Toshiba KIDS at NTCIR-3: Japanese and English-Japanese IR

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

Abstract

Toshiba participated in the Japanese monolingual track and the English-Japanese cross-language track at NTCIR-3, and achieved the highest retrieval performances among the DESCRIPTION runs. Our pre-submission experiments using the KIDS system focused primarily on effective Japanese monolingual retrieval, and addressed several problems including term selection enhancement and ranked output combination. Moreover, our post-submission reruns using the recently developed BRIDJE system substantially outperformed our official runs. In addition, our post-submission experiments include a new cross-language run which used the internal translation data of our machine translation system, which further improved performance on average.

Keywords: *KIDS, BRIDJE, pseudo-relevance feedback, ranked output combination, machine translation.*

1 Introduction

Toshiba participated in the Japanese monolingual track and the English-Japanese cross-language track at NTCIR-3. Three automatic DESCRIPTION runs were submitted to each track. (All runs described in this paper are automatic DESCRIPTION runs.)

At NTCIR-1, Toshiba participated in the Japanese-English cross-language track only [9]. At NTCIR-2, Toshiba (the first author) collaborated with Microsoft Research Cambridge to participate in the same track [10]. Thus this is our first year at NTCIR in which we dealt with Japanese document retrieval.

At *dry-run*, we achieved the highest performances among all runs in both Japanese monolingual and English-Japanese tracks, in terms of both *rigid* and *relaxed* relevance. As for our *official* runs:

- Our three English-Japanese runs `TSB-E-J-D-0[123]` were by far the best among all English-Japanese runs (including non-DESPCRIPTION runs), in terms of both *rigid* and *relaxed* relevance.
- In terms of *rigid* relevance, our Japanese monolingual run `TSB-J-J-D-01` achieved the highest performance among all Japanese DESCRIPTION runs. (It came third among *all* Japanese runs, but the top two runs used TITLE,NARRATIVE and CONCEPT fields in addition.)
- In terms of *relaxed* relevance, our Japanese monolingual run `TSB-J-J-D-03`, which achieved the mean average precision of 0.3910, came third among all Japanese DESCRIPTION runs. Although the top two runs achieved 0.3998 and 0.3946, respectively, the differences among these three runs are *not* statistically significant. (For example, `TSB-J-J-D-03` outperformed the best run for as many as 21 topics out of 42.)

Thus, our official runs are clearly of the highest standard. We will show that they can be improved even further in our post-submission experiments.

Our *pre-submission* experiments using the *KIDS retrieval system* [9, 17] focused primarily on effective Japanese monolingual retrieval, and addressed several problems including *term selection enhancement* and *ranked output combination*. Our *post-submission* experiments reran our official search strategies using the recently developed *BRIDJE retrieval system* [14, 16] that uses the well-known Okapi/BM25 formula [7, 11]. In addition, our post-submission experiments include a new cross-language run which uses the *internal translation data* of the *Toshiba machine translation (MT) system* [1, 16], which improves average performance considerably.

The remainder of this paper is organised as follows. Section 2 describes the KIDS/BRIDJE retrieval systems. Section 3 describes our pre-submission experiments using KIDS, and Section 4 provides our official results. Section 5 describes our post-submission experiments using BRIDJE. Finally, Section 6 concludes this paper. In addition. the Appendix contains a "separate paper", which attempts to improve existing performance measures that use multiple relevance levels.

Throughout our experiments, TREC Mean Average Precision was used as our primary retrieval performance measure, and the sign test was used for testing statistical significance.

2 KIDS/BRIDJE

KIDS (Knowledge and Information on Demand System) [9, 17] is a Japanese monolingual retrieval system based on morphological analysis. BRIDJE (Bi-directional Retriever/Information Distiller for Japanese and English) [14, 16] is a recently developed, bilingual/cross-language extension of KIDS. BRIDJE has several functions which are useful for cross-language information retrieval (CLIR), such as *synonym operator* handling and *transliteration* [16].

BRIDJE employs the Okapi/BM25 term weighting and PRF schemes, as described fully in [7, 11]. Thus, a document score is a sum of *combined iterative weights* (ciw), where ciw for a term-document pair is computed as the product of Equations (1) and (2):

$$rw = \log \frac{(r+0.5)(N-n-R+r+0.5)}{(n-r+0.5)(R-r+0.5)} \qquad (1)$$

$$\frac{tf*(k1+1)}{k1*((1-b)+b*ndl)+tf} \qquad (2)$$

where

$N =$ number of documents in the collection;

$n =$ number of documents containing the term;

$R =$ number of known/pseudo relevant documents;

$r =$ number of known/pseudo relevant documents containing the term.

$tf =$ number of occurrences of the term within the document;

$ndl =$ normalised document length [7, 11];

$k1, b =$ BM25 parameters [7, 11].

KIDS uses same algorithms as those of BRIDJE except that it uses a simpler term frequency component. That is, it uses Equation (3) instead of (2):

$$0.75 + 0.25 * tf/tf_{max} \qquad (3)$$

where

$tf_{max} =$ maximum number of occurrences of any term within the document.

Thus, our official runs used Equation (3), while our post-submission runs used Equation (2).

At NTCIR-3, KIDS/BRIDJE were used as “bag-of-words” systems. Although KIDS/BRIDJE can perform *semantic role analysis* so as to promote documents that contain specific word patterns [14], this function was not used as our primary objective at NTCIR-3 was to retrieve as many relevant documents as possible, including partially relevant ones.

3 Pre-Submission Experiments

Our *pre-submission* experiments using the KIDS retrieval system focused primarily on effective Japanese monolingual retrieval: our English-Japanese runs simply copied our monolingual strategies after search request translation using the Toshiba MT system. The following subsections summarise various search strategies which we explored, namely, document reranking, term selection enhancement and ranked output combination. Our pre-submission experiments used the Japanese/English dry-run topic sets with the NTCIR-3 Japanese document collection, through which we selected promising search strategies and parameter settings.

3.1 Document Reranking

Because our term weighting schemes do not utilise term proximity/co-occurrence, we considered reranking the *initial* (i.e. pre-PRF) ranked output based on *within-sentence co-occurrence* of initial search terms: For a document d, let $s(d)$ denote the number of sentences which contain any two distinct initial search terms. If $s(d) > 0$, then let

$$score_{new}(d) = (1 + p * \log_2(1 + s(d))) * score_{old}(d) \qquad (4)$$

We used $p = 0.1$ throughout our experiments. If $s(d) = 0$, then we let $score_{new}(d) = score_{old}(d)$.

3.2 Term Selection Enhancement

3.2.1 Relative/Absolute Criteria

Based on the arguments in [6], the *offer weight* (ow) is often used as the term selection criterion in Okapi/BM25-based PRF [7, 11]:

$$ow = r * rw \qquad (5)$$

Whereas, Sakai and Robertson [13] have suggested the following alternative to ow, based on their regression analysis experiments:

$$ow2 = \sqrt{r} * rw \qquad (6)$$

While the above two criteria treat all (pseudo) relevant documents equally, it may be better to incorporate the *rank* information or even the actual *document scores* into the criteria, so that documents at the very top of the initial ranked output can be weighted more heavily. Thus, for each candidate expansion term t, we define two new term selection criteria as follows:

$$ow3 = sr * rw \qquad (7)$$

$$ow4 = \sqrt{sr} * rw \qquad (8)$$

$$sr = \sum_{d \in rel} score(d) \qquad (9)$$

where rel is the set of (pseudo) relevant documents containing t. Since $|rel| = r$, $ow3$ and $ow4$ reduce to ow and $ow2$, respectively, when $score(d) = 1$ for every (pseudo) relevant document.

The term selection criteria described so far are *relative* criteria, i.e. they are only for *ranking* candidate expansion terms. In contrast, *chi-square* is an *absolute* criterion, which is also known to be effective:

$$\chi^2 = \frac{N(r(N-R-n+r)-(R-r)(n-r))^2}{R(N-R)n(N-n)} \tag{10}$$

Since the chi-square values are comparable across topics, a fixed *threshold* (c) can be used as a parameter instead of the number of expansion terms (T) [13]. Thus, with chi-square, the number of expansion terms can vary across topics.

3.2.2 Kanji Overlap Promotion

For Japanese IR, the term selection selection criteria described above may possibly be enhanced further via *Kanji Overlap Promotion* (KOP). Let K_I denote the set of distinct kanji characters extracted from an Initial query, and let K_E denote the set of distinct kanji characters extracted from a candidate Expansion term t. We define $k = |K_I \cap K_E|$ as the number of *kanji overlaps* for t. Then, for example, $ow4$ can be generalised as follows:

$$ow4k = (1 + p_{KOP} * k) * ow4 \tag{11}$$

We used $p_{KOP} = 0.5$ throughout our experiments, which heavily favours expansion terms that contain the same kanji characters as those of the initial terms. For example, for the dry-run topic 006 (English translation: "diagnosis and therapy of breast cancer"), the initial Japanese search terms were "*nyu-gan*" (breast cancer), "*shin-dan*" (diagnosis) and "*chi-ryo*" (therapy). (Here, the kanji terms are represented in *romaji*, and the boundaries between two kanjis are represented by "-".) For this topic, KOP successfully replaced two inappropriate expansion terms with appropriate ones, namely, "*chi-busa*" (breast) and "*sho-shin*" (first diagnosis). This is because "*sho-shin*" overlaps with "*shin-dan*", while "*chi-busa*" overlaps with "*nyu-gan*" as "*chi*" and "*nyu*" in fact represent the same kanji (which means "breast" or "milk"). KOP was motivated by the fact that many kanji words are like concise summaries: For example, from "*ji-shin*" (earthquake) and "*sai-gai*" (disaster), we obtain "*shin-sai*" (earthquake disaster), a perfectly legitimate kanji word. However, since most kanji characters are polysemous, the effect of KOP may not always be positive. The question was whether KOP is beneficial in terms of overall performance or not.

3.3 Ranked Output Combination

In Sakai's experiments with TREC data [12], combining the ranked output from full-text and summary indexes improved retrieve performance significantly. We therefore created a "summary" index using only the *title* and the *first sentence* of each document, along with a separate full-text index, in order to combine the runs using these two indexes. In effect, this is like exploiting the *document structure* of Japanese newspaper articles [8]. We combined component document scores simply by taking a weighted average:

$$score_{merged}(d) = \sum_i m_i * score_i(d) \tag{12}$$

where i represents the i-th ranked output, and $\sum_i m_i = 1$. Optionally, the component document scores can be normalised prior to merging:

$$score_{normalised}(d) = \frac{score(d) - score_{min}}{score_{max} - score_{min}} \tag{13}$$

where $score_{max}$ and $score_{min}$ are the maximum/mininum document scores in the ranked output, respectively.

Sakai [15] showed that *parallel pseudo-relevance feedback* (parallel PRF) may be comparable to *collection enrichment* [4] for retrieval of English documents. Parallel PRF searches an *external* document collection to perform query expansion, and the expanded query is used to search the target collection. Finally, the above run is combined with a traditional PRF run using Equations (12) and optionally (13). Thus, Parallel PRF employs a parallel, ranked output combination approach in contrast to *reference database feedback* which serially expands a single query [2]. At NTCIR-3, we re-examined parallel PRF in the context of *Japanese* IR by using Mainichi Newspaper articles from the year 2000 as an external collection.

4 Official Results

This section summarises our official NTCIR-3 results. We firstly describe the search *strategies* we selected through our pre-submission experiments, three of which were used to generate the official runs.

F: This is a basic strategy that uses the Full-text index, with document reranking. For PRF, $ow4k$ is used as the term selection criterion.

Fχ^2: This is the same as F except that χ^2 is used as the term selection criterion.

S: This strategy uses the aforementioned "Summary" index, using $ow4k$ for PRF without document reranking. The summary index is used for the initial search *and* the final search [12].

F+S, Fχ^2+S: Ranked output combination of F (Fχ^2) and S, without score normalisation. That is, these strategies combine a full-text run and a "summary" run [12].

Table 1. Official results (42 NTCIR-3 topics).

Strategy	RunID (?=[J|E])	J-J (relaxed)	E-J (relaxed)	J-J (rigid)	E-J (rigid)
F	-	0.3631	0.3055	0.3131	0.2660
Fχ^2	-	0.3608	0.3052	0.3074	0.2630
S	-	0.2570	0.2224	0.2243	0.1848
F+S	TSB-?-J-D-01	**0.3903**$\star$	**0.3394**$\ast\ast\star$	**0.3457**$\star$	**0.2916**$\ast\star$
Fχ^2+S	-	**0.3889**$\star$	**0.3366**$\ast\ast\star\star$	**0.3399**	**0.2885**$\ast\ast\star$
E	-	0.3291	0.2694	0.2766	0.2307
F+E	-	0.3612	0.2989	0.3099	0.2598
Fχ^2+E	-	0.3610	0.3002	0.3064	0.2619
C	-	0.3605	0.2925	**0.3191**	0.2589
F+S+E	TSB-?-J-D-02	**0.3904**$\ast\star$	**0.3404**$\ast\ast\star$	**0.3432**	**0.2926**$\ast\ast\star$
Fχ^2+S+E	TSB-?-J-D-03	**0.3910**$\star\star$	**0.3378**$\ast\ast\star\star$	**0.3416**	**0.2891**$\ast\ast\star\star$

Values in **boldface**: those that outperform F *on average*;
Values with $\ast$, $\ast\ast$: those that are significantly better than F ($\alpha = 0.05, 0.01$);
Values with $\star$, $\star\star$: those that are significantly better than Fχ^2 ($\alpha = 0.05, 0.01$).

E: This strategy uses the aforementioned External collection for initial search and query expansion, and the target collection for final search. It uses χ^2 for PRF after document reranking. That is, this is the "external" component of parallel PRF [15].

F+E, Fχ^2+E: Ranked output combination of F (Fχ^2) and E, with score normalisation. That is, these are *parallel PRF* strategies.

C: This is *Collection enrichment*, for comparison with parallel PRF. An index that consists of the target collection *plus* the aforementioned external collection is used for initial search with document reranking and query expansion using $ow4k$. The target collection index is used for final search.

F+S+E, Fχ^2+S+E: Ranked output combination of F (Fχ^2), S and E without score normalisation.

Table 1 summarises our official NTCIR-3 results. Here, a combination of a *strategy* and a *language pair* gives a particular *run*. For example, F+S (J-J) corresponds to an official run known as TSB-J-J-D-01. The upper half of Table 2 shows the parameter values that we used.

The findings from our official results can be summarised as follows:

1. By comparing F and Fχ^2, we can observe that our new *relative* term selection criterion $ow4k$ is at least as effective as the *absolute* χ^2 criterion. (The differences between F and Fχ^2 are not statistically significant.)

2. By comparing F+S (Fχ^2+S) with F (Fχ^2), it is clear that our full-text/summary ranked output combinations are effective. As indicated in Table 1, these gains are mostly statistically significant. This is in agreement with Sakai's results with the TREC English data [12].

3. By comparing F+E (Fχ^2+E) and C with F (Fχ^2), we can observe that neither parallel PRF nor collection enrichment is successful. (Although C appears to outperform F in the "J-J (rigid)" column, C actually hurts as many as 23 topics out of 42.) Possibly, these methods might have been more successful if we used a newspaper articles *other than Mainchi* from 1998-1999 as an external collection, since "topical overlap" is probably more important than "stylistic resemblance" between the target and the external document collections. (Unfortunately, we did not have such a document collection at hand.)

4. By comparing F+S (Fχ^2+S) and F+S+E (Fχ^2+S+E), we can observe that the effect of adding the "external" component E is very small. Nevertheless, this may be beneficial in the sense that the gain over F (Fχ^2) is generally more highly significant for F+S+E (Fχ^2+S+E) than for F+S (Fχ^2+S), as indicated by the "$\ast$"s and "$\star$"s. That is, adding E appears to *stabilise* the positive gain obtained by the full-text/summary combination.

5. Although not indicated in Table 1, neither document reranking nor kanji overlap promotion had any significant positive effect. Although document reranking does consistently improve the *initial* retrieval performance, we found that its effect becomes negligible after PRF. As for kanji overlap promotion, it worked in some cases but not in others due to the aforementioned polysemous nature of kanji.

Table 2. Parameter values.

F	$R = 15, T = 30$
Fχ^2	$R = 15, c = 300$
S	$R = 15, T = 40$
E	$R = 5, c = 200$
F+S	$m_F = 0.7, m_S = 0.3$
Fχ^2+S	$m_{F\chi^2} = 0.7, m_S = 0.3$
F+E	$m_F = 0.7, m_E = 0.3$
Fχ^2+E	$m_{F\chi^2} = 0.7, m_E = 0.3$
C	$R = 15, T = 50$
F+S+E	$m_F = 0.6, m_S = 0.3, m_E = 0.1$
Fχ^2+S+E	$m_{F\chi^2} = 0.6, m_S = 0.3, m_E = 0.1$
F$'$, E$'$, C$'$	$k1 = 1.2, b = 0.25$ (initial search), $b = 0.50$ (final search)
S$'$	$k1 = 1.2, b = 0.75$ (default)
F+S$'$	$m_F = 0.9, m_S = 0.1$
Fχ^2+S$'$	$m_{F\chi^2} = 0.9, m_S = 0.1$
F+E$'$	$m_F = 0.7, m_E = 0.3$
Fχ^2+E$'$	$m_{F\chi^2} = 0.7, m_E = 0.3$
F+S+E$'$	$m_F = 0.7, m_S = 0.1, m_E = 0.2$
Fχ^2+S+E$'$	$m_{F\chi^2} = 0.7, m_S = 0.1, m_E = 0.2$

5 Post-Submission Experiments

Our *post-submission* experiments using BRIDJE include reruns that correspond to our official runs, as well as a *new* English-Japanese cross-language run. This new run utilises a recently implemented function of BRIDJE: during the search request translation process, BRIDJE can now access the *internal translation data* of the Toshiba MT system. Thus, while it was not possible with the traditional MT approach to obtain more than one translation for each source language term, BRIDJE can obtain sets of "synonyms" in target language. Although the use of such internal translation data has been explored by Jones *et al.* [3], they included the translation candidates as *distinct* search terms in the queries, which hurt retrieval performance considerably. In contrast, we treat a set of "synonyms" as a *single* term using the synonym operator, as this strategy appears to be useful for CLIR [5].

Section 5.1 reruns our pre-submission experiments by replacing KIDS with BRIDJE, still primarily focusing on Japanese monolingual retrieval. Section 5.2 describes our new CLIR experiments using internal translation data.

5.1 Reruns using BRIDJE/BM25

Table 3(a) shows the results of our post-submission reruns using BRIDJE. Unlike our official runs, our reruns did not use document reranking and kanji overlap promotion. Thus, for example, F$'$ corresponds to F, but $ow4$ was used instead of $ow4k$ (See Equation (11)). The BM25 parameters and the merging factors m_i (See Equations (2) and (12)) were optimised for the 6 dry-run topics (*not* for the 42 test topics). The details are given in the bottom half of Table 2. All other parameter values were copied from our pre-submission experiments. For a preliminary comparison with Mean Average Precision (MAP), Table 3 also contains *Mean Average Gain Ratio* (MAGR) values as well: See the Appendix for further details.

The results of our re-runs are quite different from those of our official ones. Our new findings can be summarised as follows:

1. BRIDJE is much more effective than KIDS. The differences between F and F$'$ and those between Fχ^2 and F$'\chi^2$ are all statistically significant. Thus, it is clear that the term frequency component of BM25 (See Equation (2)) is very important. (The differences between F$'$ and F$'\chi^2$ are not statistically significant in terms of MAP.)

2. Unlike our official runs, the full-text/summary combinations are *not* successful. Because of this, F+S+E$'$ and Fχ^2+S+E$'$ are not successful either. (Although some runs appear to outperform F$'$, these differences are not statistically significant.) This probably means that the full-text/summary combination does not help when the full-text component alone is already very good. It is also worth noting that the differences between S and S$'$ are very small: that is, BM25 does not handle the summary index very well.

3. Parallel PRF and collection enrichment appear to be more successful in our post-submission experiments than in our pre-submission ones, although none of the gains over F (Fχ^2) are statistically significant. Moreover, collection enrichment appears to be a little more successful than parallel PRF here, as C$'$ significantly outperforms F+E$'$ in the "J-J (relaxed)" case, as well as Fχ^2+E$'$ in the "J-J (rigid)" case. However, as this tendency could not be observed in Table 1, these results are not conclusive.

We now take a closer look at our term selection criteria. Table 4 compares F$'$ and F$'\chi^2$ with runs that used other term selection criteria instead. For example, F$'ow$ represents the *traditional* PRF strategy using the offer weight. The performance differences are very small, and the gains over F$'ow$ are not statistically significant. Thus, although the "⇑"s in Table 4 suggest that using $ow4$ instead of ow was a good choice for NTCIR-3, our results do not fully confirm that incorporating the document score into the term selection criterion is effective. This may be because the initial document scores do not reflect the true probability of relevance with accuracy.

Table 3. Reruns using BRIDJE/BM25 (42 NTCIR-3 topics).

	Strategy	J - J	E - J	J - J	E - J	J - J	E - J
		(relaxed)	(relaxed)	(rigid)	(rigid)	(MAGR)	(MAGR)
(a)	F'	0.4308	0.3575	0.3715	0.3103	0.6130$\star$	0.5561
	$F'\chi^2$	0.4284	0.3571	0.3641	**0.3121**	0.6084	0.5553
	S'	0.2711	0.2258	0.2370	0.1883	0.3710	0.3205
	$F+S'$	**0.4312**	**0.3587**	0.3707	**0.3117**	0.6118	0.5558
	$F\chi^2+S'$	**0.4312**	**0.3591**	0.3673	0.3083	0.6111	0.5551
	E'	0.4197	0.3224	0.3583	0.2654	0.5971	0.5099
	$F+E'$	**0.4346**	0.3546	**0.3726**	0.3078	**0.6147**	0.5527
	$F\chi^2+E'$	**0.4331**	0.3560	0.3682	0.3067	0.6110	0.5539
	C'	**0.4440**	**0.3588**	**0.3805**	**0.3142**	**0.6340**	**0.5632**
	$F+S+E'$	**0.4327**	0.3572	**0.3724**	0.3084	**0.6138**	0.5541
	$F\chi^2+S+E'$	**0.4330**	**0.3585**	0.3705	0.3070	0.6127	0.5545
(b)	F'-internal	–	**0.3846**	–	**0.3365**	–	**0.5835**

Values in **boldface**: those that outperform F' *on average*.
Values with $\star$: those that are significantly better than $F'\chi^2$ ($\alpha = 0.05$).

Table 4. Comparison of term selection criteria (42 NTCIR-3 topics).

Strategy	J - J (relaxed)	E - J (relaxed)	J - J (rigid)	E - J (rigid)
$F'ow$	0.4318	0.3571	0.3636$\Downarrow$	0.3053$\Downarrow$
$F'ow2$	0.4303$\Downarrow$	0.3574	0.3690	0.3134$\Uparrow$
$F'ow3$	0.4327$\Uparrow$	0.3569$\Downarrow$	0.3647	0.3055
F' (i.e. $F'ow4$)	0.4308	0.3575$\Uparrow$	0.3715$\Uparrow$	0.3103
$F'\chi^2$	0.4284	0.3571	0.3641	0.3121

Values with $\Uparrow$: highest value within column;
Values with $\Downarrow$: lowest value within column.

5.2 New Cross-Language Runs using Internal Translation Data

Our pre-submission and post-submission experiments described so far have focused primarily on Japanese monolingual retrieval: the cross-language runs were mere copies of the monolingual strategies. In contrast, this subsection addresses the cross-language problem directly, by utilising the *internal translation data* of the Toshiba MT system.

The Toshiba MT system employs the *transfer method*, and its disambiguation mechanism consists of several stages [1, 3]. In our experiments, we accessed the internal translation data between the *semantic analysis stage* and the *final disambiguation stage*, where the latter means obtaining single best translation candidates for full machine translation. In this way, a set of "synonyms" were obtained for each source language term. For example, from "prime minister" (Topic 022), we obtained "*shu-sho*", "*so-ri-dai-jin*" and "*so-ri*", which are all correct translations. From "pitcher" (Topic 016), we obtained "*to-shu*" (correct translation) and "*piccha*" (*katakana transliteration* of pitcher). These were included in the initial query using the synonym operator. On the other hand, since semantic analysis is not always perfect, the "synonym" set may contain inappropriate translations as well. For example, from "Turkey (the country)" (Topic 021), we obtained not only "*toruko*" (correct) but also "*shichimen-cho* (the bird)"! (Needless to say, the Toshiba MT system successfully filtered out the latter in the final disambiguation stage.) Another interesting example is Topic 017, for which several different *kanji* spellings of "Takeshi (Kitano, the film director)" were obtained, along with a *katakana* transliteration. Unfortunately, the correct *kanji* spelling for that particular Takeshi Kitano was not among them.

Table 3(b) shows the performance of our new cross-language run, denoted by F'-internal. Thus, just like F', it uses the full-text index only, using $ow4$ as the term selection criterion. Although the differences between F' and F'-internal are not statistically significant, it can be observed that the use of internal translation data improves the average retrieval performance considerably. (F'-internal is the best run among all cross-language runs described in this paper, and therefore outperforms *all* official English-Japanese runs at NTCIR-3 by far). The lack of statistical significance is due to the aforementioned noise in the "synonym" sets: obtaining synonym sets of better quality for more effective CLIR will be one of the subjects of our future work.

6 Conclusions

This paper described our pre-submission and post-submission experiments at NTCIR-3. Our pre-submission experiments used the KIDS retrieval system, focusing primarily on Japanese monoligual retrieval: our official runs achieved the highest retrieval performances among the DESCRIPTION runs. Our post-submission experiments further improved our retrieval performances using the BRIDJE system with BM25 term weighting, and examined a new cross-language run in addition. Our main findings can be summarised as follows:

1. BRIDJE significantly outperforms KIDS. Thus, the term frequency component of BM25 is very effective.

2. Our *relative* term selection criteria are at least as effective as χ^2, which is an *absolute* criterion. However, the effect of incorporating the document score into the criterion is not clear.

3. Combining the ranked output from a full-text index and a summary index may significantly improve performance, but it may not work when the full-text component is already very effective.

4. Using the Mainichi newspaper 2000 collection as an external collection at NTCIR-3 for parallel PRF and collection enrichment is probably not a very good idea.

5. Using the internal translation data of our machine translation system for cross-language information retrieval improves average retrieval performance considerably and deserves further investigation.

Acknowledgment

The authors would like to thank the NTCIR-3 organizers for their effort, and Akira Kumano for his help in our internal translation data experiments.

References

[1] Amano, S., *et al.*: The Toshiba Machine Translation System, Future Computing Systems, Vol. 2, No. 3 (1989).

[2] Fujita, S.: Reflections on “Aboutness” TREC-9 Evaluation Experiments at Justsystem, *TREC-9 Proceedings* (2001).

[3] Jones, G. *et al.*: A Comparison of Query Translation Methods for English-Japanese Cross-Language Information Retrieval, *ACM SIGIR ’99 Proceedings*, pp. 269–270 (1999).

[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] 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).

[6] Robertson, S. E.: On Term Selection for Query Expansion, *Journal of Documentation* 46, pp. 359–364 (1990).

[7] Robertson, S. E. and Sparck Jones, K: Simple, Proven Approaches to Text Retrieval, Computer Laboratory, University of Cambridge (1997).

[8] Sakai, T., Kajiura, M. and Sumita, K.: Generation and Evaluation of Search Queries using Boolean Expressions and Document Structure for Information Filtering (*in Japanese*), *IPSJ Journal*, Vol. 39, No. 11 pp. 3076–3083 (1998).

[9] Sakai, T. *et al.*: Cross-Language Information Retrieval for NTCIR at Toshiba, *NTCIR-1 Proceedings*, pp. 137–144 (1999) `http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings/017-IR-Sakai.pdf`

[10] Sakai, T., Robertson, S. E. and Walker, S.: Flexible Pseudo-Relevance Feedback for NTCIR-2, *NTCIR-2 Proceedings*, 2001, pp.5-(59-66). `http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings2/sakai.pdf`

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

[12] Sakai, T.: Combining the Ranked Output from Fulltext and Summary Indexes, *ACM SIGIR 2001 Workshop on Text Summarization*, pp. 27–34 (2001). `http://www-nlpir.nist.gov/projects/duc/duc2001/agenda_duc2001.html`

[13] Sakai, T. and Robertson, S. E.: Relative and Absolute Term Selection Criteria: A Comparative Study for English and Japanese, *ACM SIGIR 2002 Proceedings*, pp. 411–412 (2002).

[14] Sakai, T. *et al.*: Retrieval of Highly Relevant Documents based on Semantic Role Analysis (*in Japanese*), *FIT 2002 Information Technology Letters*, pp. 67–68 (2002).

[15] Sakai, T.: The Use of External Text Data in Cross-Language IR based on Machine Translation, *IEEE SMC 2002 Proceedings* (2002).

[16] Sakai, T., Kumano, A. and Manabe, T.: Generating Transliteration Rules for Cross-Language Information Retrieval from Machine Translation Dictionaries, *IEEE SMC 2002 Proceedings* (2002).

[17] Sumita, K. *et al.*: Know-How Sharing Using a Knowledge Sharing System KIDS – A Knowledge Management Practice at a Research Labaratory – *Practical Aspects of Knowledge Management 2000*, pp.21.1–21.7 (2000).

Appendix: Average Gain Ratio based on Multiple Relevance Levels

This Appendix discusses an attempt at improving existing retrieval performance measures that use multiple relevance levels. These modified measures may be useful for NTCIR evaluations.

Järvelin and Kekäläinen [1] proposed (discounted) *cumulative gain* for evaluation with multiple relevance levels. Their basic idea is that an imaginary user scans the ranked output from the top, and "gains" an additional score each time he finds a relevant document. The gain is large if the document is highly relevant, and small if it is only partially relevant. Thus, let l denote a *relevance level*, and let $gain_l$ denote the *gain* obtained by finding an l-relevant document. For example, let $gain_S$=3, $gain_A$=2 and $gain_B$=1 for NTCIR. Moreover, let $l(i)$ denote the relevance level of the document retrieved at rank i. (Here, $1 \leq i \leq j \leq J$, where j is the actual size of the ranked output, and J is the maximum size allowed.) Then, the *gain at rank* i is given by $g(i) = gain_{l(i)}$, and the *cumulative gain at rank* i is given by $cg(i) = g(i) + cg(i-1)$ $(i > 1)$, $cg(1) = g(1)$.

Just like *precision* at a fixed document cut-off, cumulative gains are averaged across topics on a *document-rank* basis. However, as Kando *et al.* [2] have pointed out, *rank-based* averaging is *not* good for sound retrieval performance evaluation, as the total number of relevant documents differ across topics, and therefore the upperbound performance at a document rank differs across topics as well. From this perspective, Kando *et al.* have proposed "*Weighted Average Precision*" (WAP) [2], which is suitable for *recall-based* averaging just like TREC Average Precision. Thus, let $icg(i)$ (> 0) denote the *ideal* cumulative gain at rank i obtained from the "best possible" ranked output [1, 2]. Then, WAP is given by:

$$WAP = \frac{1}{R} \sum_{1 \leq i \leq j, g(i) > 0} \frac{cg(i)}{icg(i)} \qquad (14)$$

where R is the number of *all* relevant documents.

Now, let R_l denote the number of l-relevant documents. Although WAP is probably a better measure than the raw cumulative gain as it absorbs the variance of R across topics, note that it does not explicitly reflect the variance of R_l for each l: In general (though not always), the number of highly relevant documents is considerably smaller than that of partially relevant ones. Whenever this is the case, WAP is affected primarily by how well the *partially* relevant documents are retrieved, even though the most important aspect that we would like to evaluate is how well the few *most* relevant ones are retrieved. For example, suppose that $R_S = 1, R_B = 9$ and therefore $R = 10$ for a topic. Let $gain_S = 3, gain_B = 1$ and suppose that a ranked output has the S-relevant document at rank 1 and nonrelevant documents at all other ranks. Since $cg(1) = icg(1) = 3$ in this case, $WAP = (3/3)/10 = 0.1$. On the other hand, consider a ranked output which has all of the 9 B-relevant documents at the very top and nothing else. In this case, $cg(i)$ $(1 \leq i \leq 9)$ are $1, 2, 3, 4, 5, 6, 7, 8, 9$, while $icg(i)$ $(1 \leq i \leq 9)$ are $3, 4, 5, 6, 7, 8, 9, 10, 11$. Therefore, $WAP = 5.96/10 = 0.596$, which is much higher than that of the first case even though this second case *missed* the S-relevant document.

We now propose a modification of WAP. Let l denote a relevance level, where $l = 0$ implies nonrelevance and higher l implies higher relevance. As with cumulative gain and WAP, we assume that the values of $gain_l$ $(l > 0)$ are given as constants, and that $gain_0 = 0$. Then, for each $l (> 0)$ and *for each topic*, we define the *adjusted gain* as follows:

$$gain'_l = gain_l - \frac{R_l}{R}(gain_l - gain_{l-1}) \qquad (15)$$

For example, in the aforementioned example, $gain'_S = 3 - 1 * (3 - 2)/10 = 2.9$ and $gain'_B = 1 - 9 * (1 - 0)/10 = 0.1$. Note that the above transformation preserves the order $gain'_l \geq gain'_{l-1}$ in order to guarantee the optimality of the "best-possible" ranked output.

Using adjusted gains instead of the raw ones, we can easily obtain adjusted versions of the gain, the cumulative gain and the ideal cumulative gain at rank i, denoted by $g'(i)$, $cg'(i)$ and $icg'(i)$, respectively. We will refer to the ratio $cg'(i)/icg'(i)$ simply as the *Gain Ratio* (GR). Then, clearly, *Average Gain Ratio* (AGR) can be used instead of WAP for performance evaluation with multiple relevance levels:

$$AGR = \frac{1}{R} \sum_{1 \leq i \leq j, g'(i) > 0} \frac{cg'(i)}{icg'(i)} \qquad (16)$$

In the aforementioned example, the AGR of the first ranked output is $(2.9/2.9)/10 = 0.1$ (i.e. equal to WAP), while that of the second one is $(0.1/2.9 + 0.2/3.0 + 0.3/3.1 + \ldots + 0.9/3.7)/10 = 0.132$ (which is much lower than WAP).

Finally, just as TREC R-precision is used besides Average Precision, *R-Gain Ratio* (R-GR), given by $cg'(R)/icg'(R)$, may be used as an alternative evaluation measure.

References

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

[2] 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).