Toshiba BRIDJE at NTCIR-5 CLIR: Evaluation using Geometric Means

Tetsuya Sakai Toshihiko Manabe Akira Kumano Makoto Koyama
Tomoharu Kokubu
Knowledge Media Laboratory, Toshiba Corporate R&D Center
tetsuya.sakai@toshiba.co.jp

Abstract

Toshiba participated in the Monolingual and Bilingual IR tasks at NTCIR-5 CLIR using the BRIDJE system. We submitted 24 runs covering three topic languages (Japanese, English and Chinese) and two document languages (Japanese and English), and achieved the highest performances in the E-J-T, E-J-D, C-J-T, C-J-D, J-E-T *and* J-E-D *subtasks. This paper (re-)examines Partial Disambiguation and the Pivot Language approach for Bilingual IR, Selective Sampling with memory Resetting for Mono/Bilingual IR and a new Monolingual IR strategy called Bounce-and-Throw, using the Geometric Mean versions of Average Precision and Q-measure in addition to the standard Arithmetic Mean ones. Our analyses show that the Geometric Mean, which focusses on the "harder" topics, provides new insight into retrieval effectiveness evaluation.*

Keywords: *Partial Disambiguation, Selective Sampling, Bounce-and-Throw, Q-measure, Geometric Mean.*

1 Introduction

Toshiba participated in the Monolingual and Bilingual IR tasks at NTCIR-5 CLIR [3] using the BRIDJE system [9, 10, 13, 14]. Through our participation, we (re-)examined the following questions:

1. For Bilingual IR based on Machine Translation (MT), can an enhanced version of *Partial Disambiguation* (PD) [8, 9, 13] outperform *Full Disambiguation* (FD)? (See Section 2.2.)

2. For Bilingual IR based on MT, is the *Pivot Language* approach [3] practically feasible? (See Section 2.3.)

3. For Mono/Bilingual IR, how does Selective Sampling with memory Resetting (SSR) [14] compare to standard Pseudo-Relevance Feedback (PRF)? (See Section 2.4.)

4. For Monolingual IR, can the *Bounce-and-Throw* (BaT) method, which uses an external foreign-language corpus, outperform a standard monolingual run? (See Section 2.5.)

Tables 1 and 2 provide the summary of our NTCIR-5 CLIR results. We officially achieved the highest performances in the E-J-T, E-J-D, C-J-T, C-J-D, J-E-T and J-E-D subtasks. The first column shows our own run labels which reflect the search strategies used, while the second column shows the official names at NTCIR-5. The bottom of each table contains a brief description of the labels, and the search strategies will be explained in Section 2. The third and the fourth columns show the Mean *Relaxed* and *Rigid* Average Precision (AveP) values [3], and the fifth columns shows the Mean *Q-measure* values [10, 11, 12, 15], computed based on graded relevance[1]. In addition, the tables provide the *Geometric* Means of Relaxed AveP and Q-measure, which will be discussed in Section 3.2.

Section 2 describes our search strategies. Section 3 summarises the relationship between AveP and Q-measure, and discusses the advantage of taking the Geometric Mean of IR metrics across topics [17]. Section 4 examines our search strategies using Arithmetic and Geometric Mean versions of AveP and Q-measure. Section 4 concludes this paper.

2 Search Strategies for NTCIR-5

2.1 Baseline Runs

All of our runs used the BRIDJE system [9, 10, 13] for document retrieval. BRIDJE uses the Okapi/BM25 term weighting [16], and the Okapi parameters were set to $k_1 = 1$ and $b = 0.6$ for NTCIR-5. By default, the term selection criterion used for Pseudo-Relevance Feedback (PRF) is the *offer weight* (ow) [8]. Top $P =$

[1] We strongly recommend that the Organisers start using graded-relevance IR metrics officially (perhaps along with Relaxed AveP), since some of them (e.g. Q-measure) have been shown to be at least as reliable as binary-relevance ones [15].

20 documents in the initial ranked output were used to extract $T = 30$ additional terms for NTCIR-5.

Our Bilingual IR runs used Toshiba Machine Translation (MT) systems for search request translation. BRIDJE uses English-Japanese (EJ) and Japanese-English (JE) MT as its components for Bilingual IR, and can perform *Partial Disambiguation* (PD) as well as *Full Disambiguation* (FD), i.e. the use of MT as a black box for search request translation. We will describe PD in Section 2.2.

We used two additional MT systems for search request translation: The Chinese-Japanese (CJ) system developed at Toshiba Knowledge Media Laboratory and the English-Chinese (EC) system developed at Toshiba China. These systems support FD only.

2.2 Partial Disambiguation for Bilingual IR

Partial Disambiguation (PD) was introduced in the post-submission experiments at NTCIR-3 CLIR [8]. While Full Disambiguation (FD) can only obtain one target-language term from each source-language term, PD preserves alternative translations that remain after the *semantic analysis* disambiguation stage in MT. At the retrieval stage, the target-language terms for each source-language term are treated as a set of synonyms, as in *Pirkola's method* [6].

The average superiority of PD over FD was consistent across different test collections [9, 13]. However, the differences were generally not statistically significant due to the noise introduced by the alternative translations. For NTCIR-5, we therefore used only up to two translation candidates for each source-language term. Moreover, because we currently have no way of assigning priorities within each synonym set, each PD run was actually generated by merging two component runs, one using the FD query and the other using the PD query, by simply taking the average of document scores. This in effect emphasises the first translation candidates in comparison to the second ones.

2.3 Using a Pivot for Bilingual IR

The pivot language approach for MT-based Bilingual IR is attractive because, if it works, we will not have to build an MT system for every language pair for the purpose of Bilingual IR. As mentioned earlier, we have EJ and JE MT which support both PD and FD, plus CJ and EC MT which support FD only. Given this constraint, we generated two sets of pivot language runs: CJE (i.e. using Japanese as a pivot language for Chinese-English IR) and ECJ (i.e. using Chinese as a pivot language for English-Japanese IR).

The CJE approach was used officially as we did not have a direct Chinese-English MT system. For the second-stage MT (i.e. JE MT), we tried both PD and FD. Thus, note that CJE-T-PD-PRF and CJE-D-PD-PRF in Table 2 used PD at the JE MT stage but not at the CJ MT stage.

The ECJ approach was not used officially because EJ runs clearly outperformed ECJ runs in our preliminary experiments with the NTCIR-4 data.

2.4 Selective Sampling with Memory Resetting

The original Selective Sampling (SS) method was proposed at NTCIR-4 [10] [2]. The idea of SS is to select a variety of document samples from the initial ranked output instead of just scooping the top P documents. A set of consecutive documents in the initial list is regarded as a "cluster" if the documents contain the same set of query terms. SS tries to collect documents from different "clusters", thereby skipping some documents at the top of the initial list. (SS does not actually perform document clustering and is therefore computationally inexpensive.)

In [14], we observed that SS may skip too many documents if there is a very large cluster at the top of the initial list, and proposed a new method called Selective Sampling with memory Resetting (SSR). When SSR encounters a very large cluster, it takes some documents, skips some documents, and then starts taking some documents again. We showed that SSR outperforms standard PRF at least as often as PRF outperforms SSR, and therefore that "taking the top P" is not always the best strategy, although the two methods are comparable in terms of average performance.

We repeated the SSR experiments at NTCIR-5. The SSR parameters, namely, the minimum/maximum number of pseudo-relevant documents required (P_{min}/P_{max}) and the maximum number of documents examined (P_{scope}) were set to $P_{min} = 10$, $P_{max} = 20$ and $P_{scope} = 50$ [14]. As with PRF, we used $T = 30$ expansion terms.

2.5 Bounce-and-Throw for Monolingual IR

Bounce-and-Throw (BaT) is a new method we tried for enhancing monolingual IR effectiveness.

Previous work (e.g. [1, 5, 7]) suggests that *data fusion* (i.e. merging several ranked document lists) is worthwhile, provided that the component runs are substantially different from each other. For example, Kwok *et al.* [5] recently proposed a successful data fusion approach that probes the Web for obtaining alternative queries.

We also decided to explore the use of an alternative query for data fusion, but did not want to rely on an external search engine. Moreover, since we *knew* that

[2] Unfortunately, the effect of SS was overestimated in [10], as a result of underestimating a standard PRF run. The correct results with the NTCIR-4 data can be found in [14].

we were searching a collection of newspaper articles from 2000-2001, we thought that an external collection of newspaper articles from the same period might be more useful than the whole Web. More specifically, we addressed the following question: *When performing monolingual IR with newspaper articles from a particular period, could a foreign-language collection of newspaper articles from the same period be of any use for obtaining an effective alternative query?*

BaT for Japanese IR simply works as follows:

1. Using an external *English* corpus as the temporary target, perform JE IR (with PRF).
2. Translate the top 10 retrieved documents by EJ MT, and extract 30 Japanese terms from them by using $tf * idf$ for term selection. tf is the total number of occurrences in the top 10 traslated documents and idf is the inverse document frequency based on the final target collection.
3. Use the 30 terms as an "alternative query" and search the target Japanese collection (with PRF).
4. Merge the ranked list (the "BaT" component) with a standard Japanese monolingual PRF run (a "direct throw").

BaT for English IR works similarly using an external Japanese corpus.

We use the following standard method for obtaining the combined document score:

$$SCORE_{BaT} = \alpha * nscore_{direct} + (1-\alpha) * nscore_{BaT} \tag{1}$$

where $\alpha (\leq 1)$ is a parameter (set to 0.8 for NTCIR-5) and $nscore$ represents a normalised component document score. The normalisation is achieved as follows:

$$nscore = (score - score_{min})/(score_{max} - score_{min}) \tag{2}$$

where $score$, $score_{min}$ and $score_{max}$ are the original/minimum/maximum document scores in the component run, respectively.

For Japanese IR, we used the NTCIR-5 Korea Times, Yomiuri and Mainichi English data as the external collection. For English IR, we used the NTCIR-5 Yomiuri and Mainichi Japanese data.

BaT differs from *Collection Enrichment* [4] and *parallel pseudo-relevance feedback* [8] in that it uses an external *foreign-language* collection. Thus, in addition to the *enrichment effect*, we hoped that the "bounce" (i.e. two-way MT described in Steps 1 and 2 above) may introduce a *query rephrasing effect*.

3 Metrics

3.1 AveP and Q-measure

Let R denote the number of relevant documents for a topic, and let $count(r)$ denote the number of relevant documents within top r of the ranked output. *Precision at Rank* r is given by $P(r) = count(r)/r$. Let $isrel(r)$ denote a flag, such that $isrel(r) = 1$ if the document at Rank r is relevant and $isrel(r) = 0$ otherwise. AveP can be expressed as:

$$AveP = \frac{1}{R} \sum_{1 \leq r \leq L} isrel(r)P(r) \tag{3}$$

where L is the ranked output size.

Let X denote a relevance level, and let $gain(X)$ denote the *gain value* for successfully retrieving an X-relevant document. As $X \in \{S, A, B\}$ for NTCIR CLIR, we use $gain(S) = 3$, $gain(A) = 2$ and $gain(B) = 1$ by default. The *gain at Rank* r is defined as $g(r) = gain(X)$ if the document at Rank r is X-relevant and $g(r) = 0$ if it is nonrelevant. The *cumulative gain* [2] at Rank r is defined as $cg(r) = \sum_{1 \leq i \leq r} g(i)$. In particular, let $g_I(r)$ and $cg_I(r)$ represent the (cumulative) gain at Rank r for an *ideal* ranked output, which exhaustively lists up all S,A,B-relevant documents in this order (in the case of NTCIR CLIR). Then, Q-measure [11, 12, 15] can be expressed as:

$$Q\text{-}measure = \frac{1}{R} \sum_{1 \leq r \leq L} isrel(r)BR(r) \tag{4}$$

where

$$BR(r) = \frac{cg(r) + count(r)}{cg_I(r) + r} . \tag{5}$$

Q-measure has the following properties:

- $Q\text{-}measure = 1$ if and only if the system output is an ideal one.
- In a binary relevance environment, $Q\text{-}measure = AveP$ holds if and only if there is no relevant document below Rank R, while $Q\text{-}measure > AveP$ holds if and only if there is at least one relevant document below Rank R.

Moreover, Q-measure is at least as stable and sensitive as other graded relevance metrics such as *normalised discounted cumulative gain* [2], and is more highly correlated with AveP than these alternatives [15].

3.2 Geometric Mean

At the TREC 2004 Robust Retrieval track [17], the *Geometric* Mean of AveP was used to focus on the "hard" topics, i.e. those with very low effectiveness. Geometric Mean is suitable for system optimisation for guaranteeing a minimal level of retrieval quality for any query input. (For example, the Arithmetic Mean of 2 and 50 is 26, while the Geometric Mean is 10.) This is very important for a practical IR system, so we apply Geometric Mean to Q-measure as well.

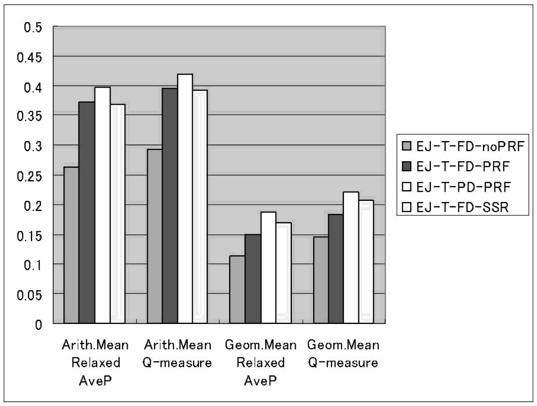

Figure 1. Arithmetic vs Geometric Means (EJ-T).

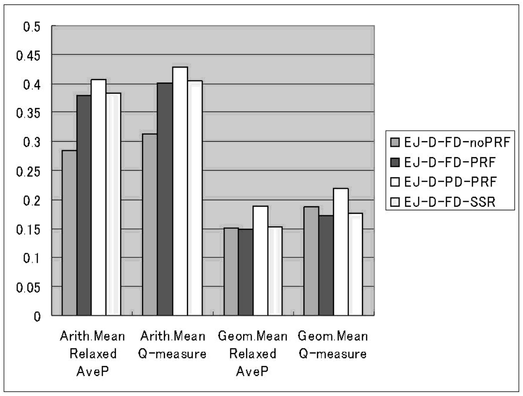

Figure 2. Arithmetic vs Geometric Means (EJ-D).

Let N be the total number of topics, and let m_i denote the value of a metric for the i-th topic (with four significant figures). Then the actual algorithm for obtaining the Geometric Mean (GM) is as follows [17]:

$$GM = \exp(\frac{\sum_{1 \leq i \leq N} \log(m_i + 0.00001)}{N}) - 0.00001\,. \tag{6}$$

The next section shows that Geometric Means provides new insight into IR evaluation.

4 Discussions

4.1 Partial Disambiguation

Table 1(c) and (d), partially visualised in Figures 1 and 2, show that PD was quite successful for the EJ subtasks: it achieved the highest mean performance in terms of *all* metrics. Although the differences between PD and FD are not statistically significant, the graphs indicate that the performance differences are greater in terms of Geometric Mean than in terms of Arithmetic Mean, suggesting that PD reduces the number of poor performers. In particular, Table 1(d) and Figure 2 are interesting: even though both EJ-D-FD-PRF and EJ-D-PD-PRF are both significantly better than EJ-D-FD-noPRF in terms of Rigid/Relaxed AveP and Q-measure, EJ-D-FD-PRF does *not* outperform EJ-D-FD-noPRF in terms of Geometric Means. That is, according to Geometric Means, EJ-D-PD-PRF is the only successful EJ-D run with query expansion. (This is not a contradiction: the sign test examines whether the medians of two unknown distributions representing the performances of two runs are different. Even if the median of system x is smaller than that of system y, x may outperform y in terms of geometric means. Similarly, Geometric Means may disagree with *parametric* significance tests as well.) Thus, the Geometric Means uncover the well-known disadvantage of standard PRF: it hurts around one-third of the topics to improve the average performance [14]. Furthermore, it can be observed that the negative effect of EJ-D-FD-PRF (compared to EJ-D-FD-noPRF) is clearer in terms of Geometric Mean *Q-measure* than in terms of Geometric Mean *Relaxed AveP*. This suggests that PRF often lowers the ranks of highly relevant documents, while perhaps raising the ranks of partially relevant ones.

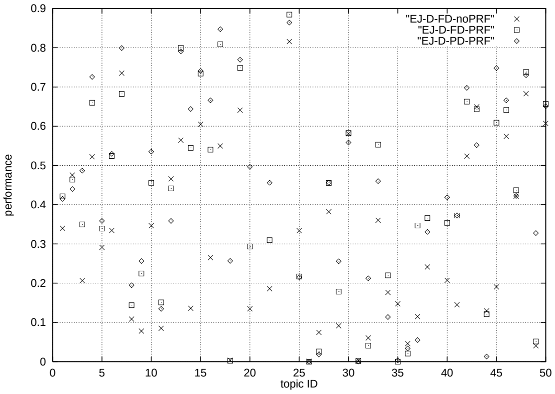

Figure 3. Per-topic analysis: EJ-D-FD-PRF **vs** EJ-D-PD-PRF **(Q-measure).**

Per-topic analysis reveals that the four worst Q-measure values for EJ-D-FD-PRF are 0.0000, 0.0000, 0.0015 and 0.0022, while those for EJ-D-PD-PRF are 0.0000, 0.0007, 0.0035 and 0.0131, which are slightly better. This explains why the Geometric Mean Q-measure of EJ-D-FD-PRF is low. Compared to EJ-D-FD-noPRF, both EJ-D-FD-PRF and EJ-D-PD-PRF improve 34 topics but hurt 13 topics. EJ-D-PD-PRF outperforms EJ-D-FD-PRF for 27 topics, while the opposite is true for 19 topics. Figure 3 shows the Q-measure values of EJ-D-FD-noPRF, EJ-D-FD-PRF and EJ-D-PD-PRF per topic.

Table 2(c) and (d) show that PD was generally successful for the JE subtasks as well, except that the

Geometric Means of JE-D-PD-PRF are not as high as those of JE-D-FD-PRF. Again, this is because Geometric Means highlight the very worst topics. Note also that our JE-T runs are almost as effective as our EE-T runs. The main reason is that JE MT was extremely successful (and more successful than EJ MT) for the NTCIR-5 CLIR test collection. A similar trend was observed at NTCIR-4 CLIR as well [10].

Table 2(e) and (f) show that PD (used in the JE MT step only) was also successful for the CJE subtasks. In particular, CJE-D-PD-PRF outperforms CJE-D-FD-PRF substantially: the differences are statistically significant in terms of Rigid/Relaxed AveP and Q-measure. In terms of Q-measure, CJE-D-PD-PRF outperforms CJE-D-FD-PRF for 32 topics, while the opposite is true for only 17 topics. Thus, PD seems to work well with the pivot language approach.

In summary, PD is more robust than FD.

4.2 Pivot Language Runs

A quick comparison of the Arithmetic Mean values of the ECJ runs with those of direct EJ runs in Table 1(c) and (d) seems to suggest that the simple approach of using two MT systems with Chinese as a pivot language is feasible: For example, in terms of Q-measure, ECJ-D-FD-PRF is $0.3301/0.4004 = 82\%$ of EJ-D-FD-PRF and $0.3301/0.4771 = 69\%$ of JJ-D-PRF, which is not discouraging. However, the Geometric Means provide a more pessimistic view: all the ECJ runs are below 0.1, revealing that the simple tandem of two MT systems fails completely for many topics. Per-topic analysis shows that many important named entities in the original English topics were lost during the EC MT phase.

The CJE runs shown in Table 2(e) and (f) look somewhat better. For example, the Geometric Mean Q-measure of CJE-D-PD-PRF is 0.2158. Moreover, it can be observed that, while the query expansion runs in Table 2(e) look comparable to those in (f) in terms of Arithmetic Means, they are much less effective in terms of Geometric Means. Thus, query expansion was not so successful with the CJE TITLE queries. One cause of this can be found in the performance values of the *initial* runs in Table 2(e) and (f): CJE-T-FD-noPRF outperforms CJE-D-noPRF in terms of Arithmetic Means, but the opposite is true in terms of Geometric Means. This means that there are extremely bad initial CJE TITLE queries, which PRF probably failed to imrove. This example also demonstrates that Geometric Means is useful for looking at results from a different angle.

4.3 Selective Sampling

Our SSR results are generally consistent with our previous results reported in [14]: SSR and PRF are comparable on average, but SSR outperforms PRF for about one-half of the topics. Thus, using the top P documents for query expansion is not always the best strategy. More specifically, EJ-T-FD-SSR outperforms EJ-T-FD-PRF for 19 topics in terms of Q-measure, while the opposite is true for 12 topics. Compared to EJ-T-FD-noPRF, EJ-T-FD-PRF improves 37 topics but hurts 10 topics; EJ-T-FD-SSR improves 39 topics but hurts 8 topics.

We now re-examine Figure 1. Interestingly, while EJ-T-FD-PRF slightly outperforms EJ-T-FD-SSR in terms of Arithmetic Means, the opposite is true in terms of Geometric Means. This is because SSR does slightly better with the hardest topics than PRF does. For example, the three worst per-topic Q-measure values for EJ-T-FD-PRF are 0.0000, 0.0003 and 0.0008, while those for EJ-T-FD-SSR are 0.0009, 0.0011 and 0.0015.

4.4 Bounce-and-Throw

Table 1(a) and (b) show that our Japanese monolingual BaT runs are only comparable to standard PRF. On the other hand, Table 2(a) and (b) show somewhat better results for the English case: The BaT runs are the best runs on average among the English monolingual runs in terms of all metrics, and EE-T-BaT significantly outperforms EE-T-PRF (and EE-T-SSR, though not explicitly indicated in the table) in terms of Rigid AveP. Even in terms of Q-measure, EE-T-BaT outperforms EE-T-PRF for 30 topics while the opposite is true for only 19 topics. In short, English BaT worked but Japanese BaT did not.

The above difference probably arose from the differences in the quantity and quality of external data used for BaT. That is, the Japanese external document set used for our English BaT run (Yomiuri and Mainichi: 858,400 documents) was probably more reliable and useful than the English external document set used for our Japanese BaT run (Korea Times, Yomiuri and Mainichi: only 60,426 documents). Another possible factor is the difference in the quality of MT: as was mentioned in Section 4.1, our JE MT is more accurate than our EJ MT at least for search request translation. Assuming that this holds for document translation as well, it is possible that our Japanese BaT suffered from noise introduced by EJ MT during the translation of the retrieved external documents.

Our current BaT method is simplistic and there is room for improvement. Nevertheless, it is good to know that an external foreign-language corpus can help improve the performance of monolingual IR.

5 Conclusions

Through our participation at NTCIR-5 CLIR, we re-examined Partial Disambiguation for Bilingual IR, the Pivot Language approach to Bilingual IR, Selective Sampling (with memory Resetting) for general IR, and the Bounce-and-Throw method for monolingual IR. We used Geometric Mean AveP and Q-measure along with the standard Arithmetic Means for evaluation, and demonstrated that Geometric Means provide new insight into IR evaluation by focussing on the "harder" topics. For example, we showed that PRF may hurt Geometric Mean values even if its effect is significantly positive in terms of a statistical test. We believe that Geometric Means are useful for building a practical IR system that never presents a disastrous output to the user. Our main findings are:

- Partial Disambiguation is generally more effective than Full Disambiguation (i.e. Black-Box MT). It seems to work well with the pivot language approach (CJE-D-PD-PRF significantly outperformed CJE-D-FD-PRF).
- A simple tandem of two MT systems for the pivot language approach is not good enough. The geometric means highlight this fact.
- Selective Sampling outperforms standard PRF as often as PRF outperforms Selective Sampling, and the two methods are comparable in terms of average performance. This is in agreement with the results reported in [14].
- Bounce-and-Throw for English monolingual IR, which used a Japanese external collection, was effective (EE-T-BaT significantly outperformed EE-T-PRF). Thus the use of an external foreign-language corpus for monolingual IR deserves further studies.

Acknowledgment

We thank Tatsuya Izuha for his help with Chinese-Japanese MT, and Zhu Jiang, Li Guo Ha and Wan Haifeng of Toshiba China for providing the English-Chinese MT output for the topics.

References

[1] Croft, W. B.: Combining Approaches to Information Retrieval, *Advances in Information Retrieval - Recent Research from the Center for Intelligent Information Retrieval* (Croft, W. B. (ed.)), Kluwer Academic Publishers, pp.1-36, 2000.

[2] Kekäläinen, J.: Binary and Graded Relevance in IR evaluations - Comparison of the Effects on Ranking of IR Systems, *Information Processing and Management*, Vol. 41, pp.1019-1033, 2005.

[3] Kishida, K. *et al.*: Overview of CLIR Task at the Fifth NTCIR Workshop, *NTCIR-5 Proceedings*, to appear, 2005.

[4] Kwok, K. L., Grunfeld, L., Dinstl, N. and Chan, M.: TREC-9 Cross Language, Web and Question-Answering Track Experiments using PIRCS, *TREC-9 Proceedings*, 2001.

[5] Kwok, K. L., Grunfeld, L., Sun, H. L. and Deng. P.: TREC2004 Robust Track Experiments using PIRCS, *TREC 2004 Proceedings*, 2004.

[6] Pirkola, A.: The Effects of Query Structure and Dictionary Setups in Dictionary-Based Cross-Language Information Retrieval, *ACM SIGIR '98 Proceedings*, pp.55-63 (1998).

[7] Sakai, T.: Combining the Ranked Output from Fulltext and Summary Indexes, ACM SIGIR 2001 Workshop on Text Summarization, pp.27-34, 2001.

[8] Sakai, T., Koyama, M., Suzuki, M. and Manabe, T.: Toshiba KIDS at NTCIR-3: Japanese and English-Japanese IR, NTCIR-3 Proceedings, 2003.

[9] Sakai, T. *et al.*: BRIDJE over a Language Barrier: Cross-Language Information Access by Integrating Translation and Retrieval, *IRAL 2003 Proceedings*, pp. 65-76, 2003.

[10] Sakai, T., Koyama, M., Kumano, A. and Manabe, T.: Toshiba BRIDJE at NTCIR-4 CLIR: Monolingual/Bilingual IR and Flexible Feedback, NTCIR-4 Proceedings, 2004.

[11] Sakai, T: New Performance Metrics based on Multigrade Relevance: Their Application to Question Answering, *NTCIR-4 Proceedings*, 2004.

[12] Sakai, T.: Ranking the NTCIR Systems based on Multigrade Relevance, *AIRS 2004 Proceedings*, pp.170-177, 2004. Also available in Myaeng, S. H. et al. (Eds.): *AIRS 2004*, Lecture Notes in Computer Science 3411, pp.251–262, Springer-Verlag, 2005.

[13] Sakai, T.: Advanced Technologies for Information Access, *International Journal of Computer Processing of Oriental Languages*, Vol. 18, No. 2, pp.95–113, 2005.

[14] Sakai, T., Manabe, T. and Koyama, M: Flexible Pseudo-Relevance Feedback via Selective Sampling, *ACM Transactions on Asian Language Information Processing*, to appear, 2005.

[15] Sakai, T.: The Reliability of Metrics based on Graded Relevance, *AIRS 2005*, Lecture Notes in Computer Science 3689, pp.1–16, Springer-Verlag, 2005.

[16] Sparck Jones, K., Walker, S. and Robertson, S. E.: A Probabilistic Model of Information Retrieval: Development and Comparative Experiments, *Information Processing and Management* 36, Part I (pp.779-808) and Part II (pp.809-840), 2000.

[17] Voorhees, E. M.: Overview of the TREC 2004 Robust Retrieval Track, *TREC 2004 Proceedings*, 2005.

Table 1. Official and unofficial NTCIR-5 Japanese document runs (47 topics).

Name used in this paper	Official Name at NTCIR-5	Arithmetic Mean			Geometric Mean	
		Rigid AveP	Relaxed AveP	Q-measure	Relaxed AveP	Q-measure
(a) Monolingual Japanese TITLE runs						
Official Top Performer		0.4193	0.5028	-	-	-
JJ-T-noPRF	-	0.2524	0.3375	0.3703	0.2465	0.2869
JJ-T-PRF	TSB-J-J-T-01	0.3569↑↑	**0.4565**↑↑	0.4806↑↑	0.3485	0.3917
JJ-T-SSR	-	0.3582↑↑	0.4498↑↑	0.4753↑↑	0.3444	0.3884
JJ-T-BaT	TSB-J-J-T-02	**0.3609**↑↑	0.4558↑↑	**0.4809**↑↑	**0.3501**	**0.3938**
(b) Monolingual Japanese DESCRIPTION runs						
Official Top Performer		0.3823	0.4707	-	-	-
JJ-D-noPRF	-	0.2742	0.3542	0.3861	0.2733	0.3204
JJ-D-PRF	TSB-J-J-D-03	0.3481↑↑	0.4560↑↑	0.4771↑↑	**0.3545**	0.3985
JJ-D-SSR	-	0.3476↑↑	0.4546↑↑	0.4764↑↑	0.3533	0.3977
JJ-D-BaT	TSB-J-J-D-04	**0.3526**↑↑	**0.4598**↑↑	**0.4821**↑↑	0.3536	**0.3997**
(c) English-Japanese TITLE runs (including pivot runs)						
Official Top Performer		0.2916 (TSB)	0.3967 (TSB)	-	-	-
EJ-T-FD-noPRF	-	0.1917	0.2630	0.2924	0.1128	0.1456
EJ-T-FD-PRF	TSB-E-J-T-01	0.2714↑↑	0.3720↑↑	0.3956↑↑	0.1499	0.1829
EJ-T-PD-PRF	TSB-E-J-T-02	**0.2916**↑↑	**0.3967**↑↑	**0.4193**↑↑	**0.1863**	**0.2208**
EJ-T-FD-SSR	-	0.2732↑↑	0.3679↑↑	0.3915↑↑	0.1684	0.2065
ECJ-T-FD-noPRF	-	0.1382	0.1922	0.2192	0.0363	0.0474
ECJ-T-FD-PRF	-	0.2275↑↑	0.3028↑↑	0.3199↑↑	0.0534	0.0677
(d) English-Japanese DESCRIPTION runs (including pivot runs)						
Official Top Performer		0.2981 (TSB)	0.4070 (TSB)	-	-	-
EJ-D-FD-noPRF	-	0.2116	0.2834	0.3122	0.1514	0.1867
EJ-D-FD-PRF	TSB-E-J-D-03	0.2752↑↑	0.3792↑↑	0.4004↑↑	0.1501	0.1729
EJ-D-PD-PRF	TSB-E-J-D-04	**0.2981**↑↑	**0.4070**↑↑	**0.4272**↑↑	**0.1887**	**0.2197**
EJ-D-FD-SSR	-	0.2855↑↑	0.3829↑↑	0.4052↑↑	0.1520	0.1757
ECJ-D-FD-noPRF	-	0.1490	0.1886	0.2195	0.0519	0.0674
ECJ-D-FD-PRF	-	0.2318↑↑	0.3085↑↑	0.3301↑↑	0.0623	0.0781
(e) Chinese-Japanese TITLE runs						
Official Top Performer		0.2684 (TSB)	0.3466 (TSB)	-	-	-
CJ-T-FD-noPRF	-	0.1540	0.2166	0.2484	0.0782	0.0977
CJ-T-FD-PRF	TSB-C-J-T-01	0.2662↑↑	0.3459↑↑	0.3694↑↑	0.1281	0.1587
CJ-T-FD-SSR	TSB-C-J-T-02	**0.2684**↑↑	**0.3466**↑↑	**0.3709**↑↑	**0.1363**	**0.1642**
(f) Chinese-Japanese DESCRIPTION runs						
Official Top Performer		0.2471 (TSB)	0.3406 (TSB)	-	-	-
CJ-D-FD-noPRF	-	0.1636	0.2175	0.2436	0.0771	0.0997
CJ-D-FD-PRF	TSB-C-J-D-03	**0.2471**↑↑	0.3405↑↑	0.3604↑↑	0.1049	0.1308
CJ-D-FD-SSR	TSB-C-J-D-04	0.2470↑↑	**0.3406**↑↑	**0.3608**↑↑	**0.1117**	**0.1407**

noPRF: no Pseudo-Relevance Feedback (i.e. initial search)
PRF: Pseudo-Relevance Feedback using the offer weight;
SSR: Selective Sampling with Memory Resetting;
BaT: Bounce-and-Throw using English corpora from 2000-2001 (Korea Times, Yomiuri and Mainichi);
FD: Full Disambiguation in search request translation;
PD: Partial Disambiguation in search request translation (for EJ-MT only).
Highest values within each subtask are indicated in bold. Wherever any of our official runs achieved the highest performance among all participants, this is indicated by "(TSB)" in the "Top Performer" row.
Using the two-sided sign test, runs that are significantly better than the corresponding noPRF run are indicated by ↑↑ ($\alpha = 0.01$) in the Arithmetic Mean columns.
None of our runs is significantly better than the corresponding standard PRF run.

Table 2. Official and unofficial NTCIR-5 English document runs (49 topics).

Name used	Official Name	Arithmetic Mean			Geometric Mean	
in this paper	at NTCIR-5	Rigid	Relaxed		Relaxed	
		AveP	AveP	Q-measure	AveP	Q-measure
(a) Monolingual English TITLE runs						
Official Top Performer		0.4539	0.5046	-	-	-
EE-T-noPRF	-	0.3513	0.3920	0.4185	0.3197	0.3561
EE-T-PRF	TSB-E-E-T-01	0.4330↑↑	0.4843↑↑	0.5072↑↑	0.4067	0.4493
EE-T-SSR	-	0.4278↑↑	0.4800↑↑	0.5038↑↑	0.4047	0.4470
EE-T-BaT	TSB-E-E-T-02	**0.4432**↑↑⇑	**0.4901**↑↑	**0.5155**↑↑	**0.4165**	**0.4573**
(b) Monolingual English DESCRIPTION runs						
Official Top Performer		0.4581	0.4981	-	-	-
EE-D-noPRF	-	0.3505	0.3764	0.4013	0.2849	0.3255
EE-D-PRF	TSB-E-E-D-03	0.4088↑↑	0.4446↑↑	0.4645↑↑	0.2853	0.3413
EE-D-SSR	-	0.4069↑↑	0.4422↑↑	0.4626↑↑	0.2862	0.3421
EE-D-BaT	TSB-E-E-D-04	**0.4198**↑↑	**0.4559**↑↑	**0.4776**↑↑	**0.2937**	**0.3507**
(c) Japanese-English TITLE runs						
Official Top Performer		0.4389 (TSB)	0.4919 (TSB)	-	-	-
JE-T-FD-noPRF	-	0.3491	0.3931	0.4160	0.2736	0.3134
JE-T-FD-PRF	TSB-J-E-T-01	0.4351↑↑	0.4846↑↑	0.5031↑↑	0.3738	0.4154
JE-T-PD-PRF	TSB-J-E-T-02	**0.4389**↑↑	**0.4919**↑↑	**0.5087**↑↑	**0.4046**	**0.4393**
JE-T-FD-SSR	-	0.4323↑↑	0.4821↑↑	0.5025↑↑	0.3733	0.4157
(d) Japanese-English DESCRIPTION runs						
Official Top Performer		0.4135 (TSB)	0.4642 (TSB)	-	-	-
JE-D-FD-noPRF	-	0.3267	0.3548	0.3811	0.2338	0.2803
JE-D-FD-PRF	TSB-J-E-D-03	0.4113↑↑	0.4518↑↑	0.4737↑↑	0.3011	0.3535
JE-D-PD-PRF	TSB-J-E-D-04	**0.4135**↑↑	**0.4642**↑↑	**0.4780**↑↑	0.2774	0.3301
JE-D-FD-SSR	-	0.4103↑↑	0.4499↑↑	0.4725↑↑	**0.3020**	**0.3537**
(e) Chinese-English TITLE runs (pivot runs only)						
Official Top Performer		0.3702	0.4130	-	-	-
CJE-T-FD-noPRF	-	0.2133	0.2406	0.2689	0.0485	0.0672
CJE-T-FD-PRF	TSB-C-E-T-01	0.3022↑↑	0.3343↑↑	0.3583↑↑	0.0748	0.0966
CJE-T-PD-PRF	TSB-C-E-T-02	0.2989↑↑	**0.3359**↑↑	**0.3613**↑↑	**0.0829**	**0.1001**
CJE-T-FD-SSR	-	**0.3042**↑↑	0.3356↑↑	0.3589↑↑	0.0771	0.0992
(f) Chinese-English DESCRIPTION runs (pivot runs only)						
Official Top Performer		0.4042	0.4496	-	-	-
CJE-D-FD-noPRF	-	0.2070	0.2275	0.2569	0.0962	0.1248
CJE-D-FD-PRF	TSB-C-E-D-03	0.3032↑↑	0.3365↑↑	0.3609↑↑	0.1448	0.1843
CJE-D-PD-PRF	TSB-C-E-D-04	**0.3411**↑↑⇑	**0.3738**↑↑⇑	**0.3946**↑↑⇑	**0.1711**	**0.2158**
CJE-D-FD-SSR	-	0.2992↑↑	0.3338↑↑	0.3590↑↑	0.1439	0.1838

noPRF: no Pseudo-Relevance Feedback (i.e. initial search)

PRF: Pseudo-Relevance Feedback using the offer weight;

SSR: Selective Sampling with Memory Resetting;

BaT: Bounce-and-Throw using Japanese corpora from 2000-2001 (Yomiuri and Mainichi);

FD: Full Disambiguation in search request translation;

PD: Partial Disambiguation in search request translation (for JE-MT only).

Highest values within each subtask are indicated in bold. Wherever any of our official runs achieved the highest performance among all participants, this is indicated by "(TSB)" in the "Top Performer" row.

Using the two-sided sign test, runs that are significantly better than the corresponding noPRF run are indicated by ↑ ($\alpha = 0.05$) and ↑↑ ($\alpha = 0.01$) in the Arithmetic Mean columns. Those that are significantly better than the corresponding PRF run are indicated by ⇑ ($\alpha = 0.05$).