.1991	0.3017

Table 7: Correlation between measure of query vector quality and retrieval effectiveness (S&A&B rel2 relevance judgement)

CORR(x,y): Pearson's correlation coefficient between x and y

MI: mutual information between occurrence of a term and relevance given either rel1 or rel2 relevance judgement

PW: Robertson and Sparck-Jones' probabilistic term weighting given either rel1 or rel2 relevance judgement

SUM: sum of each term measure in a query vector

AP: Average precision of the retrieval result of a topic

AVG: average of topic 101 to 149

Mean Avg. Prec.: Mean average precision of the run against either rel1 or rel2 relevance judgement

0.41 in long query runs.

These measures do not take term frequency into account so that the correlation coefficient is not so high especially in long query runs where term weighting factors other than presence/absence of terms affect more the effectiveness. Even though, correlation coefficient between AVG(SUM(MI)) and mean average precision of each run accounts for 0.655, and AVG(SUM(PW)) 0.666. It may not be surprising that the two measures are almost equivalent since correlation coefficient between AVG(SUM(MI)) and AVG(SUM(PW)) is 0.998. These two different term precision measure seem to behave very similarly in practical size of collections.

With S&A judgement, AVG(SUM(PW)) of JES-BASE is –4.2% from EES-BASE while EJS-BASE –15.7% from JJS-BASE. JEL-BASE is –18.7% from EEL-BASE while EJL-BASE –31.2% from JJL-BASE. This supports the assumption that E-J translation was much worse than J-E in our system.

Another question in suspense is that the query vector of JES-BASE is so good that we tend to expect that our J-E short run might be much better than JSCB5. More observation is needed on this issue.

5. Conclusions

NTCIR-2 experiments at Justsystem group are described.

The following answers to the three questions posed in the chapter 2 are drawn from these experiments:

1) Query translation quality should be ideally perfect in order that the CLIR effectiveness is comparable to the monolingual baseline. Loss of information about relevance contained in original queries causes big degradation.

2)Even the query translation is ideally perfect, pre-translation feedback and other techniques may improve effectiveness further more but not always.

3)Pre-translation feedback clearly helps to compensate for lost information by the translation. Phrasal translation is crucial for better query translation irrespective of usage of phrasal terms in the retrieval engine. Translated phrases may used as decomposed single words for the sake of robustness.

On the other hand, we need more experiments as well as careful observation on the correlation between the quality of query vectors and retrieval effectiveness. It is also interesting to introduce more sophisticated measures that may take account of term frequency and inter-term dependency.

For the future work, it is desirable to compare the more elaborated query translation techniques and to evaluate their outputs by means of both "query vector quality measures" and of traditional precision/recall measures of run results.

References

[1] Ballesteros, L. and Croft, B. Phrasal Translation and Query Expansion Techniques for Cross—Language Information Retrieval, in Proceedings of the 20th Annual International ACM SIGIR Conference(Philadelphia, July 1997), ACM Press, 84-91.

[2] Chen, A., Gey, C. F., Kishida, K., Jiang, H. and Liang, Q. Comparing multiple methods for Japanese and Japanese-English text retrieval, in Proceedings of NTCIR-1 workshop, Tokyo, 1999.

[3] Evans, D.A. and Lefferts, R.G., Grefenstette, G., Handerson, S.K., Hersh, W.R., and Archbold, A.A., CLARIT TREC Design, Experiments and Results, in Proceedings of the First Text REtrieval Conference(TREC-1), NIST Special Publication 500-207, Washington D.C., 1993, 494-501.

[4] Fujita, S. Notes on Phrasal Indexing—JSCB Evaluation Experiments at NTCIR AD HOC, in Proceedings of NTCIR-1 workshop, Tokyo, 1999.

[5] Fujita, S. Evaluation of Japanese Phrasal Indexing with a Large Test Collection, in RIAO2000 Conference proceedings, Paris, 2000, 1089-1098.

[6] Fujita, S. Discriminative Power and Retrieval Effectiveness of Phrasal Indexing Terms, in Proceedings of the ACL-2000 Workshop on Recent Advances in Natural Language Processing and Information Retrieval, Hong Kong, 2000, 47-55.

[7] Fujita, S. Reflections on "Aboutness"—TREC-9 Evaluation Experiments at Justsytem, TREC-9 (Notebook version), Gaithersburg, 2000.

[8] Greiff, W.R. A Theory of Term Weighting Based on Exploratory Data Analysis, SIGIR '98, Melbourne, 1998, 11-19.

[9] Robertson, S.E. and Spark Jones, K. Relevance weighting of search terms. Journal of the American Society for Information Science 1976, 27, 129-146.

[10] Robertson, S.E., Walker S., Jones S., Hancock-Beaulieu, M.M., Gatford, M. Okapi at TREC-3, in Proceedings of the Third Text REtrieval Conference(TREC-3), NIST Special Publication 500-225, Washington D.C., 1995, 109-126.

[11] Xu, J and Weischedel, R. TREC-9 Cross-lingual Retrieval at BBN, TREC-9 (Notebook version), Gaithersburg, 2000.