情 報 学 基 礎 51－7
自 然 言 語 処 理 127－7
（1998. 9. 17）

Cross-Language Information Access: a case study for English and Japanese

Gareth Jones Nigel Collier Tetsuya Sakai Kazuo Sumita Hideki Hirakawa
Communication and Information Systems Research Laboratories
Research and Development Center, Toshiba Corporation
1, Komukai Toshiba-cho, Saiwai-ku, Kawasaki 210-8582, Japan

Abstract

Internet search engines allow access to online information all over the world. The current assumption, that users are fluent in the language of the query and document collection, is clearly unreasonable given the growing number of languages used on the Internet. This paper shows how information retrieval and machine translation can be combined for cross-language information retrieval in an information-access frame-work. We present encouraging experimental results using English queries to retrieve Japanese documents from the BMIR-J2 test collection. A complete example demonstrates that technology already exists to provide effective and potentially useful cross-language information access applications.

1 Introduction

The quantity of textual material available online is currently increasing very rapidly. The most dramatic example of this is the recent explosion in use of the Internet and the World Wide Web. In principle users can download all online material to which they are permitted access. This is becoming increasingly possible with the advent of query translation facilities in the front-end of retrieval systems. However, a significant issue is that many users can currently only make use of information contained in documents written in languages in which they have some degree of fluency. This effectively blocks their access to information provided in other languages and limits their ability to exploit online information resources.

The importance of this issue is demonstrated by the ongoing evolution in people's approach to accessing information. Traditionally individuals requiring timely access to information in their work relied on material provided by various agencies. For example, current affairs material provided by international news agencies such as Reuters, and financial information from company and stock market reports. However, users increasingly no longer wait for information to arrive on their desk, but rather with the assistance of search engines, they can look for it online themselves. Thus workers are empowered to seek and make use of all available pertinent information and not just that provided by professional services.

In this paper we explore a complementary technology to enable content providers and consumers to make use of all available information sources. This paradigm is already well developed for information retrieval in an individual language in which the reader is fluent. However the only sources of information in other languages remain foreign correspondents and international news services. Advances in cross-language information retrieval and machine translation suggest that this problem may be eased by the development of translingual information access applications.

This paper explores issues for cross-language information access for one of the most challenging tasks: information access between European and Asian languages. In our case we take the example of English and Japanese. We analyse the importance of information retrieval and machine translation in achieving this objective and describe our ongoing work.

The remainder of this paper is organised as follows: Section 2 defines information access and explores the challenges for translingual technology, Section 3 summarises the state-of-the-art for English-Japanese machine translation, Section 4 outlines current information retrieval procedures, and Section 5 explores the approaches to cross-language information retrieval and access. Section 6 describes an initial experiment in cross-language access to Japanese news texts using the BMIR-J2 test collection. Finally, Section 7 describes current conclusions and further research directions.

2 Information Access

2.1 Definitions

When using an *information retrieval* (IR) system a user is, in general, primarily interested in *accessing* information contained in documents indexed by the retrieval system. IR is usually taken to

be the location and retrieval of documents potentially relevant to a user's information need. It is assumed within this scenario that when a document has been retrieved, the user will be able to decide if it is relevant, and if it is, to extract useful information from the document. In this paper we extend this definition of IR to the complete process of knowledge acquisition which we refer to as *information access* (IA). In IA we view the extraction of information from retrieved documents as an integral part of the information seeking process.

Interest in information retrieval research has expanded significantly in recent years, although much research is still focussed on several well established models including: the vector-space model [1], the probabilistic model [2] and inference networks [3]. These models have been extensively researched and evaluated. In recent years much of this work has concentrated on the US NIST TREC (Text REtrieval Conference) [4]. Commercial online text retrieval systems, such as Alta Vista and Lycos, have proliferated rapidly and the larger ones now contain index information for millions of documents. Using these systems information seekers are able to enter search requests in natural language and receive an interactive response.

While most information retrieval systems are restricted to single language or *monolingual* operation, there is increasing interest in the ability to query between different languages [5] [6] [7]. In this scenario, referred to as *cross-language information retrieval* (CLIR), queries are entered in one language and documents retrieved in one or more other languages. A detailed review of methods for CLIR is contained in [8].

One possible method for CLIR is to use automatic *machine translation* (MT) to translate the documents into the query language. Unfortunately there are practical, as well as technical, drawbacks to this approach. To implement this approach various translation scenarios might be considered. For example, all search engines could translate and index all documents into all possible query languages. However, this is rather impractical. Maintaining index files in multiple languages may not be feasible due to their overall size, and the maintenance overhead on the index files would potentially be very large. It is possible that special interest subscribers could pay for such a service in a specific domain, but the overall market would be limited. A more practical alternative is to translate the query into the original document language online and use this to retrieve these documents. This option is much more flexible since it allows the query to be translated into any desired document language, subject to the availability of an MT system for this language pair.

2.2 Cross-Language Information Access

Our scenario for cross-language information access (CLIA) extends the CLIR paradigm to incorporate various possible post-retrieval processes including: full MT, text summarisation, MT for content gisting, information extraction, and graphical content-visualisation. Figure 1 shows an example CLIA system which includes post-retrieval MT. The first stages follow a standard CLIR path: the user enters a query in their native language, the query is translated into the desired document language, and applied to the IR engine. Current IR systems typically present the user with a list of documents ranked by retrieval matching score, and the title and often the first sentence of each document in the list. Using this information the user selects potentially relevant documents.

The scenario as described so far is a standard CLIR system. For our CLIA system to assist the user with the initial stage of relevance judgement we could use MT to translate the title and first sentence of each document provided in the document language into the language of the query. This additional information could be presented to the user in an augmented ranked list.

When the user selects a document it could be automatically translated into the query language. Although the long term goal of fully-automatic high quality MT has not yet been achieved, today's MT systems offer an invaluable tool for gisting the document contents. A practical strategy to do this would have to be designed carefully since MT is in general computationally expensive, and the translation output will usually contain at least stylistic flaws and sometimes mistakes. However, it is important to remember that the user is interested in *information*, not perfect prose or necessarily a translation of the complete document. Fluent readers can usually easily read a document despite stylistic clumsiness and spot translation errors due to contextual inconsistency. One could view MT as assisting the user in extracting the required information.

One of the most challenging CLIA domains is content access between languages with different scripting systems. For example, Europeans can often make reasonable guesses at the contents of documents in another European language, whereas they are often completely unable to access information in Asian language documents. The same is often true to some extent in reverse for many Asian language speakers.

The foregoing discussions assume that the required component technologies are in place and that their performance levels are sufficient to pro-

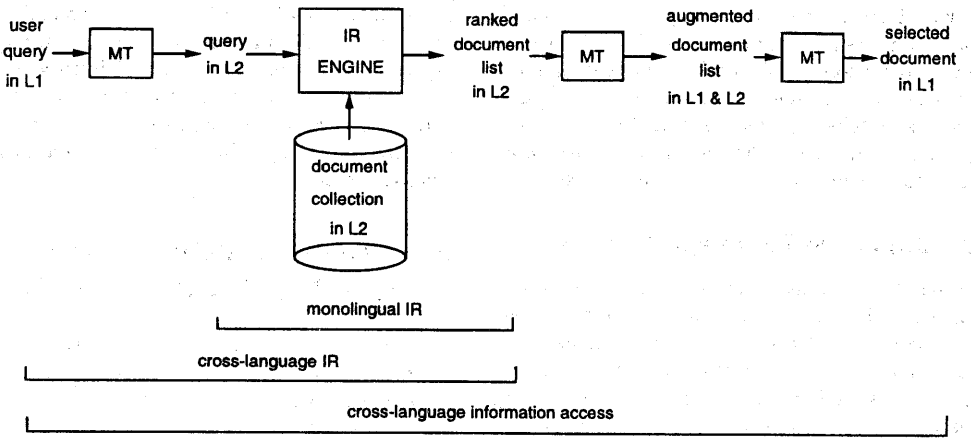

Figure 1: Flow diagram for a basic Cross-Language Information Access (CLIA) system.

vide useful IA. These scenarios can only be properly explored in the laboratory if suitable test collections are available to assess CLIR, MT and most importantly how useful a system is in assisting non-native speakers in accessing information. Unfortunately, such test collections do not currently exist. As a starting point for our work we describe a preliminary experiment using a simulated cross-language information-access task.

3 Machine Translation

The role of translation in CLIR is essentially to bridge the gap between surface forms of terms in the query and document languages. Much of the previous work in CLIR such as [4] [5] [6] has looked at CLIR for European language pairs and has avoided many of the challenges which we face in processing European-Asian language pairs. In the latter case particular difficulties arise because the language pairs are not cognates, so for example, a word in English may appear as a phrase or a word and a particle (*bunsetsu*) in Japanese. We also find that the level of lexical transfer ambiguity, i.e. the number of different translations which a word can have, is higher in such language pairs than in say English-French or English-Spanish.

The three major practical challenges which we face in CLIR are: *coverage*: providing sufficient bilingual knowledge; *disambiguation*: how to identify conceptually different forms from the set of possible translations of a query word; and *synonym selection*: how to identify conceptually equivalent forms of a translation.

MT using deep linguistic analysis is a core-technology for providing solutions in all of these areas. The main limitations which arise in adapting MT to IR are in the coverage of the bilingual dictionaries and in the amount of context available in short IR queries, where it is difficult for linguistic analysis to succeed. These problems are non-trivial and increase as the scope of language to be processed increases. Although it is generally supposed that MT quality is too unreliable to translate queries, we show in this paper that reasonable CLIR performance can be achieved without any modification to an existing MT system.

4 Information Retrieval

In this section we briefly review the techniques typically adopted for monolingual English and Japanese text retrieval.

4.1 English Language Retrieval

The vast majority of information retrieval research has been carried out on English language document collections. For this reason English text retrieval is the best understood and the techniques adopted the most extensively evaluated, recently within the TREC program [4]. TREC has provided large retrieval test collections which have enabled different approaches to information retrieval to be compared and contrasted.

The techniques used for English language retrieval by different researchers vary to some extent, but most generally employ the following stages. First, the text is conditioned to remove standard common stop words (usually short function words). The remaining content words are then suffix stripped to encourage matching between differ-

ent word forms. The suffix stripped words (search *terms*) are then statistically weighted based on their distribution within each document and the overall document archive. In retrieval the search request is matched against document and a matching score computed. The user is than presented with a document list ranked by matching score.

4.2 Japanese Language Retrieval

Compared to English, relatively little work has been carried out on information retrieval for Japanese text. All test retrieval collections available to date have been very small so it is not possible to draw definite conclusions about existing research results. However, the work which has appeared suggests that weighting schemes developed for English language transfer well to Japanese [9]. Although further experimentation with larger collections is obviously required to establish these findings, for the purposes of this work we assume the existing indicative results to be reliable.

The most basic problem for Japanese text retrieval is that it is an *agglutinating* language. In order to index the documents for retrieval some method of extracting representative indexing units must be adopted. Two methods are generally available for extracting indexing units:

- Morphological Segmentation: the continuous string of characters is divided into words using a dictionary-based morphological analyser. The morphological analyser extracts whole words by comparing each character string against entries in a dictionary. The analyser tends to extract component words (or *morphemes*) from compound words as separate indexing units. Unfortunately segmentation errors can arise due to ambiguity of word boundaries and limitations in the morphological analyser, such as its inability to identify words outside its dictionary.

- Character-based Indexing: individual characters or (usually overlapping) fixed length character n-grams are automatically extracted from the character strings and used as the indexing units. In this approach no linguistic analysis is performed and possible word boundaries are ignored.

Once the indexing units have been extracted appropriate text conditioning can be carried out. A description of the possible requirements for Japanese language text conditioning in IR and potential strategies is beyond the scope of this paper, but a good review is contained in [10]. A detailed analysis of text conditioning for Japanese language IR is an important area for future research when suitable collections become available. After applying text conditioning, a standard term weighting method can be applied to Japanese text.

In retrieval a request is processed to produce appropriate indexing units, which are then matched against the documents.

5 Cross-Language Information Retrieval Methods

CLIR is a relatively new area of research and various methods are under investigation by researchers around the world. These methods used can generally be divided into the following categories:

- Dictionary Query Term Lookup: individual terms in the query are replaced by one or more possible translations in the document language taken from a bilingual dictionary [5]. The principal advantages of this approach are that online bilingual dictionaries are becoming increasingly common, and that the translation process is computationally very cheap due to its low level of analysis. Its main disadvantage is that ambiguity is introduced when individual terms are replaced by several alternative terms. These extra terms are sometimes semantically unrelated to the original term in its current context. Techniques based on relevance feedback to overcome these ambiguity problems are explored in [11].

- Parallel-corpora based Query Translation: terms occurring in similar contexts in aligned "parallel" (more often "comparable") corpora in different languages are identified. When the user enters a query a number of related terms in the other language can be generated in a form of query expansion [6]. The main advantage of this method is that it is less prone to the ambiguity problems of dictionary based methods. Its main disadvantage is that parallel corpora are not widely available in domains outside international news services.

- Machine Translation: the query and/or the document are translated using full machine translation with linguistic analysis. The main attraction of this approach is that the ambiguity of terms should be greatly reduced by taking their context into account via the linguistic analysis in the translation process. The main disadvantages of this method are the computation expense of the MT process, and the inaccuracy of current translation systems

when used outside specific domains. Inaccuracy in translation is particularly a problem where there is little contextual information, which unfortunately is exactly the situation often encountered in search requests for information retrieval. Although widely discussed in the context of CLIR, little work on this method has appeared publically [8] [12].

In the experiment described in the next section we investigate the use of both dictionary-based term lookup and full MT of each query.

6 A CLIA Experiment

In this section we describe a preliminary simulation experiment exploring the scenario of an English speaking researcher searching for information in Japanese news stories. Our researcher will need to make use of CLIR to locate potentially relevant documents, and require the assistance of MT to decide whether a document is relevant and to access the information it contains.

For this experiment we use the Toshiba Japanese language NEAT IR system [13] and Toshiba ASTRANSAC MT system [14].

6.1 The NEAT System

The NEAT Information Retrieval system is being developed for the retrieval of online Japanese text articles. Documents are indexed separately using both morphological segmentation and character-based analysis. When a Japanese language request is entered it is morphologically segmented, the request-article matching score is computed independently for the two forms of indexing and these scores summed for each document. A list of articles ranked by the query-document summed matching scores is returned to the user.

6.1.1 Term Weighting

In this experiment the NEAT System makes use of the BM25 probabilistic combined weight (cw) [15]. The BM25 weight has been shown to be effective not only for English text retrieval, but also where documents have been imperfectly indexed, for example in Chinese text retrieval [16].

The BM25 cw weight is calculated as follows,

$$cw(i,j) = \frac{cfw(i) \times tf(i,j) \times (K1+1)}{K1 \times ((1-b) + (b \times ndl(j))) + tf(i,j)}$$

where $cw(i,j)$ represents the weight of term i in document j, $cfw(i)$ is the standard collection weight (often referred to as inverse document frequency weight), $tf(i,j)$ is frequency of term i in document j, and $ndl(j)$ is the normalised length of document j. $ndl(j)$ is calculated as,

$$ndl(j) = \frac{dl(j)}{\text{Average } dl \text{ for all documents}},$$

where $dl(j)$ is the length of j. $K1$ and b are empirically selected tuning constants for a particular collection. $K1$ is designed to modify the degree of effect of $tf(i,j)$, while constant b modifies the effect of document length. High values of b imply that documents are long because they are verbose, while low values imply that they are long because they are multitopic.

The query-document matching score is computed by summing the weights of terms present in query and the document.

6.2 The ASTRANSAC Machine Translation System

The ASTRANSAC MT system is widely used for translating Internet pages from English to Japanese and so we feel it offers the necessary general language coverage to succeed for a news domain. Translation is fully automatic and this frees the user to concentrate on the information seeking task. The translation model in ASTRANSAC is the *transfer method* (for example see [17]), following the standard process of morphological analysis, syntactic analysis, semantic analysis and selection of translation words. Analysis is top-down and uses ATNs (Augmented Transition Networks) on a context-free grammar. In our simulation we used a 65,000 term common word bilingual dictionary and 14,000 terms from a proper noun bilingual dictionary which we consider to be relevant to news events covered in the document collection used in our experiment.

For this experiment ASTRANSAC is used to automatically translate queries from English into Japanese, and also to translate individual documents for user browsing after retrieval.

6.3 BMIR-J2 Japanese Retrieval Test Collection

Our simulation experiment uses the BMIR-J2 Japanese retrieval collection. The BMIR-J2 collection consists of 5080 articles taken from the Mainichi Newspapers in the fields of economics and engineering, and a total of 50 main search requests[1] . Each request consists of a natural language phrase describing a user's information need.

[1] Data in BMIR-J2 is taken from the Mainichi Shimbun CD-ROM 1994 data collection. BMIR-J2 was constructed

Relevant documents for each query were identified as follows. A broad Boolean expression was used to identify most possible relevant documents. The retrieved documents were manually assessed for relevance to the query and the assessment cross-checked by another assessor. The average number of relevant documents for each query is 33.6.

BMIR-J2 was designed so that some search requests can be satisfied very easily, while for some others it is very difficult to retrieve the relevant documents using the request.

6.3.1 English Language Queries

For our simulation the BMIR-J2 requests were translated into English by a bilingual native Japanese speaker. The objective of this translation process was to produce queries which used reasonably good native English while preserving the meaning of the original Japanese. In this experiment we assume that these requests have been generated by the English speaking information seeker hypothesised as the start of this section.

6.3.2 Example Query

The original text of one of the BMIR-J2 Japanese queries is,

電話料金の値下げ

this was manually translated as the English phrase,

reduction of telephone rates

after translation using the ASTRANSAC MT system the following Japanese query was produced

電話料金の縮小

Clearly after translation this is somewhat different to the original Japanese query, although the basic meaning of the query has been preserved.

Inspection of the machine translated queries showed that while some were identical to the original query others were quite different. Some of these variations will have been introduced due to problems in the MT process, however others will be due to the inexact nature of the manual Japanese-English translation.

6.4 Experimental Results

In our experiments we compare retrieval performance using the original Japanese queries with those generated using automatic translation. Retrieval performance is measured in terms of *precision*, the proportion of retrieved documents which are relevant to the search request, and *recall*, the proportion of relevant documents which have been retrieved. Table 1 shows retrieval performance for the original Japanese queries and the automatically translated versions. Precision is shown at ranked list cutoff levels of 5, 10, 15, and 20 documents and the average precision which is calculated by averaging the precision values at the position of each relevant documents for each query, and then taking the average across the query set. Figure 2 shows a corresponding recall-precision curve.

The results in Table 1 and Figure 2 show that as expected retrieval performance is degraded for the automatically translated queries. For the cross-language queries MT is clearly superior to the DTL method. To the best of our knowledge this is the

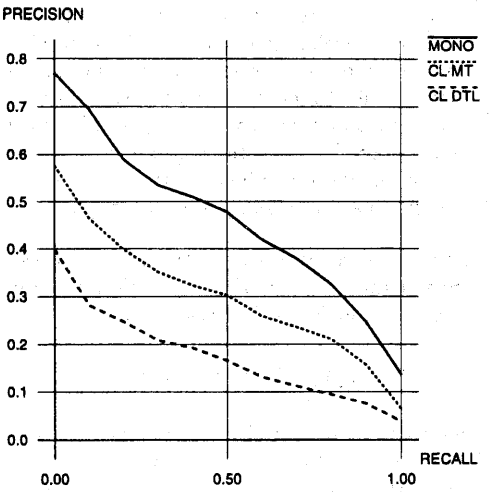

Figure 2: Recall-Precision curve for BMIR-J2 using: monolingual IR (MONO); CLIR with full query MT (CL MT); CLIR with query dictionary lookup (CL DTL).

		MONO	CL	
			MT	DTL
Prec.	5 docs	0.588	0.396	0.196
	10 docs	0.508	0.342	0.194
	15 docs	0.463	0.333	0.192
	20 docs	0.425	0.307	0.185
Av Precision		0.451	0.289	0.161
% change		—	-35.9%	-64.3%

Table 1: Retrieval Precision for BMIR-J1 using: monolingual IR (MONO); CLIR with full query MT (MT); CLIR with query dictionary lookup (DTL).

by the SIG-Database Systems of the Information Processing Society of Japan, in collaboration with the Real World Computing Partnership.

first CLIR result to indicate this clearly. Although we realise that our result must be treated with caution due to the very small size of the retrieval collection, and the inexact nature of the design of our simulation experiment. However, we feel this result suggests that MT should be explored further for CLIR than so far appears to be the case [8]. Finally, it should be emphasised that we have made no attempt to modify the translation dictionaries to the BMIR-J2 task, and thus we feel overall that the results for CLIR relative to monolingual IR are quite encouraging.

6.5 Selection and Browsing of Retrieved Documents

In order for the user to select a potentially relevant document in an informed way, and to access the information it contains, MT must be used. Figure 3 shows the top two ranked documents retrieved in response to the example query given previously. The document headings and their first sentence are shown in their original Japanese and then in English as generated by the ASTRANSAC MT system. The original Japanese information is assumed to be supplied by the search engine supplying the document list. Headings and similar short statements are a challenging translation domain for MT systems since they are often written in a terse style which is not typical of the language. There will be some computational overhead associated with this generating these translations online, but the amount of text involved is very small and the translation overhead should not noticeably interfere with the user's interaction with the retrieval system.

From the English language information shown in Figure 3 our information seeker is able to gist the possible subject of each document. The user is able only to gist the possible contents even in a monolingual system, our hope here is that the translation quality will be sufficient not to mislead the user. In Figure 3 it is clear from the MT translations that document 001339 is unlikely to be relevant to the query, but that document 001077 may be relevant.

Individual documents can be retrieved from their original server for browsing, as is common practise when using monolingual search engines. Once the document has been retrieved there are a number of translation options. MT could be used to generate the best possible translation of the whole document, but this may prove too slow for interactive information seeking. Alternatively less precise translation methods could be used to provide a faster rough translation and the user allowed to select portions for more careful translation.

7 Conclusions and Further Work

In this paper we have described a paradigm for cross-language information access. We demonsrated that, by combining existing approaches to CLIR and MT, it is already possible to build potentially useful CLIA applications, even for the difficult task of access between Asian and European languages. Improvements in any of the component technologies can be expected to improve the effectiveness of the overall system.

In our ongoing work we are exploring alternative dictionary-based translation methods and application of appropriate feedback techniques [11].

Specific further work should include the development of test collections of sufficient size and diversity to enable alternative strategies for IR, MT and CLIA to be thoroughly evaluated and contrasted individually and in combination. Design of such collections is itself a challenging and expensive task [18]. The evaluation requirements of the existing individual technologies, as well as any novel ones introduced by the CLIA paradigm would have to be considered carefully in the design of suitable collections.

The importance of applications of this type can only increase with the ever increasing volume of online digital information. Improvements to CLIA systems will be supported by ongoing advances in IR and MT technology.

References

[1] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24:513–523, 1988.

[2] S. E. Robertson. The probability ranking principle in IR. *Journal of Documentation*, 33:294–304, 1977.

[3] H. Turtle and W. B. Croft. Inference networks for document retrieval. In *Proceedings of ACM SIGIR 90*, pages 1–24, New York, 1990. ACM.

[4] D. K. Harman and E. M. Voorhees, editors. *The Sixth Text REtrieval Conference (TREC-6)*, Gaithersburg, MD, 1998. NIST.

[5] D. A. Hull and G. Grefenstette. Querying across languages: A dictionary-based approach to multilingual information retrieval. In *Proceedings of ACM SIGIR 96*, pages 49–57, Zurich, 1996. ACM.

RANK 1 ARTICLE: 001339
HEADING:
ORIGINAL: [ニッポン政治考] 消費者が利益享受できる規制緩和を＝スミス佐々木三根子。
TRANSLATED: It is deregulation which can carry out profits enjoyment of the [ニッポン political idea] consumer = Smith 佐々木 Mine child.
FIRST SENTENCE:
ORIGINAL: <モルガン・スタンレー証券東京支店シニアエコノミスト>。日米の経済関係は、必然的に衝突するコースを進む、と言わざるを得ない。
TRANSLATED: ¡モルガン Stanley security Tokyo branch senior economist¿. A Japan-U.S. economical relation cannot but say that the course which collides inevitably progresses.

RANK 2 ARTICLE: 001077
HEADING:
ORIGINAL: 4月から4－5割前後の値下げ　自動車・携帯電話の基本料金－－売り切りに対応。
TRANSLATED: price reduction before and behind [April to] 4-50percent basic charge- of a car and a portable telephone - selling out - correspondence
FIRST SENTENCE:
ORIGINAL: 自動車・携帯電話サービスを実施しているＮＴＴ移動通信網（ＮＴＴドコモ）など各社は五日、郵政省に基本料金など各種料金の変更認可申請を提出した。
TRANSLATED: Each company, such as NTT mobile-communications network (NTT DoCoMo) which is carrying out a car and portable-telephone service, submitted on fifth the change approval application of various charges, such as a basic charge, to the Ministry of Posts and Telecommunications.

Figure 3: Example of ranked retrieval list. Article header and first sentence shown in original Japanese and machine translated English. (Article 001339 is judged not-relevant and 001077 is judged relevant.)

[6] P. Sheridan and J. P. Ballerini. Experiments in Multilingual Information Retrieval using the SPIDER system. In *Proceedings of ACM SIGIR 96*, pages 58–65, Zurich, 1996. ACM.

[7] J. Carbonell, Y. Yang, R. E. Frederking, R. D. Brown, Y. Geng, and D. Lee. Translingual Information Retrieval: A Comparative Evaluation. In *Proceedings of IJCAI 96*, pages 708–714, Nagoya, 1996. IJCAI.

[8] D. W. Oard. A survey of multilingual text retrieval. Technical Report UMIACS-TR-9619, University of Maryland, 1996.

[9] G. J. F. Jones, T. Sakai, M. Kajiura, and K. Sumita. Experiments in Japanese Text Retrieval and Routing using the NEAT System. In *Proceedings of ACM SIGIR 98*, Melbourne, 1998. ACM.

[10] H. Fujii and W. B. Croft. A Comparison of Indexing Techniques for Japanese Text Retrieval. In *Proceedings of ACM SIGIR 93*, pages 237–246, Pittsburgh, 1993. ACM.

[11] L. Ballesteros and W. B. Croft. Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval. In *Proceedings of ACM SIGIR 97*, pages 84-91, Philadelphia, 1997. ACM.

[12] N. Collier, H. Hirakawa, and A. Kumano. Machine translation vs. dictionary term translation - a comparison for English-Japanese news article alignment. In *Proceedings of COLING-ACL'98*, Montreal, 1998. ACL.

[13] M. Kajiura, S. Miike, T. Sakai, M. Sato, and K. Sumita. Development of the NEAT Information Filtering System. In *Proceedings of the 54th IPSJ National Conference*, pages 3–(299–300), Tokyo, 1997. IPSJ.

[14] H. Hirakawa, H. Nogami, and S. Amano. EJ/JE machine translation system AS-TRANSAC - extensions towards personalization. In *Proceedings of the Machine Translation Summit III*, pages 73–80, 1991.

[15] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In *Proceedings of ACM SIGIR 94*, pages 232–241, Dublin, 1994. ACM.

[16] M. M. Beaulieu, M. Gatford, X. Huang, S. E. Robertson, S. Walker, and P. Williams. Okapi at TREC-5. In *Proceedings of the Fifth Text REetrieval Conference (TREC-5)*. NIST, 1997.

[17] W. J. Hutchins and H. L. Somers. *An Introduction to Machine Translation*. Academic Press Limited, London, 1992.

[18] K. Sparck Jones and J. R. Galliers. *Evaluating Natural Language Processing Systems*, volume 1083 of *Lecture Notes in Artificial Intelligence*. Springer, 1996.