情 報 学 基 礎 63−12
(2001. 7. 25)

情報検索システム評価における対訳テストコレクションの意義について

藤田澄男
株式会社 ジャストシステム
Sumio_Fujita@justsystem.co.jp

Abstract
対訳テストコレクションの利用方法に注目して、NTCIR ワークショップ 2 評価実験の経験を再検討した。各種 CLIR 方式の比較実験の報告を通して、日英検索と英日検索での精度の違いに着目して分析を行った。

NTCIR-2 as a Rosetta Stone in Laboratory Experiments of IR Systems

Sumio FUJITA
JUSTSYSTEM Corporation
Sumio_Fujita@justsystem.co.jp

Abstract
Our NTCIR workshop 2 evaluation experiments are reviewed with a focus on usage of parallel test collection, which is one of the most characteristic features of NTCIR-2. Evaluation experiments of CLIR related techniques are reported and asymmetry between J-E and E-J CLIR performance is analyzed.

1. Introduction

A test collection is no more than a small fragment of vast real world just like the Rosetta stone is. What can be observed from such a small fragment of the world?

A researcher can say "method A is significantly better than B in the experiments using the test collection X and the topic set Y and relevance judgement provided by Z" from his/her direct observation. This researcher might want to generalize his/her claim to "method A is significantly better than B with collections having such features" or "method A is significantly better than B with query set having such features". What types of observation allow him/her to claim such generalization?

Despite essential simplicity of their principle, scoring processes of IR systems lack in accountability. Empirical observation is not able to directly approach mechanisms of information ranking. The experimental results are subject to be biased by features like:

-Written language: language of topic description and collection.

-Collection features: average document length of the collection, diversity of document length, target domain(newspapers, patents, academic publications, heterogeneous like web).

-Query features: length of topic description, structure of the topic description, simulated user models.

Instead of trying any combinations of such features examining through test collections around the world, we are trying to accumulate empirical evidences in carefully controlled environments to explain how information ranking mechanisms work.

Without generalization, an observation remains in discrete suggestions but they can wait for more evidences by followed researchers and this may be important significance of focused evaluation workshops annually organized.

2. Cross Language IR test collections

An interesting point of NTCIR-2 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 of CLIR effectiveness. Normally the cases where CLIR effectiveness surpasses the monolingual one are either that the monolingual run is not very

effective [10] 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. This leads to the possibility of pivot language for multi-lingual CLIR case given that certain amount of parallel or comparable collection is available.

2.2 Which language is more difficult?

Advantages of parallel test collections are not only limited to CLIR related evaluation but also better understanding about IR might be obtained by utilizing parallel collections. For example a naive question like "English IR and Japanese IR, which one is more difficult?" is never answered without examining a parallel test collection.

3. System Description

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 workshop 1 system, is utilized so that the retrieval performance is comparable with our previous system [4][7]. 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

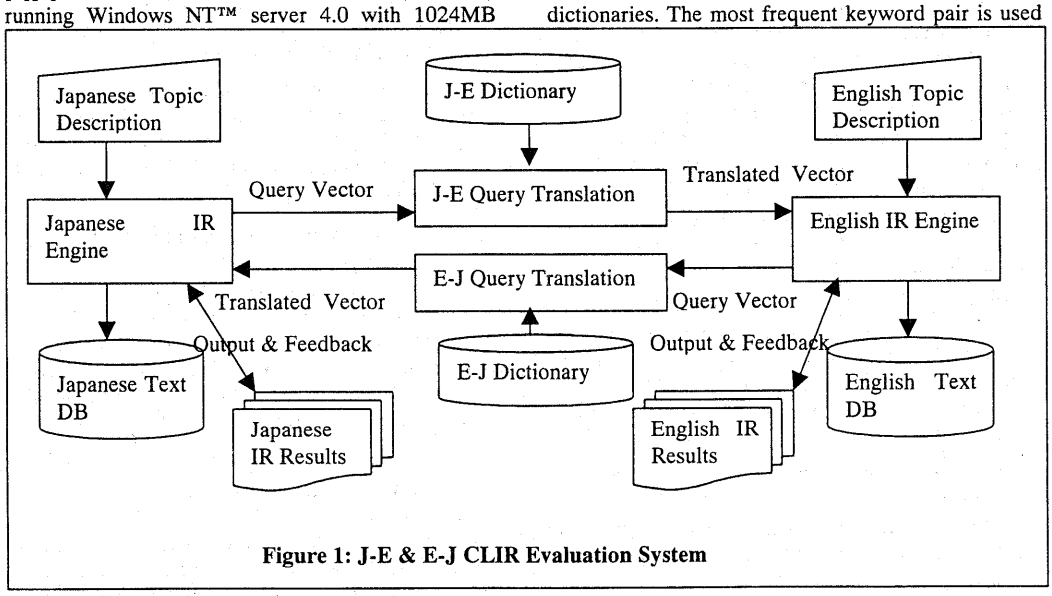

Figure 1: J-E & E-J CLIR Evaluation System

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 modifiers.

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.

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.

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 NTCIR-1 queries (topic 31-83), NTCIR-1 collection and their relevance judgement provided by NACSIS.

3.8 Parallel Collection Usage

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

Because some parts 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 NTCIR-1 test set and the mean average precision of J-E retrieval of NTCIR-1 set reaches 36.80% (long query, rel1 judgement, 39 topics of NTCIR-1 test set) without using any query translation method. This seems to be better than any NTCIR workshop 1 CLIR systems.

Among 332,918 Japanese documents in the NTCIR-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 NTCIR-2 test collections where the portion of parallel documents is much smaller (25%).

4. NTCIR workshop 2 Experiments

Our NTCIR workshop 2 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 J-E and E-J CLIR Runs

Table 1 shows our CLIR official runs and other runs.

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

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 2 compares the following different treatment of phrasal translation with monolingual baseline of target language side and with CLIR baseline runs listed in Table 1.

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

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

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.

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.

4.2 Quality of Translated Query Vectors

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. 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].

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

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

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

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 3: Correlation between measure of query vector quality and retrieval effectiveness (S&A rel1 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

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

Table 3 shows the quality of query vectors thus measured.

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 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 existence of asymmetry between J-E and E-J CLIR as observed in CLIR run experiments.

5. Reiterative Translation Experiments

The sources of difficulties of information retrieval seem to fall into either one of the following:
1)discrepancy between user aboutness of the query and author aboutness of the collection,

2)essential difficulty of the information needs that can not be adequately expressed by bag of words representation.

The second case might be out of the scope of this paper. Introducing the notion of information needs that are behind search requests, the first issue can be decomposed into two aspects:

1)Query discrepancy

Discrepancy between information needs and the query. Users can not always express the information needs in an adequate manner.

2)Collection discrepancy

Discrepancy between information needs and relevant documents. The target collection does not necessarily contain ideal relevant documents and relevance judgement may be compromise.

The first case is more common and the query should be refined through interaction.

In the second case it is the collection to be changed. Although, in the test collection case, existence of certain number of relevant documents is assured, choice of collection largely affects precision/recall rate as well as absolute relevant document numbers.

Both mean average precision of runs and query quality measures introduced in the previous chapters address the resultant of these two issues. But the fact that even the query is the same, evaluation measures are not comparable if the target collection is not identical, suggests us existence of two distinct issues.

The following experiment hopefully helps to analyze these resultant measures.

Query vectors utilized in CLIR baseline runs, JES-BASE, EJS-BASE, JEL-BASE and EJL-BASE are re-translated into another language so that for each queries both J-E translation and E-J translation are intervened. Thus constructed queries are submitted against the target language collections and these runs are named JEJS-BASE, EJES-BASE, JEJL-BASE and EJEL-BASE respectively.

These queries are re-translated again and submitted against respective target collections. These runs are named JEJES-BASE, EJEJS-BASE, JEJEL-BASE and EJEJL-BASE respectively.

Table 4 shows thus obtained results.

Comparison of measures(MAP or any query quality measures) of the run sequence JJS-BASE, JES-BASE, JEJS-BASE and JEJES-BASE with EES-BASE, EJS-BASE, EJES-BASE and EJEJS-BASE, enables us to evaluate query discrepancy in each language topic set eliminating collection discrepancy problem since each run sequence is inputting either J or E topic description and both target collections are used in each run sequence so that differences are neutralized.

Comparing measures of the run sequence JJS-BASE, EJS-BASE, JEJS-BASE and EJEJS-BASE with EES-BASE, JES-BASE, EJES-BASE and JEJES-BASE, we might evaluate collection discrepancy in each language collection eliminating query discrepancy problem.

In each run sequence, either number of runs with the same target collection or number of runs with the same topic description is controlled so that discrepancy is neutralized.

Unfortunately, this assumes that J-E translation and E-J translation are equally good in their ability of preserving information since the order and times of application of translation in both run sequences are different and uncontrollable.

Compare average of MAP of each run sequence:

AVG(JJS-BASE,JES-BASE,JEJS-BASE,JEJES-BASE)=0.2780 >

AVG(EES-BASE,EJS-BASE,EJES-BASE,EJEJS-BASE)=0.2462

AVG(JJL-BASE,JEL-BASE,JEJL-BASE,JEJEL-BASE)=0.3308 >

Run tag (REL1)	JJS-BASE	JES-BASE	JEJS-BASE	JEJES-BASE
AVG(SUM(PW))	33.84	26.85	24.47	20.7
Mean Avg. Prec.	0.3611	0.2634	0.2556	0.2319
Run tag (REL1)	EESBASE	EJS-BASE	EJES-BASE	EJEJS-BASE
AVG(SUM(PW))	28.02	28.52	17.8	18.73
Mean Avg. Prec.	0.3637	0.2151	0.2208	0.1853
Run tag (REL1)	JJL-BASE	JEL-BASE	JEJL-BASE	JEJEL-BASE
AVG(SUM(PW))	137.19	90.97	76.50	60.79
Mean Avg. Prec.	0.402	0.3519	0.3034	0.2657
Run tag (REL1)	EEL-BASE	EJL-BASE	EJEL-BASE	EJEJL-BASE
AVG(SUM(PW))	111.93	94.29	53.81	60.42
Mean Avg. Prec.	0.4137	0.2805	0.2858	0.2438

Table 4: measure of query vector quality and retrieval effectiveness in reiterative translation experiment(S&A rel1 relevance judgement)

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

SUM: sum of each term measure in a query vector

AVG: average of topic 101 to 149

Mean Avg. Prec.: Mean average precision of runs

AVG(EEL-BASE,EJL-BASE,EJEL-BASE,EJEJL-BASE)=0.3060

These results suggest that Japanese topic description is better representative of the information needs than English one irrespective of target collections.

AVG(JJS-BASE,EJS-BASE,JEJS-BASE,EJEJS-BASE)=0.2543 <

AVG(EES-BASE,JES-BASE,EJES-BASE,JEJES-BASE)=0.2700

AVG(JJL-BASE,EJL-BASE,JEJL-BASE,EJEJL-BASE)=0.3074 <

AVG(EEL-BASE,JEL-BASE,EJEL-BASE,JEJEL-BASE)=0.3293

On the other hands, these results suggest that the English collection is better information repository against the information needs irrespective of the quality of topic description.

6. Conclusions

NTCIR-2 is a Rosetta stone of which hieroglyphics are not yet deciphered.

Through CLIR evaluation experiments, asymmetry between J-E and E-J CLIR performance is observed. In order to explain observed asymmetry of effectiveness, two types of discrepancies are assumed and existence of such discrepancies is suggested by reiterative translation experiments.

Such experimental works hopefully bring us better understanding of the mechanisms of information ranking, which is difficult to directly observe. To do so, our experimental research continues to be back to the Cranfield style approaches carefully preparing controlled environments.

Acknowledgements

The author thanks NII for providing us of NTCIR-1 and NTCIR-2. We participated in the NTCIR workshops utilizing NTCIR-1 and NTCIR-2, that are developed by NII thanks to understanding of academic societies (http://research.nii.ac.jp/~ntcadm/acknowledge/thanks1-en.html) who provided the data.

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