BRIDJE over a Language Barrier: Cross-Language Information Access by Integrating Translation and Retrieval

Tetsuya Sakai **Makoto Koyama** **Masaru Suzuki** **Akira Kumano** **Toshihiko Manabe**
Toshiba Corporate R&D Center Knowledge Media Laboratory,
1 Komukai-Toshiba-cho, Saiwai-ku, Kawasaki 212-8582, JAPAN
tetsuya.sakai@toshiba.co.jp

Abstract

This paper describes two new features of the BRIDJE system for cross-language information access. The first feature is the *partial disambiguation* function of the Bi-directional Retriever, which can be used for search request translation in cross-language IR. Its advantage over a "black-box" machine translation approach is consistent across five test collections and across two language permutations: English-Japanese and Japanese-English. The second new feature is the Information Distiller, which performs interactive summarisation of retrieved documents based on Semantic Role Analysis. Our examples illustrate the usefulness of this feature, and our evaluation results show that the precision of Semantic Role Analysis is very high.

1 Introduction

Cross-Language Information Retrieval (CLIR) (Grefenstette, 1998) has received a lot of attention recently. TREC currently studies English-Arabic IR, CLEF studies CLIR across European languages, and NTCIR studies CLIR across Asian languages (Chen *et al.*, 2003; Kando, 2001). As with monolingual IR, CLIR evaluations usually rely on the use of static test collections: The system accepts a *source language* search request and outputs a ranked list of *target language* documents, and this list is evaluated using metrics such as Average Precision. However, CLIR solves only part of the Language Barrier Problem: if the user cannot express his information need in target language, then he probably cannot make much use of the retrieved documents written in the same language. (If the source and target languages are reasonably similar, then the user may find such plain CLIR useful. However, this is certainly not the case for pairs of disparate languages such as English and Japanese.) Thus, what deserves more attention is Cross-Language Information *Access* (CLIA), which subsumes CLIR *and* provides useful information to the user in source language (e.g. (Frederking *et al.*, 1997)).

This paper describes two new features of the BRIDJE (Bi-directional Retriever/Information Distiller for Japanese and English) system (Sakai *et al.*, 2002a; Sakai *et al.*, 2002b; Sakai *et al.*, 2003) which integrates machine translation (MT) with information retrieval (IR) to support both English-Japanese (E-J) and Japanese-English (J-E) CLIA. The first feature is the *partial disambiguation* function of the *Bi-directional Retriever* part of BRIDJE for enhancing retrieval performance in the traditional sense. While most of the traditional MT-based CLIR systems use MT as a "black box", partial disambiguation accesses the internal data structures of a commercial MT system for search request translation so that multiple translation candidates can be used as search terms. We present positive results that are consistent across five test collections (or six topic sets), and across two language permutations: English-Japanese and

Japanese-English. To our knowledge, BRIDJE is the first system that truly integrates MT with IR *and* performs well in terms of standard measures. The second new feature is the entire *Information Distiller* part of BRIDJE, which can provide generic or query-specific summaries of the retrieved documents, as well as their translations in source language. Based on *Semantic Role Analysis* (SRA) originally designed for enhancing retrieval performance (Sakai *et al.*, 2002a; Suzuki *et al.*, 2001), the Information Distiller extracts important text fragments from a retrieved document on the fly. Preliminary evaluations suggest that SRA can classify text fragments with very high precision, and that it is useful for efficient information access. We regard our Information Distiller feature as one step towards Cross-Language Question Answering.

BRIDJE is an enhanced, fully bilingual version of the KIDS Japanese retrieval system that recently achieved the highest performances in the English-Japanese and Japanese monolingual IR tasks at NTCIR-3 (Chen *et al.*, 2003; Sakai *et al.*, 2003).

The remainder of this paper is organised as follows: Section 2 compares some of the previous work on CLIR/CLIA with our present study. Section 3 provides an overview of the BRIDJE system. Section 4 describes an extensive set of retrieval experiments that compares partial disambiguation with the black-box MT approach. Section 5 provides examples to illustrate the advantages of the Information Distiller for efficient information access, as well as some evaluation results of SRA. Finally, Section 6 provides conclusions.

2 Previous Work

This section reviews some previous work on CLIR/CLIA, focussing primarily on those that deal with English and Japanese or those that use MT in some way or other.

The early J-E CLIR systems (Susaki *et al.*, 1996; Yamabana *et al.*, 1998) employed *dictionary-based* search request translation, with *corpus-based* disambiguation. However, it is not clear how effective these systems are in terms of retrieval performance. In contrast, for both J-E and E-J CLIR, *MT-based* search request translation has been combined successfully with *Pseudo-Relevance Feedback* (Sakai *et al.*, 1999a; Sakai, 2001). The recent CLIR results at NTCIR-3 also showed that this approach, which BRIDJE also employs, is very promising (Chen *et al.*, 2003; Sakai *et al.*, 2003).

In our *partial disambiguation* experiments, we go beyond the use of MT as a black box by accessing the internal data structures of a commercial MT system in order to use multiple translation candidates as search terms. Jones *et al.* (Jones *et al.*, 1999) have explored this approach to some extent, but they observed performance degradation when compared to *full disambiguation* (i.e. black-box MT), as they treated the multiple candidates as *distinct* search terms. In contrast, BRIDJE treats these candidates as a group of *synonyms*, which is now known to be effective in CLIR (Pirkola, 1998). Thus, while dictionary-based approaches start from the maximum ambiguity state and performs vigorous disambiguation, we start from the minimum ambiguity state reached through full MT and take a step backward for obtaining alternative translations. The SYSTRAN NLP browser (Gachot *et al.*, 1998) also integrates MT with IR at a deep level, but this was designed for retrieving *sentences* that match specific grammatical features, and its effectiveness as a *document* retrieval system is not clear.

Regarding CLIA (as opposed to CLIR), Frederking *et al.* (Frederking *et al.*, 1997) proposed a framework in which retrieved documents could be summarised and translated by multiple MT engines in parallel. MULINEX (Capstick *et al.*, 2000), which is a dictionary-based CLIR system for French, English and German, uses the LOGOS MT system for document/summary presentation. Oard and Rensik (Oard and Resnik, 1999) have used *word-by-word* J-E translations in their study on interactive document selection. PRIME (Higuchi *et al.*, 2001), yet another J-E/E-J CLIR system based on dictionaries and corpora, translates retrieved patent documents *phrase-by-phrase*. Compared to these CLIA systems, the Information Distiller part of BRIDJE is unique in that it performs SRA for interactive document summarisation/presentation.

3 The BRIDJE System

This section describes the general features of the BRIDJE system. Sections 3.1 and 3.2 describe the request translation and indexing/retrieval components of the Bi-directional Retriever. Section 3.3 introduces the Information Distiller that provides summaries and translations of retrieved documents.

3.1 Request Translation

The default search request translation strategy of BRIDJE is *full disambiguation*: a source language request is fed to a commercial MT system (Amano *et al.*, 1989), and the output is treated as a monolingual request written in target language. BRIDJE is also capable of performing *transliteration* for treating words that are outside of the MT dictionary (Sakai *et al.*, 2002b), but this feature is not used in the present study.

In order to describe *partial disambiguation*, we first illustrate the disambiguation process of MT with an example. Suppose that a search request contains the word "play" in the context of E-J CLIR. Firstly, by *dictionary lookup*, we recognise that "play" can either be a noun or a verb, and obtain all possible Japanese translations accordingly. Secondly, we determine its part-of-speech through *syntactic analysis*. Here, suppose that "play" was used as a verb, and that all Japanese translations for the noun "play" have been filtered out. Thirdly, we perform *semantic analysis* using *transfer rules*. These rules contain knowledge such as "IF the object of the verb "play" is a *sport*, THEN it should be translated as "*suru*". IF the object is a *musical instrument*, THEN it should be translated as "*hiku*" or "*ensō-suru*". Here, suppose that the object was "violin", and therefore that "*hiku*" and "*ensō-suru*" have been selected as translation candidates. Then, at the final output stage, only one translation is selected for each source language word. For the word "play" in the above example, "*hiku*" is selected as the final translation since this is the first entry in the aforementioned transfer rule. (The above explanation is a simplified version of what really goes on inside our MT system.)

Partial disambiguation takes all candidate translations that are left just after the *semantic analysis* stage ("*hiku*" and "*ensō-suru*" in the above example), and treats them as a set of *synonyms* in retrieval. As it is well known, disambiguation in MT is far from perfect, and the synonym groups thus produced often contain some inappropriate terms. Despite this, our experiments described in Section 4 show that partial disambiguation is effective.

3.2 Indexing and Retrieval

Our default retrieval strategy is *Okapi/BM25* with *Pseudo-Relevance Feedback* (PRF) (Robertson and Sparck Jones, 1997; Sakai, 2001). The term selection criterion used in all of our experiments involving PRF is $ow4$, which incorporates the initial document scores into the traditional offer weight (Sakai *et al.*, 2003).

BRIDJE employs word-based indexing, prior to which synonyms and phrases can be defined. Synonym groups can also be defined at query time, which is useful for partial disambiguation and transliteration (Sakai *et al.*, 2002b): the term frequency (tf) and the document frequency (df) are counted as if all the members of a synonym group are one identical term.

Optionally, BRIDJE can perform indexing and retrieval based on SRA (Sakai *et al.*, 2002a; Suzuki *et al.*, 2001), which was originally designed for going beyond the "bag-of-words" approach to IR. Although the present study uses SRA for interactive document summarisation/presentation and not for retrieval, it works as follows: During document indexing and search request processing, BRIDJE can examine the output of the *parser* (morphological analyser for Japanese; part-of-speech tagger/stemmer for English), based on a set of hand-written *SRA rules* which specify the following:

- How to break up the text into *fragments*, based on regular expression pattern matching. A fragment can be a sentence, a paragraph, a prepositional phrase, and so on. For example, an English sentence of the form "A for B with C" can be broken into "A", "for B" and "with C" if prepositions are used as fragment boundaries.
- How to assign *Semantic Roles* (SRs) to each fragment, again based on regular expression pattern matching. Using prepositions as *SRA triggers* in the above example, it is possible to tag the fragment "for B" with "`PURPOSE`", the fragment "with C" with "`MEANS`", and the fragment "A" with "`UNDETERMINED`" (which means that no SRA rule matched).
- The *SR correlation weights* for document score

calculation in retrieval (*not* used for document presentation). For example, if a term occurs in a PURPOSE context in the request *and* occurs in a PURPOSE context in the document as well, its BM25 term weight can be upweighted.

The above simple example of using PURPOSE and MEANS as SRs has been shown to be effective for retrieving highly relevant documents from the NTCIR-2 English test collection (Sakai *et al.*, 2002a). However, we have also found that it is difficult to devise SRA rules that work across different relevance levels or across different document types. This lead us to explore an alternative way of utilising SRA, i.e. for document summarisation/presentation.

3.3 Information Distiller

The Information Distiller can interactively generate generic or query-specific summaries of a retrieved document by extracting fragments (usually sentences) that meet specified criteria. The unique feature of BRIDJE as a CLIA system is that it supports interactive selection of fragments based on SRA: For example, the user can select fragments whose SRs are TOPIC/AIM, BACKGROUND, RESULT/CONCLUSION, OPINION, and so on. In addition, the Information Distiller can generate *lead* and tf-idf-based extracts, and these criteria can be combined with SRA requirements. The summaries thus produced (or the original document text) can be translated back into source language using MT.

4 Evaluation of Partial Disambiguation

This section compares *partial disambiguation* with traditional *full disambiguation* for search request translation in CLIR.

4.1 Experimental Setting

We used three *English-Japanese* test collections: NTCIR-3 E-J (Chen *et al.*, 2003), NTCIR-2 E-J (Kando, 2001) and BMIR-J2 E-J (Sakai *et al.*, 1999a; Sakai *et al.*, 1999b) [1]. As BMIR-J2 E-J has two English topic sets X and Y translated from a single Japanese topic set (Sakai *et al.*, 1999a), we performed four sets of E-J CLIR experiments in total. On the other hand, we used two *Japanese-English* test collections: NTCIR-3 J-E (Chen *et al.*, 2003) and NTCIR-2 J-E (Kando, 2001). Table 1 summarises the features of these five test collections. As they provide multiple relevance levels, the number of relevant documents summed across topics are shown for each relevance level: S-relevant (highly relevant), A-relevant (relevant), and B-relevant (partially relevant). There are no S-relevant documents for BMIR-J2.

Following the practice at NTCIR, we computed Mean Average Precision (MAP) based on "relaxed" relevance (which treats S,A and B-relevant documents as relevant) and on "rigid" relevance (which treats S and A-relevant documents as relevant) (Chen *et al.*, 2003; Kando, 2001). In addition, for unified evaluation with multiple relevance levels, we used *Average Gain Ratio* (AGR) (Sakai *et al.*, 2003; Sakai, 2003). For testing statistical significance, we used the sign test.

For each test collection, *full disambiguation* and *partial disambiguation* queries were generated using the topic *descriptions*. Although our MT system has several domain-specific dictionaries, we used the general dictionary only. For the NTCIR-3 E-J experiment, the BM25 and PRF parameters were tuned using the dryrun topics (Sakai *et al.*, 2003). For all other experiments, Okapi/BM25 defaults were used, and the number of pseudo-relevant documents and that of expansion terms were fixed to 10 and 30, respectively (Sakai, 2001).

4.2 Results

Table 2 summarises the results of our CLIR experiments. Full disambiguation runs with and without PRF are denoted by FD+PRF and FD, while the corresponding partial disambiguation runs are denoted by PD+PRF and PD, respectively. The *monolingual* performances with and without PRF, denoted by ML+PRF and ML, are also shown. Columns (i), (ii) and (iii) show performances in MAP with relaxed relevance, MAP with rigid relevance, and Mean AGR (MAGR), respectively. Runs that significantly outperform FD are indicated by "$*$"s, while those that significantly outperform PD are indicated by "†"s. For example, Table 2**A** Column (i) include the following information in terms of relaxed MAP:

[1] Data in BMIR-J2 is taken from the Mainichi Shimbun CD-ROM 1994 data collection. BMIR-J2 was constructed by the SIG Database Systems of the Information Processing Society of Japan, in collaboration with the Real World Computing Partnership.

Table 1: Test Collections

name	#topics	#S/A/B-rel docs	#docs	document type
English-Japanese test collections				
NTCIR-3 E-J	42	330/1324/884	236,664	Mainichi newspaper 1998-1999
NTCIR-2 E-J	49	465/2815/1813	736,158	conference paper abstracts + grant-in-aid research report abstracts
BMIR-J2 E-J	50(X) 50(Y)	0/624/1057	5,080	Part of Mainichi newspaper 1994
Japanese-English test collections				
NTCIR-3 J-E	32	116/328/297	22,927	Mainichi Daily News 1998-1999 + Taiwan News / Chinatimes English News 1998-1999
NTCIR-2 J-E	49	214/1196/726	322,058	conference paper abstracts + grant-in-aid research report abstracts

(a) PD+PRF is 8% better than FD+PRF, and significantly better than FD ($\alpha = 0.01$) and PD ($\alpha = 0.05$); (b) FD+PRF is *not* significantly better than FD and PD (hence the lack of "$*$"s and "†"s); and (c) PD is 6% better than FD, but this difference is not statistically significant (hence the lack of "$*$"s).

The following are general observations made from Table 2:

- PD outperforms FD for all test collections in terms of all three evaluation measures. The improvements are 1-11%. The differences are statistically significant in Table 2**C** Columns (i) and (iii), for Topic Set Y ($\alpha = 0.05$).

- PD+PRF outperforms FD+PRF for all test collections in terms of all three evaluation measures. The improvements are 1-8%. Although these differences are not statistically significant by direct comparison, the "$*$"s and "†"s, which indicate superiority over FD and PD, suggest that PD+PRF is superior to FD+PRF in general. (Table 2**B** is an exception: here, there is no statistical evidence which suggests that PD+PRF is superior to FD+PRF. However, note that PD+PRF outperforms FD+PRF on average even for this test collection.)

- In general, PRF preserves the positive effect of partial disambiguation. For example, Table 2**A** Column (ii) shows that PD is 8% better than FD, and that PD+PRF is also 8% better than FD+PRF. (Table 2**C** is an exception: For example, in Column (ii), PD+PRF(Y) is only 2% better than FD+PRF(Y) even though PD(Y) is 11% better than FD(Y).)

Thus, the small advantage of partial disambiguation over full disambiguation is consistent across the five test collections (six topic sets) and across the two language permutations.

Table 3(a) compares FD and PD for each test collection in terms of average number of query terms per topic. For the *E-J* topics, the PD queries are approximately three times longer than the FD ones. Compared to this, the FD and PD queries for the *J-E* topics are relatively similar in length. That is, semantic analysis in J-E MT generally yields fewer translation candidates than that in E-J MT does. However, the relationship between the number of alternative translations and the success of partial disambiguation is not straightforward: While the E-J results are more successful than the J-E results for NTCIR-3 (Table 2**A** vs **D**), suggesting that "adding more terms is better", this is not true for NTCIR-2 (Table 2**B** vs **E**). This is probably because the quality of the alternative translations vary widely, as we shall see later.

Table 3(b) shows the total number of out-of-vocabulary words for each topic set. Neither FD nor PD could translate these words as our general MT dictionary was not tuned in any way for our experiments. The NTCIR-3 E-J topic set contained three out-of-vocabulary words, including "Tomiich" which is a misspelling of "Tomiichi (Murayama, a former Japanese prime minister)", while the NTCIR-3 J-E topic set contains five, including "*konpyutā*" which is a misspelling of "*konpyūtā*" or "*konpyūta*". Meanwhile, almost all of the out-of-vocabulary words in the NTCIR-2 E-J and J-E topic sets were technical terms. However, there was no topic that contained more than one out-of-vocabulary

Table 2: Partial disambiguation vs full disambiguation.

	(i) relaxed MAP	(ii) rigid MAP	(iii) MAGR
A. NTCIR-3 E-J			
ML+PRF	0.4308	0.3715	0.6130
ML	0.3953	0.3368	0.5854
PD+PRF	**0.3846**(+8%) $**\dagger$	**0.3365**(+8%) $**\dagger\dagger$	**0.5835**(+5%) $**\dagger\dagger$
FD+PRF	0.3575	0.3121	0.5561 $**$
PD	**0.3351**(+6%)	**0.2874**(+8%)	**0.5400**(+3%)
FD	0.3158	0.2672	0.5250
B. NTCIR-2 E-J			
ML+PRF	0.2903	0.3039	0.4076
ML	0.2462	0.2650	0.3478
PD+PRF	**0.2461**(+3%) $*\dagger$	**0.2769**(+3%) $**\dagger$	**0.3562**(+1%) $*\dagger\dagger$
FD+PRF	0.2391 $**\dagger\dagger$	0.2691 $**\dagger\dagger$	0.3536 $**\dagger\dagger$
PD	**0.1898**(+3%)	**0.2229**(+3%)	**0.2839**(+1%)
FD	0.1845	0.2157	0.2810
C. BMIR-J2 E-J			
ML+PRF	0.4653	0.4135	0.6736
ML	0.4345	0.3792	0.5485
PD+PRF(X)	**0.3816**(+4%) $**\dagger\dagger$	**0.3532**(+3%) $**\dagger\dagger$	**0.5870**(+2%) $*\dagger\dagger$
FD+PRF(X)	0.3658 $**$	0.3434 $**\dagger\dagger$	0.5751
PD(X)	**0.3380**(+6%)	**0.3009**(+2%)	**0.4192**(+6%)
FD(X)	0.3196	0.2936	0.3939
PD+PRF(Y)	**0.3522**(+6%) $**\dagger\dagger$	**0.2949**(+2%) $**$	**0.5660**(+4%) $**$
FD+PRF(Y)	0.3333 $**$	0.2879 $**$	0.5423 $**$
PD(Y)	**0.2965**(+7%) $*$	**0.2538**(+11%)	**0.4307**(+11%) $*$
FD(Y)	0.2772	0.2291	0.3888
D. NTCIR-3 J-E			
ML+PRF	0.4620	0.4141	0.6698
ML	0.4237	0.3809	0.6322
PD+PRF	**0.4103**(+3%) $*$	**0.3735**(+3%) $*$	**0.6112**(+1%) $*\dagger$
FD+PRF	0.3973	0.3617	0.6038 $*\dagger$
PD	**0.3676**(+3%)	**0.3396**(+2%)	**0.5603**(+2%)
FD	0.3584	0.3329	0.5497
E. NTCIR-2 J-E			
ML+PRF	0.2644	0.3075	0.4496
ML	0.2279	0.2729	0.3811
PD+PRF	**0.2202**(+4%) $**\dagger\dagger$	**0.2440**(+4%) $**\dagger\dagger$	**0.3984**(+3%) $**\dagger\dagger$
FD+PRF	0.2112 $**$	0.2344 $*\dagger$	0.3861 $**\dagger\dagger$
PD	**0.1870**(+5%)	**0.2212**(+4%)	**0.3352**(+1%)
FD	0.1780	0.2120	0.3297

Runs that significantly outperform FD are indicated by "$*$" ($\alpha = 0.05$) and "$**$" ($\alpha = 0.01$).
Those that significantly outperform PD are indicated by "$\dagger$" ($\alpha = 0.05$) and "$\dagger\dagger$" ($\alpha = 0.01$).
The percentages in the PD+PRF rows represent the gain over FD+PRF, while those in the PD rows represent the gain over FD.

Table 3: (a)#terms per topic / (b)#out-of-vocabulary words.

	(a)		(b)
	FD	PD	
NTCIR-3 E-J	8.6	26.9	3
NTCIR-2 E-J	6.5	18.1	4
BMIR-J2 E-J (X)	3.1	11.9	1
BMIR-J2 E-J (Y)	3.5	12.3	0
NTCIR-3 J-E	11.0	20.1	5
NTCIR-2 J-E	13.5	16.0	8

word (with one exception), and it appears that out-of-vocabulary words did not directly affect our experiments: partial disambiguation managed to improve five of the eight NTCIR-2 J-E topics that contained an out-of-vocabulary word.

Given that the effect of out-of-vocabulary words is negligible, one may hypothesize that the advantage of partial disambiguation over full disambiguation may be smaller with technical papers than with newspapers, as technical papers contain more technical terms and full disambiguation may be sufficient for translating them. By comparing the *E-J* results (Table 2**A-C**), it can be observed that partial disambiguation was indeed a little less successful for NTCIR-2 E-J (technical papers) than for NTCIR-3 E-J and BMIR-J2 E-J (newspapers). However, as our *J-E* results (Table 2**D-E**) show comparable performances for both document types, the above hypothesis is not fully supported.

4.3 Per-topic Analyses

While partial disambiguation improves retrieval performance on average for all test collections, per-topic analyses show that it hurts performance for some topics, and that there is room for improvement. Table 4 provides some per-topic comparisons of the Average Precision values of FD and PD. Two examples are given for NTCIR-3 E-J and for J-E, respectively, where the Japanese *descriptions* for the latter are shown here in English. Words that are mentioned in the discussion below are underlined.

As Table 4 shows, PD was hugely successful for E-J Topic 017: The only word that FD translated correctly was "Kitano": "Director" was mistranslated as "*kanrisha*" (manager), "Takeshi" was transliterated into *katakana* (which is not appropriate in this case), and "films" was transliterated into "*firumu*" (which means "camera films", not "movies"). In contrast, PD successfully recaptured the correct translations "*kantoku*" (director) and "*eiga*" (films). However, some inappropriate translations such as "*shikisha*" (orchestra director) and "*torishimariyaku*" (managing director) were added as well, as semantic analysis is not perfect. Moreover, PD did not help in translating "Takeshi": although it obtained two *kanji* spellings for it, they were incorrect for this particular Takeshi Kitano. (There are more than 40 possible *kanji* spellings for "Takeshi"!) Nevertheless, the use of synonym operators seems to have absorbed the negative effect of such inappropriate translations for this topic. On the other hand, PD was *not* successful for E-J Topic 004: while FD obtained the correct translations "*denshishōtorihiki*" (E-commerce) and "*naiyō*" (contents), PD added the *acronym* of "E-commerce", which happened to hit many nonrelevant documents that mention "European Community". (The roman alphabet is often used in Japanese texts for representing foreign acronyms.) Moreover, PD added inappropriate translations for "contents" such as "*mokuji*" (table of contents) and "*yōryō*" (capacity).

PD was also successful for J-E Topic 031: FD obtained "optimal" instead of "best", and "place" instead of "spot". In contrast, PD successfully recaptured "best" and "spot", even though it also added some possibly harmful terms such as "position" and "space". On the other hand, PD was *not* successful for J-E Topic 050: FD obtained "dress" instead of "clothing", to which PD added "appearance", "clothes", "costume" and "garment". Meanwhile, FD obtained "hairstyle" instead of "hair styles", to which PD added "hairdo" and "coiffure". FD obtained "makeup" instead of "cosmetics", to which PD added "dressing" and "toilet".

As mentioned earlier, while semantic analysis in J-E MT generally yields fewer translations than that in E-J MT does, this difference is not clearly reflected in terms of retrieval performance. One possible cause of this is that the partial disambiguation terms in the J-E case are more polysemous, e.g. "space" and "toilet", though fewer in number.

The above examples suggest that partial disambiguation may be improved by adopting a more *selective* strategy. Although we have conducted additional experiments by limiting the number of terms added by partial disambiguation, this did not improve performance as the candidate terms obtained after semantic analysis have no priority information in our MT system. One possible solution to this is to utilise the corpus statistics such as the document frequency so that polysemous words can be filtered out, but this is beyond the scope of this paper.

Finally, by looking across the columns of Table 2, it can be observed that the results in terms of MAGR are generally consistent with those in terms of relaxed/rigid MAP. This suggests that MAGR is a good

Table 4: Per-topic comparison of Average Precision: FD vs.PD.

NTCIR-3 E-J				
TopicID	DESCRIPTION	FD		PD
017	Articles relating [related] to Director Takeshi Kitano's films.	0.038	<	0.278
004	Find [out] what E-Commerce is and its contents.	0.368	>	0.283

NTCIR-3 J-E				
TopicID	Official English translation of DESCRIPTION	FD		PD
031	Where are the best spots in Kyoto for viewing of Japanese maples in their fall color?	0.504	<	0.607
050	To retrieve documents describing teenagers' fashion trends in clothing, hair styles, cosmetics.	0.330	>	0.306

Japan-Nepal Health Scientific Expedition -Comparative **Epidemiological** Studies on the Genesis of Hypertension- (UNDETERMINED)

S1: It is generally thought that the increase in **blood pressure** with age may be avoided by the extremely low sodium intake. (BACKGROUND)

S2: we, however, have noticed that **blood pressure** hardly increased with age in some communities briefly studied in Nepal since 1978, even though the inhabitants take salty foods and beverages.(TOPIC/AIM)

S3: Our purpose was to ascertain this and to clarify the factor(s) influencing no increase in **blood pressure** with age in terms of extensive **epidemiological** standpoint. (TOPIC/AIM)

S9: RESULTS : **blood pressure** for both sexes was statistically significantly **higher** in villagers in Bhadrakali than in Kotyang. (RESULT/CONCLUSION)

S16: CONCLUSION : In spite of consuming more than 10g per day of **salt** in both Kotyang and Bhadrakali, the **blood pressure** hardly increased with age only in the former, suggesting that the **blood pressure** may be influenced by physical activity, fat free mass and nutrient consumption rather than **salt** intake in these villages in Nepal. (RESULT/CONCLUSION)

Figure 1: A summary for NTCIR-2 J-E Topic 0103 (official English translation: "Correlations between the onset of hypertension and diet, such as salt intake, based on epidemiological surveys in countries other than Japan")

substitute for MAP in evaluations using multiple relevance levels.

5 Information Distiller

This section provides some example summaries and translations to illustrate the usefulness of the Information Ditiller for efficient information access, as well as some preliminary evaluation results of SRA.

5.1 Example Summaries and Translations

As this paper is intended primarily for the English speaking community, we provide one example summary in the context of J-E CLIR *before* they are translated by MT into Japanese, and two in the context of E-J CLIR *after* they have been translated by MT into English. The summaries shown here are all query-specific, and are based on full disambiguation queries.

Figure 1 shows a sample summary of an English technical paper abstract that is S-relevant to NTCIR-2 J-E Topic 0132, which is about "correlations between hypertension and diet." Of the 16 sentences in the original document (excluding the title, shown at the top of this summary), those which do not contain any of the English query terms and those whose SRs were UNDETERMINED have been filtered out, leaving only five sentences. The words that matched the query terms are shown in boldface, and the English expressions which acted as SRA triggers are underlined. For example, Sentence 1 (S1) was tagged with BACKGROUND because the string "It is generally thought" matched a regular expression in one of the SRA rules. If the user desires an *indicative* summary, BRIDJE can present S2 and S3, which are tagged with TOPIC/AIM. Subsequently, if he desires an *informative* summary, BRIDJE can present S9 and S16, which are tagged with RESULT/CONCLUSION. Of course, the user can specify multiple SRs, choose to read UNDETERMINED sentences, or combine such usage with *tf-idf*-based sentence filtering. Real-time response is easy because SRA for the documents are done at index time. Finally, the summary can be translated into Japanese by MT for the Japanese user.

In the above example, only one SR was assigned

The Cannes International **Film** Festival is challenged shortly. – **Director** **Takeshi Kitano** and "chrysanthemum Jiro's summer" are sent. (TITLE)

S1: The [Paris 22-day cooperation] **Director** **Takeshi Kitano** who won Golden Lion (Grand Prix) at the Venice **Film** Festival in 1997 will send new work "chrysanthemum Jiro's summer" into a world's largest **film** festival and the competition section of the Cannes International **Film** Festival. (TOPIC,DATE,TITLE)

S4: It is **Director** **Kitano** 8 Motome's work in "the summer of chrysanthemum Jiro." (TITLE)

Figure 2: A translated summary for NTCIR-3 E-J Topic 017: "Articles relating [related] to Director Takeshi Kitano's films."

The **Emperor** and Empress' ◯ ◯ ◯ is decided. (UNDETERMINED)

S1: The schedule on which the **Emperor** and Empress visit Britain and **Denmark** as a guest of the nation was reported to the cabinet meeting on the 17th, and outlines, such as a welcome event, solidified. (DATE)

S3: Periods are 13 nights and 14 days of May 23 start and June 5 homecoming. (DATE)

S4: After dropping in at Portugal, it will arrive in Britain for 25 days, and from the next day, starting with Queen Elizabeth's welcome ceremony, a **Japanese** company besides a start, 3 times of dinner meetings, and 2 times of luncheons visits Wales to which it has advanced mostly by day's trip, or a formal event has a friendly talk [scientists / of a royal association]. (DATE)

S5: Arriving in **Denmark** is on the afternoon of the 31st. (DATE)

S6: Although there is no formal event, events, such as a welcome ceremony, start the royal palace of lodgings on a visit and following the **2nd**, and Queen Margrethe inspects the Copenhagen university, the National Museums, a welfare institution for the aged, etc., and she will go back Denmark earnestly on the **afternoon of the 5th** on the night of the **4th** on the next day. (DATE)

Figure 3: A translated summary for NTCIR-3 E-J Topic 030: "When, if ever, has the Japanese Emperor been to Denmark?"

to each fragment (i.e. sentence) for simplicity. However, SRs can be used as orthogonal features, as shown in the next example.

Figure 2 shows a sample *translated* summary (i.e. an MT output) of a Japanese newspaper article that is S-relevant to NTCIR-3 E-J Topic 017: "Articles relating [related] to Director Takeshi Kitano's films." Words that match those from the source language request are indicated in boldface, and words that correspond to the Japanese SRA triggers are underlined. As the search request ends with the word "films", it is possible for BRIDJE to guess that movie titles may be useful to the user. Thus, BRIDJE can show S1 and S4 to the user by default, as they have been tagged with TITLE based on SRA triggers such as "*kantoku*" (director), "*sakuhin*" (work) and Japanese brackets. Note also that S1 has two more SRs, TOPIC and DATE. Unfortunately, the correct English translation of the movie title is "Kikujiro's summer": *Kikujiro* is a very unusual Japanese name, while *Jiro* is a common first name. Besides, *kiku* does mean chrysanthemum the flower! Nevertheless, we view this example as one step towards cross-language question answering, which deals with questions such as "List up Takeshi Kitano's Japanese films - give me rough translations in English."

Similarly, Figure 3 shows a *translated* summary of a Japanese newspaper article that is S-relevant to NTCIR-3 E-J Topic 030: "When, if ever, has the Japanese Emperor been to Denmark?". By performing SRA on this "when" question, it is possible for BRIDJE to guess that the user is looking for DATE-type information, as in question answering. In such a case, BRIDJE can present a list of sentences containing dates as shown. Even though the translations are far from perfect, the English speaking user can probably guess, by reading S3, S5, and S6 that the answer to the question is "from May 31st to June 4th or 5th". Combining this with the *meta-data* of this document, a sophisticated cross-language question answering system would output the year "1998" as well. Unfortunately, the document title in Figure 3 contains a kanji word which MT failed to translate: "*hōōbi*" (dates for visiting Europe), shown as "◯ ◯ ◯" in this paper to avoid Japanese fonts.

The above three examples illustrate the usefulness of SRA for efficient access to the desired information within an English or a Japanese document, and for allowing different *views* for summarising a document. Moreover, we have argued that current MT technology can be useful for CLIA despite its limited quality. By performing SRA-based sentence filtering, the amount of MT output that the user has to go through can be kept to a minimum.

Table 5: SRA precision for NTCIR-2.

Semantic Role	#fragments	Precision
A. NTCIR-2 English documents		
TOPIC/AIM	72 (12%)	100%
RESULT/ CONCLUSION	58 (10%)	98% (57/58)
BACKGROUND	8 (1%)	100%
OPINION	0 (0%)	-
SubTotal	**138 (23%)**	**99% (137/138)**
UNDETERMINED	454 (77%)	-
Total	592 (100%)	-
B. NTCIR-2 Japanese documents		
TOPIC/AIM	82 (15%)	96% (79/82)
RESULT/ CONCLUSION	42 (8%)	100%
BACKGROUND	27 (5%)	93% (25/27)
OPINION	7 (1%)	100%
SubTotal	**158 (31%)**	**97% (153/158)**
UNDETERMINED	355 (69%)	-
Total	513 (100%)	-

Table 6: SRA precision for NTCIR-3.

Semantic Role	#fragments	Precision
C. NTCIR-3 English documents		
TOPIC	157 (6%)	98% (154/157)
OPINION	24 (1%)	100%
MONEY	43 (2%)	100%
YEAR	62 (2%)	100%
PERCENTAGE	29 (1%)	100%
SubTotal	**315 (13%)**	**99% (312/315)**
UNDETERMINED	2201 (87%)	-
Total	2516 (100%)	-
D. NTCIR-3 Japanese documents		
TOPIC	123 (6%)	98% (121/123)
COMMENT	171 (8%)	89% (153/171)
TITLE	5 (0%)	100%
DATE	81 (4%)	98% (79/81)
MONEY	27 (1%)	100%
PERCENTAGE	22 (1%)	100%
SubTotal of unique fragments	**401 (18%)**	**95% (379/401)**
UNDETERMINED	1772 (82%)	-
Total	2173 (100%)	-

5.2 Precision of SRA

As the Information Distiller is an interactive subsystem, user-oriented, overall usefulness evaluations should be performed. As a first step, however, we evaluate the *precision* of the SRA rule sets that were actually used to generate the aforementioned examples. The results reported here are only indicative as the rule sets are experimental versions.

By using S-relevant documents as training data, one SRA rule set was devised for each of the four document collections: NTCIR-2 English/Japanese (technical papers) and NTCIR-3 English/Japanese (newspapers). Then, for each collection, we prepared a test set of documents A' as follows: Let S and A be the set of unique S-relevant and A-relevant documents, respectively. Take 100 A-relevant documents at random from the set $A - S$ and let this set be A'. This ensures that the test sets consist of unknown documents only, even if a document that is A-relevant for a certain topic is S-relevant for another topic.

SRA was performed on all fragments (i.e. sentences) from A' for each collection. Then, the first author examined each fragment and judged whether the assigned SR was "acceptable" or not. Here, "unacceptable" SRs are those that would probably mislead the user: For example, if a fragment from a technical paper is tagged with RESULT/CONCLUSION even though the fragment in fact provides a BACKGROUND information, this may cause a misunderstanding and is therefore unacceptable. Note also that this evaluation concerns *generic* summary fragments, as they subsume *query-focused* ones.

We define *SRA precision* as the number of fragments with acceptable SRs divided by the total number of fragments. We do not consider its *recall* counterpart here, because the Information Distiller is supposed to filter out many sentences and present "typical" sentences only. Although our current SRA rule sets assign SRs to only a small fraction of given fragments, the user can choose to read UNDETERMINED sentences or select multiple SRs at any time, as discussed earlier. Thus, the aim of this lenient evaluation is to ensure that the output of Information Distiller will "look okay" to the user.

Tables 5 and 6 summarise our SRA precision results. Table 5**A** corresponds to the NTCIR-2 *English* SRA rule set (which was used to generate the summary in Figure 1), and 5**B** corresponds to the

Table 7: Unacceptable vs desirable SRs.

Assigned SR	Desirable SR	#frags
A. NTCIR-2 English docs		
RESULT/ CONCLUSION	BACKGROUND	1
B. NTCIR-2 Japanese docs		
TOPIC/AIM	BACKGROUND	2
TOPIC/AIM	RESULT/ CONCLUSION	1
BACKGROUND	TOPIC/AIM	1
BACKGROUND	RESULT/ CONCLUSION	1
C. NTCIR-3 English docs		
TOPIC	OPINION	2
TOPIC	MONEY	1
D. NTCIR-3 Japanese docs		
TOPIC	UNDETERMINED	2
COMMENT	UNDETERMINED	13
COMMENT	TITLE	5
DATE	UNDETERMINED	2

NTCIR-2 *Japanese* SRA rule set. Table 6**C** corresponds to the NTCIR-3 *English* SRA rule set, and 6**D** corresponds to the NTCIR-3 *Japanese* SRA rule set (which was used to generate the original Japanese summaries for Figures 2 and 3). For example, Table 5**A** includes information such as: (a) Of the 592 fragments extracted from the test set A', only 138 (23%) were tagged with an SR; (b) Of the above 138 fragments, 58 were tagged with RESULT/CONCLUSION, but one of them was judged as unacceptable. Hence the Precision for this SR is 57/58=98%; (c) As the abovementioned fragment was the only unacceptable one, the overall precision is 137/138=99%.

The one unacceptable fragment tagged with RESULT/CONCLUSION was: "*As Kipps's recognition algorithm does not give us a way to extract any parsing result, his algorithm is not considered as a practical parsing algorithm.*" which accidentally matched an SRA rule that included "result" as a trigger. As shown in Table 7**A**, this fragment should probably be tagged with BACKGROUND, since it discusses previous work rather than the author's present work.

Similarly, Tables 5**B** and 7**B** show the results for the NTCIR-2 Japanese SRA rule set which is very similar to its English counterpart. As there were five unacceptable cases, its overall precision is 97%.

Tables 6**C** and 7**C** show the results for the NTCIR-3 English SRA rule set. Three fragments tagged with TOPIC were judged as unacceptable, one of which was: "*But I've always said that I won't compromise when it comes to demanding that the facts surrounding the incident come out in the open.*" This fragment accidentally matched an SRA trigger "said that", designed to match fragments such as "The prime minister said that. . . "

Tables 6**D** and 7**D** show the results for the NTCIR-3 Japanese SRA rule set. Unlike the other SRA rule sets, the SRs in this rule set were defined as orthogonal features, which allowed multiple SRs per fragment as in Figure 2. (Hence the **SubTotal** row provides information on *distinct* fragments.) As many as 18 fragments were incorrectly tagged with COMMENT, due to our reliance on Japanese brackets as SRA triggers: words surrounded by brackets are often technical terms or movie/book titles, rather than quoted comments. Moreover, two fragments were incorrectly tagged with DATE as they contained the word "*ichinichi*" (one day), whose spelling is the same as "*tsuitachi*" (first [of January]).

To summarise our preliminary results: Our experimental SRA rule sets can assign SRs to 13-31% of completely unknown English/Japanese sentences (but from known document genres), with precision of 95-99%. Moreover, SRA precision would be even higher for query-focused summaries. Thus, the SRs presented by the Information Distiller would probably look satisfactory to the user.

6 Conclusions

This paper introduced two new features of the BRIDJE system, namely, partial disambiguation for effective CLIR and document summarisation/presentation based on Semantic Role Analysis. We showed that the advantage of partial disambiguation over full disambiguation is consistent across five test collections, with four English-Japanese and two Japanese-English topic sets. As for document presentation using the Information Distiller, we have provided examples as well as preliminary evaluations to show that it can be useful for efficient and interactive cross-language information access. Topics of our future work include:

- Improving partial disambiguation by being more selective;

- Extensive and user-oriented evaluations of the Information Distiller;
- User-oriented query expansion using the Information Distiller;
- Expanding our language scope, for example, to Chinese; and
- Building a true cross-language question answering system.

References

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

Capstick, J. *et al.*: A System for Supporting Cross-Lingual Information Retrieval, *Information Processing and Management*, Vol. 36, No. 2, pp. 275–289 (2000).

Chen, K.-H. *et al.*: Overview of CLIR Task at the Third NTCIR Workshop, *NTCIR-3 Proceedings* (2003).

Frederking, R. *et al.*: Translingual Information Access, *AAAI Spring Symposium Cross-Language Text and Speech Retrieval*.

Gachot, D. A., Lange, E. and Yang, J.: The SYSTRAN NLP Browser: An Application of Machine Translation Technology in Cross-Language Information Retrieval, In (Grefenstette, 1998).

Grefenstette, G. (ed.): *Cross-Language Information Retrieval*, Kluwer Academic Publishers (1998).

Higuchi, S. *et al.*: PRIME: A System for Multi-lingual Patent Retrieval, *MT Summit VIII*, pp. 163–167 (2001).

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

Kando, N.: Overview of Japanese and English Information Retrieval Tasks (JEIR) at the Second NTCIR Workshop, *NTCIR-2 Proceedings*, pp. 73–96 (2001).

Oard, D. W. and Resnik, P.: Support for Interactive Document Selection in Cross-Language Information Retrieval, *Information Processing and Management*, Vol. 35, No. 3, pp. 363–379 (1999).

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).

Robertson, S. E. and Sparck Jones, K: Simple, Proven Approaches to Text Retrieval, Computer Laboratory, University of Cambridge (1997). `http://www.ftp.cl.cam.ac.uk/ftp/papers/reports/#TR356`

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).

Sakai, T. *et al.*: BMIR-J2: A Test Collection for Evaluation of Japanese Information Retrieval Systems, *ACM SIGIR Forum*, Vol. 33, No. 1 (1999). `http://www.acm.org/sigir/forum/F99/tetsuya.sakai.pdf`

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

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

Sakai, T. *et al.*: Generating Transliteration Rules for Cross-Language Information Retrieval from Machine Translation Dictionaries, *IEEE SMC 2002* (2002).

Sakai, T. *et al.*: Toshiba KIDS at NTCIR-3, *NTCIR-3 Proceedings* (2003). `http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings3/NTCIR3-CLIR-SakaiT`

Sakai, T.: Average Gain Ratio: A Simple Retrieval Performance Measure for Evaluation with Multiple Relevance Levels, *ACM SIGIR 2003 Proceedings*, *to appear* (2003).

Susaki, S., Hayashi, Y. and Kikui, G: Navigation Interface in Cross-Lingual WWW Search Engine, TITAN, *AUUG '96 & Asia Pacific World Wide Web*, `http://www.csu.edu.au/special/auugwww96/proceedings/susaki/susaki.html` (1996).

Suzuki, M. *et al.*: Customer Support Operation with a Knowledge Sharing System KIDS: An Approach based on Information Extraction and Text Structurization, *IIIS SCI 2001 Proceedings*, pp. 89–94 (2001). World Multiconference on Systemics, Cybernetics and Informatics, International Institute of Informatics and Systemics

Yamabana, K. *et al.*: A Language Conversion Front-End for Cross-Language Information Retrieval, In (Grefenstette, 1998).