Toshiba BRIDJE at NTCIR-4 CLIR: Monolingual/Bilingual IR and Flexible Feedback

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

Abstract

Toshiba participated in the Monolingual/Bilingual tasks at NTCIR-4 CLIR using our CLIR system called BRIDJE. 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-D*,* C-J-D*,* C-J-T*,* E-E-D*,* J-E-D*,* J-E-T *subtasks. We had 12 more runs which we were not allowed to submit due to the limitation on the number of runs, with which we would have achieved the highest performances in the* J-J-D*,* C-E-D *and* C-E-T *subtasks as well. Based on our formal run results, this paper discusses (a) the feasibility of the MT-based pivot language approach; (b) the effectiveness of our new Flexible Pseudo-Relevance Feedback methods; and (c) the advantages of Q-measure, which is a new retrieval performance metric based on multigrade relevance.*
Keywords: *BRIDJE, pivot language, Flexible Pseudo-Relevance Feedback, Q-measure.*

1 Introduction

Toshiba participated in the Monolingual/Bilingual tasks at NTCIR-4 CLIR using our CLIR system called BRIDJE [10, 11]. The objectives of our participation this year were: (a) To study the feasibility of the *Pivot Language* (or Transitive Translation) approach using Machine Translation (MT) systems; and (b) To devise new methods for *Flexible Pseudo-Relevance Feedback* [5, 6, 7, 8, 9]. In addition, this paper has a third purpose: (c) To illustrate the advantages of *Q-measure*, which is a new retrieval performance metric based on multigrade relevance [13].

We submitted 24 runs covering three topic languages (Japanese, English, and Chinese) and two document languages (Japanese and English). In addition, there were 12 runs which we generated but did not submit, because only up to two runs were allowed for each language pair and topic field, i.e. TITLE or DESCRIPTION. (We did not submit a fifth run by mixing different topic fields because we believe that this is not practical.) Table 1 provides a summary of our official and unofficial runs. “TOP” indicates that our official performance was the highest among all participants, and “(TOP)” indicates that our *Selective Sampling* runs (See Section 3.3), which we did not submit, actually outperform the official top performers. Thus, we achieved the highest performances in the E-J-D, C-J-D, C-J-T, E-E-D, J-E-D and J-E-T subtasks, and would have achieved the highest performances in the J-J-D, C-E-D and C-E-T subtasks as well with the Selective Sampling runs. We used Japanese as a pivot language for the last two subtasks (See Section 2.1). Throughout this paper, we prefer to use the Unofficial Names listed in the third column of Table 1, as they better reflect the search strategies used.

The remainder of this paper is organised as follows. Section 2 describes the search request translation process of our bilingual runs, including the pivot runs, and briefly discusses their effectiveness. Section 3 introduces two Flexible Pseudo-Relevance Feedback methods and discusses their effectiveness using our monolingual results. It also discusses the advantages of Q-measure as a retrieval performance metric based on multigrade relevance. Finally, Section 4 concludes this paper. We report on our work for the NTCIR-4 QAC2 task in a separate paper [12].

2 Search Request Translation

2.1 BRIDJE and MT

The BRIDJE Cross-Language Information Access System [10, 11] accepts Japanese or English search requests and retrieves documents from Japanese or English text databases using the Okapi/BM25 algorithm [14]. All of our NTCIR-4 runs used the *default* Okapi parameter values [8]: that is, we did not tune the Okapi parameters at all (due to lack of time). Our *traditional* Pseudo-Relevance Feedback (PRF) runs used the *offer weight* (ow) for term selection, with $P = 10$

Table 1. TSB Formal Run Results at NTCIR-4 CLIR.

Topic Field	Official Name	Unofficial Name	Relaxed MAP	Rigid MAP	Description
(a) Monoglingual Japanese runs (55 topics)					
DESC	Top Performer at NTCIR-4		0.4838	0.3804	
	TSB-J-J-D-01	J-J-D-PRF	0.4759	0.3667	Traditional PRF
	TSB-J-J-D-03	J-J-D-TE	0.4683	0.3578	Flexible PRF (Term Exhaustion)
	not submitted	J-J-D-SS	**0.4854** (TOP)	**0.3677**	Flexible PRF (Selective Sampling)
TITLE	Top Performer at NTCIR-4		0.4864	0.3890	
	TSB-J-J-T-02	J-J-T-PRF	0.3863	0.2834	Traditional PRF
	TSB-J-J-T-04	J-J-T-TE	0.3829	0.2802	Flexible PRF (Term Exhaustion)
	not submitted	J-J-T-SS	**0.4538**$\uparrow\uparrow\Uparrow\Uparrow$	**0.3460**$\uparrow\uparrow\Uparrow\Uparrow$	Flexible PRF (Selective Sampling)
(b) English-Japanese runs using E-J MT (55 topics)					
DESC	Top Performer at NTCIR-4		0.3688	0.2674	
	TSB-E-J-D-01	E-J-D-PRF	**0.3688** TOP	0.2672$\Uparrow$	Traditional PRF
	TSB-E-J-D-03	E-J-D-TE	0.3620	0.2615	Flexible PRF (Term Exhaustion)
	not submitted	E-J-D-SS	0.3673	**0.2715**	Flexible PRF (Selective Sampling)
TITLE	Top Performer at NTCIR-4		0.3525	0.2735	
	TSB-E-J-T-02	E-J-T-PRF	0.3244	0.2388	Traditional PRF
	TSB-E-J-T-04	E-J-T-TE	0.3134	0.2284	Flexible PRF (Term Exhaustion)
	not submitted	E-J-T-SS	**0.3486**	**0.2557**	Flexible PRF (Selective Sampling)
(c) Chinese-Japanese runs using C-J MT (55 topics)					
DESC	Top Performer at NTCIR-4		0.3008	0.2309	
	TSB-C-J-D-01	C-J-D-PRF	0.2986	0.2269	Traditional PRF
	TSB-C-J-D-03	C-J-D-TE	**0.3008** TOP	**0.2309** TOP	Flexible PRF (Term Exhaustion)
	not submitted	C-J-D-SS	0.2997	0.2282	Flexible PRF (Selective Sampling)
TITLE	Top Performer at NTCIR-4		0.3193	0.2458	
	TSB-C-J-T-02	C-J-T-PRF	0.3193 TOP	**0.2458** TOP	Traditional PRF
	TSB-C-J-T-04	C-J-T-TE	0.3055	0.2324	Flexible PRF (Term Exhaustion)
	not submitted	C-J-T-SS	**0.3198**	0.2423	Flexible PRF (Selective Sampling)
(d) Monolingual English runs (58 topics)					
DESC	Top Performer at NTCIR-4		0.4368	0.3469	
	TSB-E-E-D-01	E-E-D-PRF	**0.4368** TOP	0.3469 TOP	Traditional PRF
	TSB-E-E-D-03	E-E-D-TE	0.4242	0.3381	Flexible PRF (Term Exhaustion)
	not submitted	E-E-D-SS	0.4366	**0.3510** (TOP)	Flexible PRF (Selective Sampling)
TITLE	Top Performer at NTCIR-4		0.4512	0.3576	
	TSB-E-E-T-02	E-E-T-PRF	**0.4404**	0.3500	Traditional PRF
	TSB-E-E-T-04	E-E-T-TE	0.4274	0.3367	Flexible PRF (Term Exhaustion)
	not submitted	E-E-T-SS	0.4378	**0.3522**	Flexible PRF (Selective Sampling)
(e) Japanese-English runs using J-E MT (58 topics)					
DESC	Top Performer at NTCIR-4		0.4227	0.3340	
	TSB-J-E-D-01	J-E-D-PRF	**0.4227**$*$ TOP	**0.3340** TOP	Traditional PRF
	TSB-J-E-D-03	J-E-D-TE	0.4110	0.3253	Flexible PRF (Term Exhaustion)
	not submitted	J-E-D-SS	0.4105	0.3288	Flexible PRF (Selective Sampling)
TITLE	Top Performer at NTCIR-4		0.4262	0.3407	
	TSB-J-E-T-02	J-E-T-PRF	**0.4262** TOP	**0.3407** TOP	Traditional PRF
	TSB-J-E-T-04	J-E-T-TE	0.4218	0.3369	Flexible PRF (Term Exhaustion)
	not submitted	J-E-T-SS	0.4074	0.3336	Flexible PRF (Selective Sampling)
(f) Chinese-English *pivot* runs using C-J MT and J-E MT (58 topics)					
DESC	Top Performer at NTCIR-4		0.2829	0.2238	
	TSB-C-E-D-01	C-E-D-PRF	0.2767	0.2183	Traditional PRF
	TSB-C-E-D-03	C-E-D-TE	0.2753	0.2169	Flexible PRF (Term Exhaustion)
	not submitted	C-E-D-SS	**0.2862** (TOP)	**0.2303** (TOP)	Flexible PRF (Selective Sampling)
TITLE	Top Performer at NTCIR-4		0.2879	0.2380	
	TSB-C-E-T-02	C-E-T-PRF	0.2873	0.2207	Traditional PRF
	TSB-C-E-T-04	C-E-T-TE	0.2780	0.2114	Flexible PRF (Term Exhaustion)
	not submitted	C-E-T-SS	**0.2969**$\Uparrow$ (TOP)	**0.2370**$\Uparrow$	Flexible PRF (Selective Sampling)

Based on the Sign Test, TE/SS runs that are significantly better than the corresponding PRF run are indicated by $\uparrow$ ($\alpha = 0.05$) and $\uparrow\uparrow$ ($\alpha = 0.01$). PRF/SS runs that are significantly better than the corresponding TE run are indicated by $\Uparrow$ ($\alpha = 0.05$) and $\Uparrow\Uparrow$ ($\alpha = 0.01$). PRF/TE runs that are significantly better than the corresponding SS run are indicated by $*$ ($\alpha = 0.05$) and $**$ ($\alpha = 0.01$). **Boldface** values indicate the best average performance within each language-pair/topic field.

pseudo relevant documents and $T = 40$ expansion terms [6, 7, 8, 10]. The algorithms for generating our *flexible* PRF runs will be described in Section 3.

For our E-J and J-E runs, the search requests were simply translated using the Toshiba MT System as in our previous work [4, 8, 10]. (A more sophisticated search request translation method using MT is described in [11].) For our C-J runs, a Chinese-Japanese MT system that is currently being developed at Toshiba was used for search request translation. As this new system is not yet complete, its translation quality is not as good as our English-Japanese and Japanese-English MT systems. For our C-E runs, we tried a *pivot language* approach instead of using a Chinese-English MT system: The Chinese requests were first translated into Japanese using the new Chinese-Japanese MT system, and the translated requests were further translated into English using our Japanese-English MT system. In short, this is a "Japanese as a pivot language" experiment.

2.2 Analysis of Bilingual Runs

Table 2 shows the *relative* performance values of our cross-language runs based on traditional PRF, where, for example, E-J-D-PRF and C-J-D-PRF are compared with the corresponding monolingual baseline J-J-D-PRF. For the E-J and C-J runs, the percentages are considerably higher for the TITLE runs than for the DESCRIPTION runs, due to the fact that the absolute performance of J-J-D-PRF was much higher than that of J-J-T-PRF. C-J-D-PRF is considerably less effective than E-J-D-PRF because our Chinese-Japanese MT system is not yet as sophisticated as our English-Japanese one. We expect this difference to disappear eventually as we continue to improve our Chinese-Japanese MT system. However, note that no such performance difference is visible for the TITLE runs, i.e., C-J-T-PRF *vs* E-J-T-PRF.

On the other hand, Table 2 shows that our J-E runs are comparable to the monolingual baselines. That is, our Japanese-English MT did an excellent job. Because of this, our pivoted (i.e. C-E) runs are also reasonably successful: the relative performance of C-E-D-PRF is comparable to that of C-J-D-PRF. Recall that our C-E runs were generated by using Chinese-Japanese MT first, and then Japanese-English MT: As the second MT did not introduce much noise, our Chinese-English translations were almost as good as the Chinese-Japanese ones. Note also that our pivot runs are among the very best C-E runs (Table 1 (f)). Thus, our experiments suggest that the Pivot Language approach using *good* MT systems is feasible.

3 Flexible Feedback

3.1 Overview on Flexible Feedback

Traditional PRF relies on at leaset two parameters: P (the number of pseudo-relevant documents scooped from the top of the initial ranked output), and T (the number of expansion terms added to the initial query). Although PRF often improves *average* performance, it typically hurts one-third of a given set of search requests [5]. Various *Flexible PRF* methods have been proposed to enable *per-request adjustment* of these parameters [5, 6, 7, 8, 9], but the results have been somewhat inconclusive. Other researchers have also tackled this problem but without clear success (e.g. [1]).

For NTCIR-4 CLIR, we tried two new Flexible PRF methods for determining P for each search request, both of which are based on which of the query terms occur in the initially retrieved documents. Sections 3.2 and 3.3 describe these methods.

3.2 Term Exhaustion

Our first Flexible PRF method is called *Term Exhaustion*. The idea behind it is simple: Scan the initial ranked output from the top, examining the query terms contained in the retrieved documents. Stop when "novel" query terms (i.e. those that were not in the previous documents) appear to have run out.

Let P_{min} and P_{max} denote the minimum/maximum number of pseudo-relevant documents required, respectively. Then, the problem is to automatically determine, for each topic, P such that $P_{min} \leq P \leq P_{max}$. Let $d(r)$ denote the document at Rank r in the initial ranked output, and let $T(d(r))$ denote the set of initial query terms contained in $d(r)$. The algorithm shown in Figure 1 determines P based on Term Exhaustion. Based on our preliminary Japanese monolingual experiments with the NTCIR-3 test collection, we let $P_{min} = 6$ and $P_{max} = 20$ for *all* NTCIR-4 Term Exhaustion (TE) runs, including the ones with English documents. As for T, we simply let $T = 40$ as in traditional PRF.

3.3 Selective Sampling

Our second method, *Selective Sampling*, is unlike any other Flexible PRF method in that it does not necessarily treat the top P documents as pseudo-relevant. That is, it can *skip* documents. The idea behind it is that there may be similar (and therefore redundant) documents among the top P documents, and it may be better in such a case to go further down the list to look for more "novel" documents.

In addition to P_{min} and P_{max}, we introduce the third parameter called P_{scope}, so that no more than P_{scope} documents are examined. The algorithm shown

Table 2. Relative performance of the cross-language PRF runs.

Unofficial name	Relaxed MAP ratio	Rigid MAP ratio	Unofficial name	Relaxed MAP ratio	Rigid MAP ratio
E-J-D-PRF	77%	73%	E-J-T-PRF	84%	84%
C-J-D-PRF	63%	62%	C-J-T-PRF	83%	87%
J-E-D-PRF	97%	96%	J-E-T-PRF	97%	97%
C-E-D-PRF	63%	63%	C-E-T-PRF	65%	63%

```
T_O = φ;
/* T_O is the set of query terms Observed already. */
i = 0;
/* i is the number of consecutive documents that do not
contain a novel query term. */
for( r = 1; r ≤ P_max; r++ ){
    if( T(d(r)) − T_O == φ ) /* no novel term in d(r)*/
        i++;
    else /* at least one novel term in d(r) */
        i = 0; /* start counting from scratch */
    if( i + 1 == P_min )
        return( r );
    T_O = T_O ∪ T(d(r));
}
return( r );
```

Figure 1. Determining R based on Term Exhaustion.

```
S = φ;
/* S is the set of Sample documents that will be
treated as pseudo-relevant. */
for( r = 1; r ≤ P_scope; r++ ){
    if( is_good_sample_document( r ) )
        S = S ∪ d(r);
    if( |S| == P_max )
        return( S );
}
return( S );

int is_good_sample_document( r )
{
i = 0;
/* i is the number of previously seen documents with
the same set of query terms */
for( r' = 1; r' ≤ r − 1; r'++ )
    if( T(d(r')) == T(d(r)) )
        i++;
if( i < P_min )
    return( 1 ); /* a good sample document */
else
    return( 0 ); /* NOT a good sample document */
}
```

Figure 2. Obtaining the set of pseudo-relevant documents based on Selective Sampling.

in Figure 2 returns a set of pseudo-relevant documents, namely S, obtained through Selective Sampling. Thus, the number of pseudo-relevant documents $P = |S|$. The essence of the algorithm is that it tries to avoid collecting too many documents with the same $T(d(r))$. For NTCIR-4, we used $P_{min} = 3$, $P_{max} = 10$, and $P_{scope} = 50$ for all Selective Sampling (SS) runs, again based on our Japanese monolingual experiments with the NTCIR-3 test collection. As with traditional PRF, we let $T = 40$. However, as mentioned earlier, these runs were not submitted due to the limitation on the number of runs.

3.4 New Evaluation Metrics: Q-measure and R-measure

This section briefly describes Average Weighted Precision (AWP), Q-measure and R-measure which we use in Sections 3.5 and 3.6 for analysing our monolingual Flexible PRF results.

At NTCIR, both *Rigid* and *Relaxed* Mean Average Precision are calculated for performance comparison, as Average Precision cannot handle multiple relevance levels. AWP (originally called *weighted average precision* [3]) proposed by Kando *et al.* can handle multigrade relevance, but has a defect: it does not give a reliable score if relevant documents are ranked below Rank R, where R is the number of known relevant documents. To solve this problem, Sakai [13] has proposed *Q-measure*, which has the reliability of Average Precision *and* the multigrade relevance capability of AWP. Sakai has also proposed *R-measure*, which can be used along with Q-measure just like R-Precision is used besides Average Precision.

Formally, let $gain(X)$ denote the *gain value* for successfully retrieving an X-relevant document. (We let $gain(S) = 3$, $gain(A) = 2$, $gain(B) = 1$ throughout this paper.) Let L denote the size of the ranked output, and let $X(r)$ denote the relevance level of the document at Rank r ($\leq L$). Then, the *gain at Rank* r is given by $g(r) = gain(X(r))$ if the document at Rank r is relevant, and $g(r) = 0$ if it is nonrelevant. The *cumulative gain at Rank* r is given by $cg(r) = g(r) + cg(r-1)$ for $r > 1$ and $cg(1) = g(1)$ [2].

Let $cig(r)$ represent the cumulative gain at Rank r for an *ideal* ranked output. (An ideal ranked output for NTCIR can be obtained by listing up all S-relevant documents, then all A-relevant documents, then all B-relevant documents.) Then, AWP is defined as:

$$AWP = \frac{1}{R} \sum_{1 \le r \le L, g(r) > 0} \frac{cg(r)}{cig(r)} \tag{1}$$

The problem with AWP arises from the fact that $cig(r)$ remains constant for $r \geq R$. That is, AWP cannot discriminate between a relevant document at Rank R and one near the bottom of the ranked list.

Let the *bonused gain at Rank* r be given by $bg(r) = g(r) + 1$ if $g(r) > 0$ and $bg(r) = 0$ if $g(r) = 0$, and its cumulative version be given by $cbg(r) = bg(r) + cbg(r-1)$ for $r > 1$ and $cbg(1) = bg(1)$. Then, Q-measure is defined as:

$$Q\text{-}measure = \frac{1}{R} \sum_{1 \le r \le L, g(r) > 0} \frac{cbg(r)}{cig(r) + r} \tag{2}$$

Q-measure is free from the problem of AWP because the denominator $cig(r) + r$ is guaranteed to increase with r.

Finally, R-measure is defined as:

$$R\text{-}measure = \frac{cbg(R)}{cig(R) + R} \tag{3}$$

Q-measure is equal to one iff a system output (s.t. $L \geq R$) is an ideal one. R-measure is equal to one iff all the top R documents are at least partially relevant. For more detailed discussions, see [13].

3.5 Analysis of Monolingual Runs

Table 3 summarises the results of our monolingual runs using the abovementioned metrics based on multigrade relevance. While the Term Exhaustion results are rather disappointing, the Selective Sampling results are very interesting: In particular, J-J-T-SS easily outperforms J-J-T-PRF, and the difference is statistically significant ($\alpha = 0.01$) with the Sign Test as it is actually better than traditional PRF for around 45 topics out of 55 regardless of the performance metric. Unfortunately, however, the *English* Selective Sampling results are not as straightforward as the Japanese ones. In Section 3.6, we shall investigate the cause of this inconsistency.

Although it is theoretically clear that Q-measure is a more reliable performance metric than AWP [13], we first illustrate its superiority over AWP using actual data. Figure 3 provides a per-topic analysis of J-J-T-SS, which is the most successful Selective Sampling run: Each "circle" represents the value of Q-measure *minus* that of Relaxed Average Precision, while each "cross" represents the value of AWP *minus* that of Relaxed Average Precision. The horizontal axis represents the number of relevant documents R. It is clear that the "circles" are closer to the horizontal axis than the "crosses", and therefore that the property of Q-measure resembles that of Average Precision more than AWP does. Moreover, it is clear that AWP *overestimates* the performance for topics with small R. This is because AWP is unreliable when relevant documents are found below Rank R.

To study the defect of AWP more closely, Table 4 provides some statitistics for Topics 009 and 006, which correspond to the two "crosses" at the top left-hand corner of Figure 3. The table shows that the AWP values are over 0.5 even though Relaxed/Rigid Average Precision values are only around 0.1 and the Q-measure ones are around 0.2. Below, we use Topic 009 to illustate how AWP overestimates performance for topics with small R.

From Table 4, an ideal ranked output for Topic 009 contains S-relevant documents from Rank 1 to 7, A-relevant documents from Rank 8 to 20, and B-relevant documents from Ranks 21 to 23. Therefore, the cumulative gain at Rank $r(\geq 23)$ for this ideal list is $cig(r) = 7*3 + 13*2 + 3*1 = 50$. Table 5 shows exactly how AWP and Q-measure are calculated for Topic 009 with J-J-T-SS, by listing up pertinent statistics for all r such that $g(r) > 0$ (i.e. for every relevant document retrieved). Thus, AWP is calculated by dividing the sum of values in Column 4 by $R = 23$, while Q-measure is calculated by dividing the sum of values in Column 6 by $R = 23$. it is clear that $cg(r)/cig(r)$ is not suitable for calculating retrieval performance: For example, even though the the twenty-third (i.e. the last) relevant document is at Rank 431, $cg(431)/cig(431)$ is equal to one, as if to imply Perfect Precision. In contrast, it can be observed that $cbg(r)/(cig(r)+r)$ penalises relevant documents retrieved at lower ranks.

3.6 Further Analysis of Selective Sampling

Having shown that Q-measure is a reliable evaluation metric, this section uses Q-measure to discuss the conditions under which Selective Sampling may outperform traditional PRF.

In Table 3, Selective Sampling is very successful for the Japanese TITLE run, moderately successful for the Japanese DESCRIPTION run, but only comparable to traditional PRF for the English runs. At first, we suspected that it may have been the particular choice of parameters that caused this inconsistency, but our post-NTCIR-4 experiments showed that this is not the case: Selective Sampling significantly outperforms traditional PRF with the Japanese TITLEs *regardless of the choice of* P and P_{max}.

Further post-NTCIR-4 experiments have suggested that the success of Selective Sampling may depend on at least three factors, namely, (1) *document collection homogeneity* (2) *number of relevant documents* and (3) *query length*.

Factor (1) is a natural candidate for explaining the

Table 3. The monolingual results in terms of Q-measure, R-measure, AWP and R-WP.

Uofficial Name	Relaxed MAP	Rigid MAP	Q-measure	AWP	R-measure
J-J-D-PRF	0.4759	0.3667	0.4823	0.5466	0.4997
J-J-D-TE	0.4683	0.3578	0.4738	0.5360	0.4906
J-J-D-SS	**0.4854**	**0.3677**	**0.4934**	**0.5597**$\Uparrow$	**0.5086**
J-J-T-PRF	0.3863	0.2834	0.4001	0.4725	0.4350
J-J-T-TE	0.3829	0.2802	0.3976	0.4718	0.4309
J-J-T-SS	**0.4538**$\uparrow\uparrow\Uparrow\Uparrow$	**0.3460**$\uparrow\uparrow\Uparrow\Uparrow$	**0.4663**$\uparrow\uparrow\Uparrow\Uparrow$	**0.5385**$\uparrow\uparrow\Uparrow\Uparrow$	**0.4816**$\uparrow\uparrow\Uparrow\Uparrow$
E-E-D-PRF	**0.4368**	0.3469	**0.4539**	**0.5471**	0.4652
E-E-D-TE	0.4242	0.3381	0.4430	0.5367	0.4532
E-E-D-SS	0.4366	**0.3510**	**0.4539**	0.5461	**0.4654**
E-E-T-PRF	**0.4404**	0.3500	**0.4570**$*$	**0.5449**	**0.4717**
E-E-T-TE	0.4274	0.3367	0.4423	0.5275	0.4612
E-E-T-SS	0.4378	**0.3522**	0.4547	0.5378	0.4696

The significance test results are given in the same way as in Table 1.

Table 4. J-J-T-SS **performance values for Topics 006, 009, 044 and 045.**

Topic ID	R	R_S	R_A	R_B	Relaxed	Rigid	Q-measure	AWP
006	15	0	11	4	0.1759	0.1168	0.2500	0.5615
009	23	7	13	3	0.1092	0.0868	0.2017	0.5043

inconsistency between our Japanese and English results, as the NTCIR-4 English document collection is more heterogeneous than its Japanese counterpart in that the former includes documents from six different sources while the latter includes Mainichi and Yomiuri articles only. As Selective Sampling aims at reducing the redundancy in the initial ranked output, it may not work if the initial ranked output already contains a variety of documents. Hence, to reduce the heterogeneity of the NTCIR-4 English collection, we conducted two *subcollection* experiments using the Xinhua (XIE) and Hong Kong Standard (HK) documents, respectively. For example, our XIE subcollection experiments used 53 topics (out of 58) and a "qrels" file that contain XIE documents only. For comparison, we also conducted subcollection experiments with the NTCIR-4 Japanese test collection, by using the Yomiuri and Mainichi subcollections separately.

We found that Selective Sampling significantly outperforms traditional PRF for both of the Japanese subcollections, but not for the English subcollections. However, we suspect that this may be due to Factor (2) mentioned above: The Yomiuri and Mainichi collections have 6,115 and 5,359 relaxed relevant documents in total (around 100-110 per topic), respectively, while the XIE and HK collections have only 2,294 and 3,427 relaxed relevant documents in total (around 40-60 per topic). Thus, the success of Selective Sampling with the Japanese case may have arose from the fact that the NTCIR-4 Japanese document collection is homogeneous *and* that the topics have many relevant documents.

Figure 4 plots the per-topic differences between Selective Sampling and traditional PRF for the Japanese monolingual runs in terms of Q-measure against the number of relevant documents. The abundance of dots above the horizontal axis represents the huge success of Selective Sampling. The graph also suggests that (a) Selective Sampling may not reliably outperform traditional PRF for topics with (say) less than 200 relevant documents; (b) The per-topic performance difference tends to be greater when Selective Sampling outperforms traditional PRF than when traditional PRF outperforms Selective Sampling. Although we cannot observe any such trends from a similar graph for the NTCIR-4 English subtask (not shown here due to lack of space), the English *subcollection* results are a little easier to interpret: For example, Figure 5 shows a similar graph for the XIE subcollection experiments. Observations (a) and (b) seem to apply for this graph as well: For example, the largest positive difference between E-E-T-XIE and E-E-T-PRF is 0.667, while the largest negative difference is only 0.193. Thus, some characteristics of Selective Sampling do seem to have emerged from the English homogeneous subcollection experiments.

As for Factor (3) mentioned above, our Japanese results suggest that Selective Sampling works better with short queries. This is because short queries imply fewer query terms, which in turn imply larger sets of "similar" documents (i.e. documents with the same $T(d(r))$), and therefore more frequent *skipping*. Table 6 shows how many top ranked documents were actually skipped for our NTCIR-4 monolingual Selective Sampling runs. It can be observed that more skipping occurs with TITLEs than with DESCRIPTIONs.

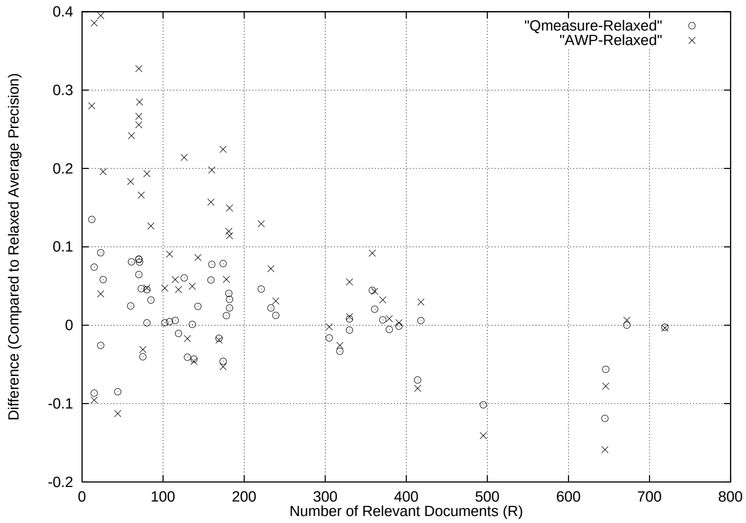

Figure 3. R **vs Q-measure (AWP) minus Relaxed Average Precision (J-J-T-SS).**

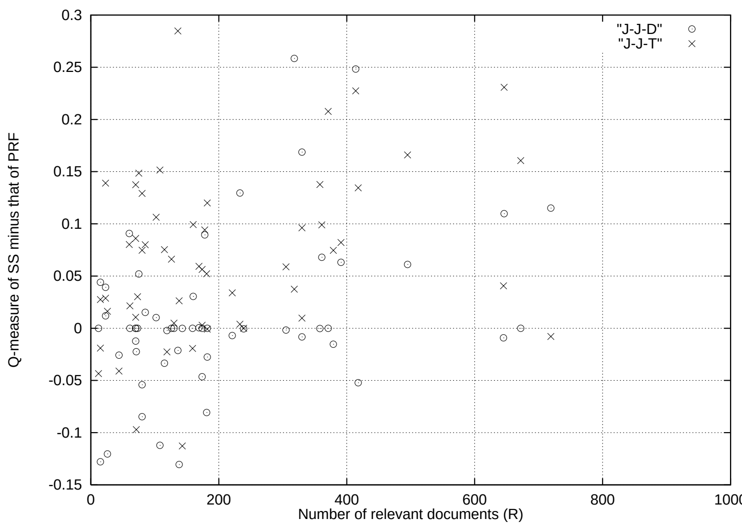

Figure 4. R **vs Q-measure of Selective Sampling minus that of PRF (J-J).**

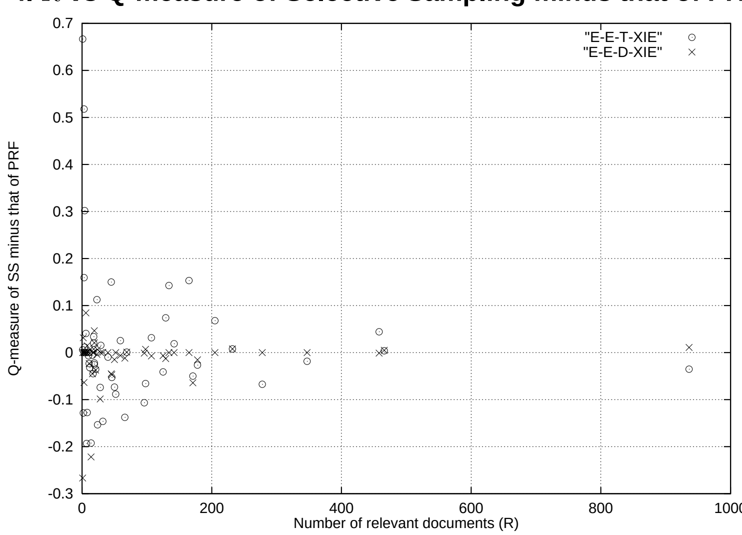

Figure 5. R **vs Q-measure of Selective Sampling minus that of PRF (E-E, Xinhua documents only).**

Table 5. AWP/Q-measure calculation for Topic 009 (J-J-T-SS).

r	$cig(r)$	$cg(r)$	$\frac{cg(r)}{cig(r)}$	$cbg(r)$	$\frac{cbg(r)}{cig(r)+r}$
12	31	1	0.0323	2	0.0465
19	45	3	0.0667	5	0.0781
37	50	4	0.0800	7	0.0805
41	50	6	0.1200	10	0.1099
43	50	9	0.1800	14	0.1505
46	50	11	0.2200	17	0.1771
48	50	14	0.2800	21	0.2143
52	50	16	0.3200	24	0.2353
56	50	19	0.3800	28	0.2642
69	50	21	0.4200	31	0.2605
88	50	23	0.4600	34	0.2464
91	50	26	0.5200	38	0.2695
103	50	28	0.5600	41	0.2680
117	50	30	0.6000	44	0.2635
126	50	32	0.6400	47	0.2670
141	50	34	0.6800	50	0.2618
168	50	37	0.7400	54	0.2477
179	50	38	0.7600	56	0.2445
196	50	41	0.8200	60	0.2439
276	50	43	0.8600	63	0.1933
309	50	45	0.9000	66	0.1838
338	50	48	0.9600	70	0.1804
431	50	50	1.0000	73	0.1518

Table 6. Number of skipped documents averaged over the 55/58 topics.

Unofficial name	#docs skipped
J-J-D-SS	9.3
J-J-T-SS	11.0
E-E-D-SS	6.9
E-E-T-SS	15.9

*The Working Notes Version of the above table showed values averaged over the original 60 topics.

4 Conclusions

Toshiba participated in the Monolingual/Bilingual tasks at NTCIR-4 CLIR. Our main findings are:

1. The "Japanese as a pivot language" approach using *two* MT systems is feasible;

2. Flexible Feedback based on Selective Sampling is effective for the NTCIR-4 *Japanese* test collection, especially with the TITLE fields. It may outperform traditional PRF for (1) homogeneous document collections (2) topics with many relevant documents and (3) short queries;

3. Q-measure is a useful metric for evaluation with multigrade relevance.

As future work, we plan to repeat our experiments using other large-scale Japanese/English test collections and to explore more robust Flexible PRF methods.

References

[1] Billerdeck, B. and Zobel, J.: When Query Expansion Fails, *ACM SIGIR 2003 Proceedings*, pp. 387-388 (2003).

[2] Järvelin, K. and Kekäläinen, J.: IR Evaluation Methods for Retrieving Highly Relevant Documents, *ACM SIGIR 2000 Proceedings*, pp. 41–48 (2000).

[3] Kando, N., Kuriyama, K. and Yoshioka, M.: Information Retrieval System Evaluation using Multi-Grade Relevance Judgments - Discussion on Averageable Single-Numbered Measures (in Japanese), *IPSJ SIG Notes*, FI–63–12, pp. 105-112 (2001).

[4] Sakai, T. *et al.*: A Study on English-to-Japanese / Japanese-to-English Cross-Language Information Retrieval using Machine Translation (in Japanese), *IPSJ Journal*, Vol. 40, No. 11, pp. 4075–4086 (1999).

[5] Sakai, T., Kajiura, M. and Sumita K: A First Step towards Flexible Local Feedback for Ad hoc Retrieval, *IRAL 2000 Proceedings*, pp.95-102 (2000).

[6] Sakai, T., Robertson, S. E. and Walker, S.: Flexible Pseudo-Relevance Feedback for NTCIR-2, *NTCIR-2 Proceedings*, pp.5-(59-66) (2001).

[7] Sakai, T., Robertson, S.E. and Walker, S.: Flexible Pseudo-Relevance Feedback via Direct Mapping and Categorization of Search Requests, *BCS-IRSG ECIR 2001 Proceedings*, pp.3-14 (2001).

[8] Sakai, T.: Japanese-English Cross-Language Information Retrieval Using Machine Translation and Pseudo-Relevance Feedback, *IJCPOL*, Vol. 14, No. 2, pp. 83–107 (2001).

[9] Sakai, T. and Robertson, S.E.: Flexible Pseudo-Relevance Feedback Using Optimization Tables, *ACM SIGIR 2001 Proceedings*, pp.396-397 (2001).

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

[11] 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).

[12] Sakai, T. *et al.*: Toshiba ASKMi at NTCIR-4 QAC2, *NTCIR-4 Proceedings*, to appear (2004).

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

[14] 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).