An Intelligent Multilingual Information Browsing and Retrieval System Using Information Extraction

Chinatsu Aone and **Nicholas Charocopos** and **James Gorlinsky**
Systems Research and Applications Corporation (SRA)
4300 Fair Lakes Court
Fairfax, VA 22033
aonec@sra.com

Abstract

In this paper, we describe our multilingual (or cross-linguistic) information browsing and retrieval system, which is aimed at monolingual users who are interested in information from multiple language sources. The system takes advantage of *information extraction* (IE) technology in novel ways to improve the accuracy of cross-linguistic retrieval and to provide innovative methods for browsing and exploring multilingual document collections. The system indexes texts in different languages (e.g., English and Japanese) and allows the users to retrieve relevant texts in their native language (e.g., English). The retrieved text is then presented to the users with proper names and specialized domain terms translated and hyperlinked. Moreover, the system allows interactive information discovery from a multilingual document collection.

1 Introduction

More and more multilingual information is available on-line every day. The World Wide Web (WWW), for example, is becoming a vast depository of multilingual information. However, monolingual users can currently access information only in their native language. For example, it is not easy for a monolingual English speaker to locate necessary information written in Japanese. The users would not know the query terms in Japanese even if the search engine accepts Japanese queries. In addition, even when the users locate a possibly relevant text in Japanese, they will have little idea about what is in the text. Outputs of off-the-shelf machine translation (MT) systems are often of low-quality, and even "high-end" MT systems have problems particularly in translating proper names and specialized domain terms, which often contain the most critical information to the users.

In this paper, we describe our multilingual (or cross-linguistic) information browsing and retrieval system, which is aimed at monolingual users who are interested in information from multiple language sources. The system takes advantage of *information extraction* (IE) technology in novel ways to improve the accuracy of cross-linguistic retrieval and to provide innovative methods for browsing and exploring multilingual document collections. The system indexes texts in different languages (e.g., English and Japanese) and allows the users to retrieve relevant texts in their native language (e.g., English). The retrieved text is then presented to the users with proper names and specialized domain terms translated and hyperlinked. The system also allows the user in their native language to browse and *discover* information buried in the database derived from the entire document collection.

2 System Description

The system consists of the Indexing Module, the Client Module, the Term Translation Module, and the Web Crawler. The Indexing Module creates and loads indices into a database while the Client Module allows browsing and retrieval of information in the database through a Web browser-based graphical user interface (GUI). The Term Translation Module is bi-directional; it dynamically translates user queries into target foreign languages and the indexed terms in retrieved documents into the user's native language. The Web Crawler can be used to add textual information from the WWW; it fetches pages from user-specified Web sites at specified intervals, and queues them up for the Indexing Module to ingest regularly.

For our current application, the system indexes names of people, entities, and locations, and scientific and technical (S&T) terms in both English and Japanese texts, and allows the user to query and browse the database in English. When Japanese texts are retrieved, indexed terms are translated into English.

This system is designed to expand to other lan-

guages besides English and Japanese and other domains beyond S&T terms. Moreover, the English-centric browsing and retrieval mode can be switched according to the users' language preference so that, for example, a Japanese user can query and browse English documents in Japanese.

2.1 The Intelligent Indexing Module

The Indexing Module indexes names of people, entities, and locations and a list of scientific and technical (S&T) terms using state-of-the-art IE technology. It uses different configurations of the same fast indexing engine called NameTag^{TM} for different languages. Two separate configurations ("indexing servers") are used for English and Japanese, and how the English and Japanese indexing servers work is described in (Krupka, 1995; Aone, 1996).

In the Sixth Message Understanding Conference (MUC-6), the English system was benchmarked against the Wall Street Journal blind test set for the name tagging task, and achieved a 96% F-measure, which is a combination of recall and precision measures (Adv, 1995),. Our internal testing of the Japanese system against blind test sets of various Japanese newspaper articles indicates that it achieves from high-80 to low-90% accuracy, depending on the types of corpora. Indexing names in Japanese texts is usually more challenging than English for two main reasons. First, there is no case distinction in Japanese, whereas English names in newspapers are capitalized, and capitalization is a very strong clue for English name tagging. Second, Japanese words are not separated by spaces and therefore must be segmented into separate words before the name tagging process. As segmentation is not 100% accurate, segmentation errors can sometimes cause name tagging rules not to fire or to misfire.

Indexing of names is particularly useful in the Japanese case as it can improve overall segmentation and thus indexing accuracy. In English, since words are separated by spaces, there is no issue of indexing accuracy for individual words. On the other hand, in languages like Japanese, where word boundaries are not explicitly marked by spaces, indexing accuracy of individual words depends on accuracy of word segmentation. However, most segmentation algorithms are more likely to make errors on names, as these are less likely to be in the lexicons. Name tagging can reduce such errors by identifying names as single units.

Both indexing servers are "intelligent" because they *identify* and *disambiguate* names with high speed and accuracy. They identify names in texts dynamically rather than relying on finite lists of names. Thus, they can identify names which they have never seen before. In addition, they can disambiguate types of names so that a person named "Washington" is distinguished from a place called Washington, and a company "Apple" can be distinguished from a common noun "apple." In addition, they can *generate* aliases of names automatically (e.g., "ANA" for "All Nippon Airline") and *link* variants of names within a document.

As the indexing servers process texts, the indexed terms are stored in a relational database with their semantic type information (person, entity, place, S&T term) and alias information along with such meta data as source, date, language, and frequency information. The system can use any ODBC (Open DataBase Connectivity)-compliant database, and form-based Boolean queries from the Client Module, similar to those seen in any Web search engine, are translated into standard SQL queries automatically. We have decided to use commercial databases for our applications as we are not only indexing strings of terms but also adding much richer information on indexed terms available through the use of IE technology. Furthermore, we plan to apply data-mining algorithms to the resulting databases to conduct advanced data analysis and knowledge discovery.

2.2 The Client Module

The Client Module lets the user both retrieve and browse information in the database through the Web browser-based GUI. In the query mode (cf. Figure 1), a form-based Boolean query issued by a user is automatically translated into an SQL query, and the English terms in the query are sent to the Term Translation Module. The Client Module then retrieves documents which match either the original English query or the translated Japanese query. As the indices are names and terms which may consist of multiple words (e.g, "Bill Clinton," "personal computer"), the query terms are delimited in separate boxes in the form, making sure no ambiguity occurs in both translation and retrieval. The user has the choice of selecting the sources (e.g, Washington Post, Nikkei Newspaper, Web pages), languages (e.g., English, Japanese, or both), and specific date ranges of documents to constrain queries.

In the browsing mode, the Client Module allows the user to browse the information in the database in various ways. As an overview of the database content, the Client Module lets the user browse the top 25 and 50 most frequent entity, person, and location names and S&T terms in the database (cf. Figure 4). Once the user selects a particular document for viewing, the client sends the document to an appropriate (i.e., English or Japanese) indexing server for creating hyperlinks for the indexed terms and in the case of a Japanese document, sends the indexed terms to the Term Translation Module to translate the Japanese terms into English. The result that the user browses is a document each of whose indexed terms are hyperlinked to other documents containing the same indexed terms (cf. Figure 2). Since hy-

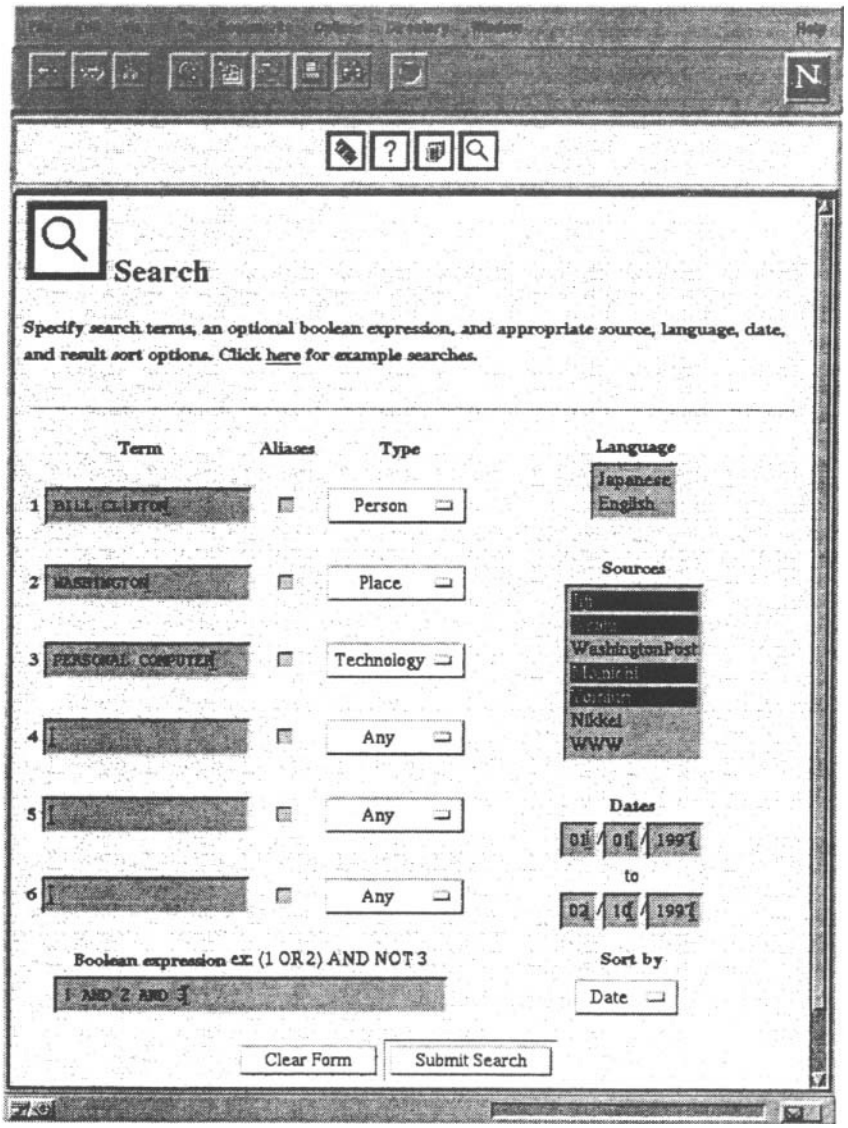

Figure 1: The Search Screen

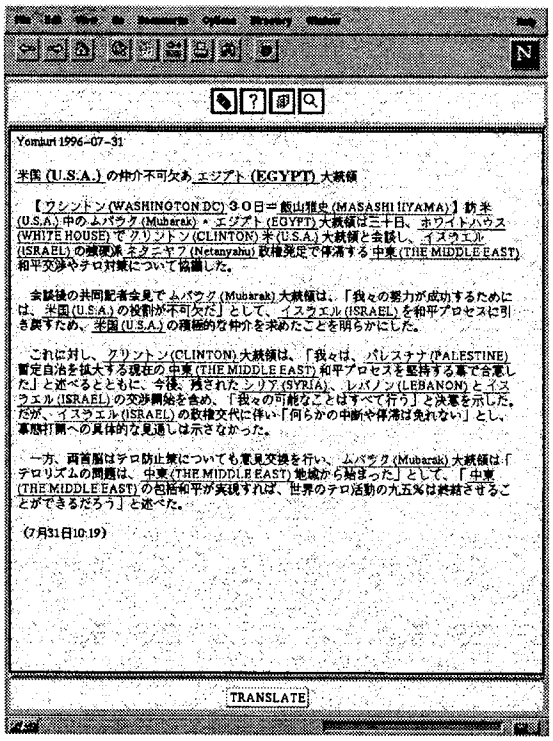

Figure 2: Translated and Hyperlinked Terms

perlinking is based on the original or translated *English* terms, the user can follow the links to both English and Japanese documents transparently. In addition, the Client Module is integrated with a commercial MT system for rough translation. A document which the user is browsing can be translated on the fly by clicking the TRANSLATE button.

2.3 The Term Translation Module

The Term Translation Module is used by the Client Module bi-directionally in two different modes. It translates English query terms into Japanese in the query mode and translates Japanese indexed terms into English for viewing of a retrieved Japanese text in the browsing mode.

This translation module is sensitive to the semantic types of terms it is translating to resolve translation ambiguity. Thus, if a term can be translated in one way for one type and in another way for another type, the Term Translation Module can output appropriate translations based on the type information. For example, in translating Japanese text into English, a single *kanji* (Chinese) character standing for England can be also a first name of a Japanese personal name, which should be translated to "Hide" and not "England." In translating an English query into Japanese, a company "Apple" should be translated into a transliteration in *katakana* and not into a Japanese word meaning a fruit apple.

The Term Translation Module uses various resources and methods to translate English and Japanese names. We use automated methods as much as possible to reduce the cost of creating a large name lexicon manually.

First, this module is unique in that it creates on the fly English translations of *hiragana* names and personal names. Hiragana names are transliterated into English using the hiragana-to-romaji mapping rules. Japanese personal names are translated by finding a combination of first and last names which spans the input.[1] Then, each of the name parts is translated using the Japanese-English first and last name lexicons.

In addition, in order to develop a large lexicon of English names and their Japanese translations, which are transliterated into *katakana*, we have automatically generated katakana names from phonetic transcriptions of English names. We have written rules which maps phonetic transcriptions to katakana letters, and generated possible Japanese katakana translations for given English names. As transliterations of the same English names may differ, multiple katakana translations may be generated for single English names.[2]

The remaining terms are currently translated using the English-Japanese translation lexicons, and we are expanding the lexicons by utilizing on-line resources and corpora and a translation aiding tool.

3 Utilizing IE in Multilingual Information Access

The system applies *information extraction* technology (Adv, 1995) to index names accurately and robustly. In this section, we describe how we have incorporated this technology to improve multilingual information access in several innovative ways.

3.1 Query Disambiguation

As described in Section 2.1, the Indexing Module not only identifies names of people, entities and locations but also disambiguates types among themselves and between names and non-names. Thus, if the user is searching for documents with the location "Washington (not a person or a company named "Washington"), a person "Clinton" (not a location), or an entity "Apple" (not fruit), the system allows the user to specify, through the GUI, the type of each query term (cf. Figure 1). This ability to disambiguate types of queries not only constrains the search and hence improves retrieval precision but also speeds

[1] The Japanese Indexing Module does not specify if an identified name is a first name, a last name, or a combination of first and last name. Since there is no space between first and last names in Japanese, this must be automatically determined.

[2] This is still an experimental effort, and we have not evaluated the quality of generated translations quantitatively yet.

up the search time considerably especially when the database is very large.

3.2 Translation Disambiguation

In developing the system, we have intentionally avoided an approach where we first translate foreign-language documents into English and index the translated English texts (Fluhr, 1995; Kay, 1995; Oard and Dorr, 1996). In (Aone et al., 1994), we have shown that, in an application of extracting information from foreign language texts and presenting the results in English, the "MT first, IE second" approach was less accurate than the approach in the reverse order, i.e., "IE first, MT second". In particular, translation quality of names by even the best MT systems is poor.

There are two cases where an MT system fails to translate names. First, it fails to recognize where a name starts and ends in a text string. This is a non-trivial problem in languages such as Japanese where words are not segmented by spaces and there is no capitalization convention. Often, an MT system "chops up" names into words and translates each word individually. For example, among the errors we have encountered, an MT system failed to recognize a person name "Mori Hanae" in kanji characters, segmented it into three words "mori," "hana," and "e" and translated them into "forest," "England" and "blessing," respectively.

Another common MT system error is where the system fails to make a distinction between names and non-names. This distinction is very important in getting correct translations as names are usually translated very differently from non-names. For example, a personal name "Dole" in katakana was translated into a common noun "doll" as the two have the same katakana string in Japanese. Abbreviated country names for Japan and United States in single kanji characters, which often occurs in newspapers, were sometimes translated by an MT system into their literal kanji meanings, "day" and "rice," respectively.

Our system avoids these common but serious translation errors by taking advantage of the Indexing Module's ability to identify and disambiguate names. In translating terms from Japanese to English in the browsing mode, the Indexing Module identifies names correctly, avoiding the first type of translation errors. Then, the Term Translation Module utilizes type information obtained by the Indexing Module to decide which translation strategies to use, thus overcoming the second type of error.

3.3 Intelligent Query Expansion and Hyperlinking

As described in Section 2.1, the Indexing Module automatically identifies aliases of names and keeps track of such alias links in the database. For example, if "International Business Machine" and "IBM" appears in the same document, the system records in the database that they are aliases.

The system uses this information in automatically expanding terms for query expansion and hyperlinking. At the query time, when the user types "IBM" and chooses the *alias* option in the search screen (see Figure 1), the query is automatically expanded to include its variant names both in English and Japanese, e.g., "International Business Machine," "International Business Machine Corp." and Japanese translations for "IBM" and their aliases in Japanese. This is especially useful in retrieving Japanese documents because typically the user would not know various ways to say "IBM" in Japanese. The automated query expansion thus improves retrieval recall without manually creating alias lexicons.

The same alias capability is also used in hyperlinking indexed terms in browsing a document. For example, when a user follows a hyperlink "United States," it takes the user to a collection of documents which contains the English term "United States" and its aliases (e.g., "US," "U.S.A." etc.), and the Japanese translations of "United States" and their aliases. The result is a truly transparent multilingual document browsing and access capability.

3.4 Information Discovery

One of the biggest advantages of introducing IE technology into information access systems is the ability to create rich structured data which can be analyzed for "buried" information. Our multilingual capability enables the merging of possibly complementary data from both English and Japanese sources and enriching the available information.

Currently the system offers the user several ways to explore and discover hidden information. Our search capability allows interactive information discovery methods. For example, using the query interface, the user can in effect ask "Which company was mentioned along with Intel in regard to microprocessors?" and the system will return all the articles which mentions "Intel," "microprocessors," *and* one or more company names. The user might see that NexGen and Cyrix often occurs with Intel and find out that they are competitors of Intel in this field. Or the user might ask "Who is related to "Shinshintou Party," a Japanese political party, and the user can find out all the people associated with this party. This type of search capabilities cannot be offered by typical information retrieval systems as they treat words as just strings and do not distinguish their semantic attributes.

Furthermore, as we discussed earlier in Section 2.2, browsing documents by following hyperlinks allows a user to discover related information effectively. For example, when the user searches for documents on "NEC Corp.", selects one of the returned documents, and finds another company name

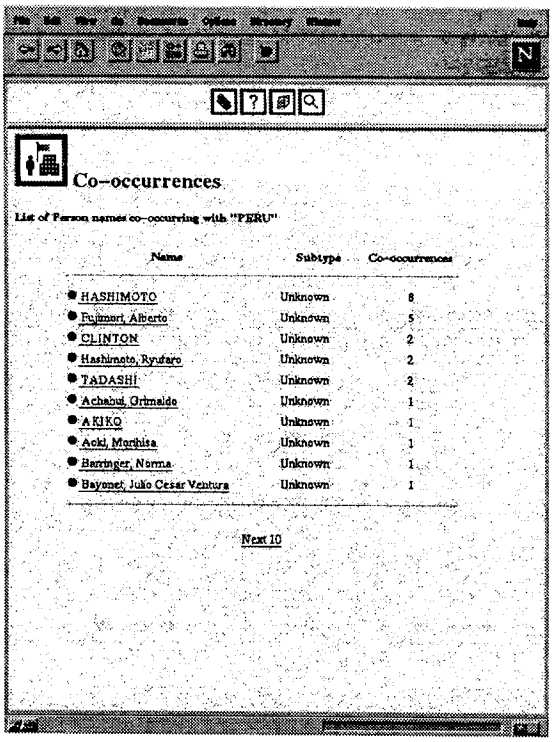

Figure 3: Person Names Co-occurring with Peru

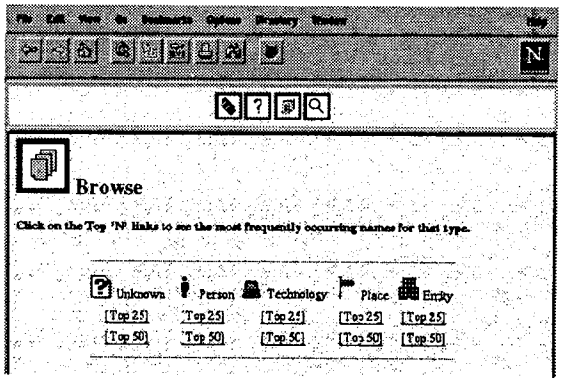

Figure 4: The 25 and 50 Most Frequent Names

"Toshiba" mentioned in this document, the user can establish an immediate connection and follow the link from "Toshiba" to other English *and* Japanese documents which contain that term.

In addition, for each indexed term, the user can explore co-occurring persons, entities, places and technology. For example, Figure 3 shows a list of people co-occurring with the place "Peru." It lists the Japanese prime minister and the Peruvian president at the top (as the Japanese embassy hostage incident occurred recently.)

4 The System Tour

In this section, we give a tour of the system. Figure 4 shows the main Browse screen where the user can browse the top 25 or 50 names of people, entities, locations, and S&T terms. This can provide the user with a snapshot of what is in the database and what types of information are likely to be available.

By following the top 50 entity name link, the user sees the list of entity names in order of frequency (cf. Figure 5). The *Subtype* column in the screen indicates more detailed types of the entity (e.g., organization, company, facility, etc.) From this screen, the user can go to a list of all English and Japanese documents which mention, for example, "Bank of Japan" by clicking the link (cf. Figure 6). The list provides information on the title, length, source, language, and date of each article.

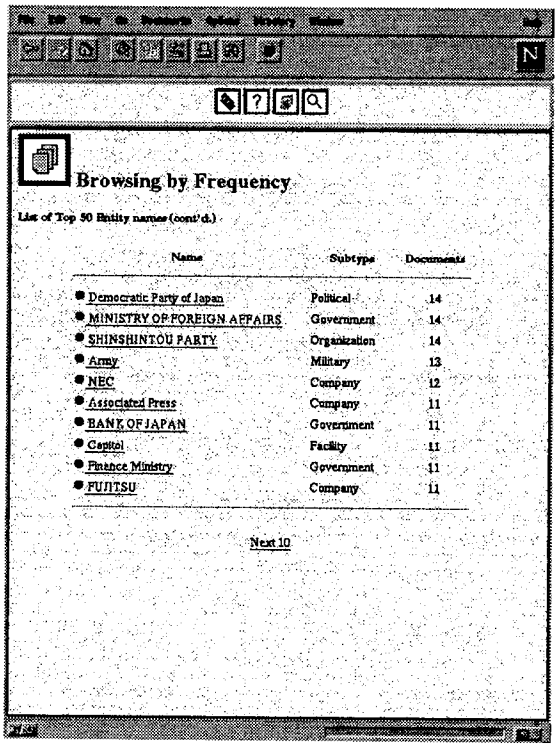

Figure 5: Top 50 Entity Names

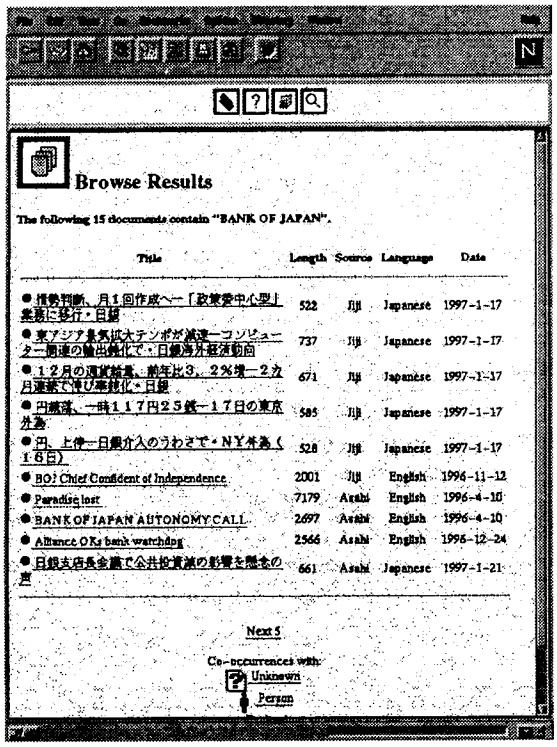

Figure 6: Documents Containing "Bank of Japan"

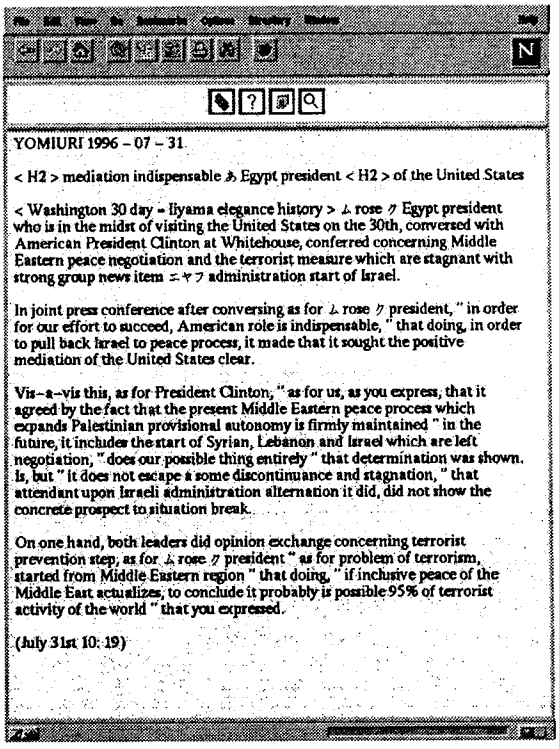

Figure 7: Translation by a Commercial MT system

In the main Search screen (cf. Figure 1), the user types in each query term, including multi-words like "personal computer," in each numbered box. The user can formulate a Boolean query using the box numbers and boolean operators. If not specified, the query terms are joined by "OR". When the **Alias** button is on, query terms are expanded to include their aliases. The **Type** menu allows the user to disambiguate types of query terms. In the **Language** box, the user has the choice of selecting documents in English, Japanese, or both. In addition, the user can constrain sources and the date range of documents, and also sort the results by date, title, and sources.

As discussed in Section 2.2, when the user selects a Japanese article, they can optionally send the article to a commercial MT system for rough translation by pushing the TRANSLATE button (cf. Figure2). Figure 7 shows the translation result for the Japanese document in Figure 2.

5 Summary

We have described an advanced multilingual cross-linguistic information browsing and retrieval system which takes advantage of information extraction technology in unique ways. In addition to its basic capability of allowing a user to send Boolean queries in English against English and Japanese documents and to view the results in semi- and fully translated forms, the system has many innovative capabilities. It can disambiguate query terms to increase precision, expand query terms automatically using aliases to increase recall, and improve translation accuracy significantly by finding and disambiguating names accurately. Moreover, the system allows interactive information discovery from a multilingual document collection by combining IE and MT technologies.

The Indexing Module is currently running on a Sun platform and is designed to scale for a multi-user operational environment. The Web browser-based user interface will work in any Web browser supporting HTML 3.0 on any platform which the Web browser supports, and this ensures a large user base. The system is customizable in several ways. For our current application, the system indexes names and S&T terms, but for other applications we can customize the system to index different types of names and terms. For example, the system can be customized to index product names and financial terms for a business application. Its ODBC-compliance makes porting of databases from one vendor to another very easy. Finally, the system does not assume any particular language combination or target language. Thus, this system can also be used for Japanese monolingual users who want to query and browse in Japanese a set of documents written in English, Japanese, and Spanish.

References

Advanced Research Projects Agency. 1995. *Proceedings of Sixth Message Understanding Conference (MUC-6).* Morgan Kaufmann Publishers.

Aone, Chinatsu. 1996. NameTag Japanese and Spanish Systems as Used for MET. In *Proceedings of Tipster Phase II.* Morgan Kaufmann Publishers.

Aone, Chinatsu, Hatte Blejer, Mary Ellen Okurowski, and Carol Van Ess-Dykema. 1994. A Hybrid Approach to Multilingual Text Processing: Information Extraction and Machine Translation. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas (AMTA).*

Fluhr, Christian. 1995. Multilingual information retrieval. In Ronald A. Cole, Joseph Mariani, Hans Uszkoreit, Annie Zaenen, and Victor Zue, editors, *Survey of the State of the Art in Human Language Technology.* Oregon Graduate Institute.

Kay, Martin. 1995. Machine translation: The disappointing past and present. In Ronald A. Cole, Joseph Mariani, Hans Uszkoreit, Annie Zaenen, and Victor Zue, editors, *Survey of the State of the Art in Human Language Technology.* Oregon Graduate Institute.

Krupka, George. 1995. SRA: Description of the SRA System as Used for MUC-6. In *Proceedings of Sixth Message Understanding Conference (MUC-6).*

Oard, Douglas W. and Bonnie J. Dorr, editors. 1996. *A Survey of Multilingual Text Retrieval.* Technical Report UMIACS-TR-96-19. Institute for Advanced Computer Studies, University of Maryland.