BACKGROUND

Much of the information in digital library or digital information organization applications is in the form of text. Even when the application focuses on multimedia objects such as images, video or audio, these objects are primarily described using text or words and phrases from controlled vocabularies such as the Library of Congress Subject Headings. Compared to the data stored in traditional database systems, text is relatively unstructured and has less well-defined semantics. Consequently, the processes of indexing (both manual and automatic) and retrieving textual information are fundamental to digital library applications. Information retrieval (IR) is an area of both computer and information science that studies these processes.

Research on effective and efficient automatic techniques for IR has been underway in the United States and Europe since the 1960s (Sparck, Jones and Willett 1997). Dramatic improvements in disk storage technology and then the growth of Internet information access led to a significant increase in IR research activity in the 1990s. The U.S. government has also provided substantial funding in this area through efforts such as the DARPA TIPSTER project and NSF Digital Library Initiative. The TIPSTER project focused on the development of new techniques for text retrieval, filtering and extraction. The TREC annual evaluation conference was started by DARPA in conjunction with TIPSTER and is now run by the National Institute of Standards and Technology. TREC has resulted in standard text collections and evaluation methodologies that have increased both the size of the text-related research community and the number of new results.

In recent years, the IR research community in the United States has broadened its interests beyond retrieval to include new areas such as filtering, distributed and scalable IR, cross-lingual IR, information extraction, summarization, visualization, text data mining, and event detection.

In Japan, information retrieval research has not been a major focus of computer and information scientists until much more recently. Because of the relative difficulty of indexing Japanese text, much of the earlier work in Japan focused on hardware and software techniques to support exact matching of Boolean combinations of text strings. The first studies on the effectiveness of ranking techniques with Japanese text databases were not done until the 1990s (e.g., Ogawa et al. 1993) and some of those were done in the United States as part of the TIPSTER project (Fujii and Croft 1993). The advent of the Internet, the visibility of the U.S. government initiatives, and the general commercial interest in text search has changed this situation considerably. Many industry laboratories and university groups are working on a range of technologies for text. In addition, an increasing number of papers from Japan are submitted to the major IR conferences, new conferences focusing on Asian language issues have been started (e.g., IRAL 1997), and government initiatives exist in areas such as digital libraries.
IR RESEARCH IN JAPAN

Japanese IR research covers basically the same areas as in the United States, although there is more emphasis on indexing techniques appropriate for Japanese and other major Asian languages (especially Chinese and Korean). Indexing issues that are common concerns include the following:

Dealing With Different Character Encodings

There are a number of standard two-byte character encodings for Japanese and Chinese (e.g., JIS and GB). Unicode is a developing standard that addresses most of the problems in this area.

Input Methods for Japanese and Other Languages

Segmentation of Japanese into “Words”

A number of techniques based on dictionaries have been proposed to identify the words and phrases in Japanese text for indexing. Some of these techniques are now available commercially and achieve high accuracy rates.

Morphological Analysis

The identification and normalization of inflected word forms is less important for Japanese than for some languages, and techniques have been developed to do this as part of segmentation.

N-gram Indexing

To avoid the segmentation process, many researchers have proposed indexing Japanese (and Chinese) using pairs (bigrams) or trigrams of characters. Experimental results with this technique have shown that retrieval based on n-gram indexing is approximately as effective as retrieval based on segmented words.

The Use of Controlled Vocabularies and Ontologies for Indexing

As more work is done in this area, the language-dependent indexing issues are becoming less important. Full text indexing for Japanese is now relatively straightforward, and the focus of Japanese IR researchers is increasingly on the core issues of effectiveness and efficiency. Some of the specific research areas mentioned during the WTEC industrial and university visits were as follows:

- retrieval models (i.e., how to rank)
- text representation (e.g., concepts vs. words)
- query expansion (using corpus analysis and thesauri to add related words to the query)
- relevance feedback (using learning to improve retrieval)
- summarization (e.g., of documents and groups of documents to support browsing and text data mining)
- interfaces (different perspectives on the database or retrieved information)
- scalability (how to achieve fast indexing and query processing times as databases become much larger)

There was also significant interest in multilingual systems. The major focus here was on supporting the major Asian languages in one system, and to a lesser extent, supporting English. There is some work being done on cross-lingual retrieval (asking questions in one language and retrieving documents in many languages), but it is currently more of an interest rather than an active research topic. Machine translation was mentioned frequently and is considered an important part of a multilingual digital library application. The primary focus in this area was on translation from English to Japanese, particularly for Internet applications. Other Asian language and English translation pairs also exist or are being developed.

A number of digital library applications based on libraries were discussed during the WTEC visit, and these applications typically made use of traditional indexing approaches with controlled vocabularies, manual indexing and catalogs for searching. The issues involved with creating and searching online public access
catalogs (OPACS) as text databases are essentially the same as those studied in the United States. As an extension of this approach, metadata standards for information on the Internet (such as the Dublin Core) are being developed and debated in the international community. Japanese information scientists are part of that debate.

OVERVIEW OF TEXT RESEARCH AT SITES VISITED BY WTEC

All of the sites visited were dealing with text in some form, but a smaller number were actually doing research in text indexing and retrieval or related techniques. These sites were as follows:

- University of Library and Information Science, Tsukuba
- National Center for Science Information Systems (NACSIS)
- HUMI Project, Keio University
- NTT Multi-Media System Laboratory Group
- OMRON Multi-Media Information Technology Research Center
- Advanced Telecommunication Research Institute (ATR)
- Fujitsu
- IBM Tokyo Research Laboratory
- Matsushita Multimedia Systems Research Laboratory
- Hitachi Central Research Laboratory
- Nara Institute of Science and Technology (NAIST)

SUMMARY

In this section, the overall impression of the text and IR research being done in Japan is summarized, and then it is compared to the work being done in the United States.

The first observation is that the Japanese community of computer and information scientists working in the IR and text-related areas is smaller than the comparable communities in the United States and Europe. As a result, Japanese research in these areas tends to follow directions and initiatives begun in the United States. Individual projects are of good quality and are producing interesting technology, but progress has been somewhat impeded by a lack of a Japanese version of TREC or equivalent test collections. Although the value of recall/precision measurements is hotly debated in the IR community, there is no doubt that the culture of experiment and comparison in IR and TREC has led to significant improvements in both the understanding and performance of text access techniques. There have been some efforts to develop test collections for Japanese and this has resulted in a recent Call for Participation for IREX (Japanese Information Retrieval and Extraction Exercise, http://cs.nyu.edu/cs/projects/proteus/irex). IREX is organized by a committee of people from Japanese companies and universities, and is modeled on the TIPSTER and TREC programs. In addition, because TREC has made Chinese collections available, there have been a large number of recent papers on Chinese text retrieval.

Text-related research in Japan covers essentially the same areas as the United States, although there continues to be a strong emphasis on indexing techniques and speed. The differences that arose from the language-dependent aspects of Japanese text are rapidly disappearing.

Japanese companies appear to be focusing on developing the best commercial Asian language search systems for applications in Japanese, Chinese and Korean. There is, however, considerable competition even in this area in that considerable research and development of Chinese IR is underway in China, Singapore, Taiwan and Hong Kong, and Korea has a substantially longer history of IR research than Japan. One general criticism is that there seems to be too much reinvention of basic IR technology in Japan. Nearly every group visited was developing its own search engine (or engines). Licensing of U.S. search engines with Japanese
capability such as Verity or Infoseek is limited but may increase as it is demonstrated that search technology is essentially language-independent.

Current Japanese research and text search techniques do not offer significant benefits for English applications. The research is complementary to that being done in the United States, and the results tend to be incremental in nature. As the community of researchers in this area increases, however, we may expect to see more innovation and exploration of new ideas.

A number of groups in Japan are studying information visualization, architectures for scalable IR systems, and the application of natural language processing (NLP) techniques to IR. These are areas that could have a significant impact on the development of text-based systems. For example, the use of NLP techniques for IR has been studied in the United States for some time because of the obvious potential benefits of a system that “understands” the query better than a word-based system. Despite those potential benefits, research using quantitative evaluation based on test collections such as TREC has never demonstrated any retrieval effectiveness improvements from NLP. On the other hand, there is some evidence that language-based techniques may work better in Japanese than in English (Fujii 1997), and this may lead to a better understanding of text retrieval in general. Information visualization is another area where the opportunity exists for substantial innovation and synergy between Japanese research groups. An example of a visualization interface being developed and deployed by IBM Japan is shown in Figure 5.1.

In conclusion, the WTEC panelists’ view was that text-related research in Japan has been lagging behind that of the United States and Europe, but that substantial recent investments by companies and universities in this area mean that this gap is rapidly narrowing. One should expect to see substantially more new techniques and research directions originating in Japan in the near future.
REFERENCES


