TRW JAPANESE FAST DATA FINDER

Matt Mettler

TRW Systems Development Division
R2/2194
One Space Park
Redondo Beach, CA 90278
matt@wilbur.coyote.trw.com

ABSTRACT

The Japanese Fast Data Finder (JFDF) is a system to load electronic Japanese text, allow the user to enter a query in Japanese, and retrieve documents that match the query. Key terms in the browser display are highlighted. The interface is targeted for the non-native Japanese speaker and a variety of tools are provided to help formulate Japanese queries. The system uses TRW/Paracel Fast Data Finder hardware as the underlying search engine.

1. DESCRIPTION OF FINAL SYSTEM

1.1. Technical Approach

As part of the Tipster program, TRW has built an interface to the TRW Fast Data Finder (FDF) to search native Japanese scripts. The program is called the Japanese Fast Data Finder (JFDF). This effort involved working out the issues of converting and loading Japanese text into the FDF's filesystem, adapting the FDF to properly search 16-bit characters, developing an X-windows application program to work in Japanese, and researching query formulation techniques that will provide good precision and recall against Japanese texts.

We aimed our design toward providing special query generation support for non-native Japanese speakers. We developed an algorithm to convert English proper nouns or terms to multiple possible Katakana representations. We setup the user interface to provide easy access to lists of predefined subqueries and to an English to Japanese thesaurus. While the prototype was implemented to work with Japanese, we believe the JFDF could easily be extended to work with other languages such as Chinese and Korean. Data loaded for searching can be converted from multiple representations to a common Unicode representation if desired. Using Unicode allows a single database to contain documents in multiple languages. Query terms entered in Japanese could match against equivalent characters in Chinese or Korean documents.

The basic query formulation paradigm is extended boolean search. The user is provided with an expandable boolean form in which to enter query terms. Kana terms are entered by selecting the appropriate entry mode and typing Romaji. Kanji terms are entered by first entering the Hiragana and then asking for a menu of choices. Terms from the browser window may be copied and pasted into the query window to help refine a query. The user may select from a menu of choices to require query terms to be within a specified proximity window. A full set of features to browse, save, recall, print, and manipulate the results of searches are provided. There are also features to select the databases to search and to create and load new databases. Each screen has a help button which brings up a description of the available functions.

Since there are a number of different encoding schemes used for Japanese, the program currently runs using either Extended Unix Coding (EUC) or Unicode. Routines are provided to convert from one to another so that all documents in the database are represented in a common encoding format. It is easier to convert data once at load time than to convert (and search) each user query in multiple data representations.

Query Generation Features for the Non-Native Speaker

One of our primary considerations during the design of the JFDF was to make it easy to operate for the non-native Japanese speakers, which we believe make up the bulk of our prospective government users. We implemented three special features to aid query generation.

First, the JFDF includes a 28,000 entry English to Japanese thesaurus which can be used to help select appropriate terms for the search. The user types the term in English and receives a menu of possible Japanese terms and their English meanings. The user may select any combination of terms which are then included by reference in the query. The thesaurus is in a simple format that may be easily extended at customer sites.

Second, the user may select from a menu of predefined subqueries. These subqueries are typically setup by the system administrator or an advanced user. They may themselves include any valid FDF query expression including proximity, nested boolean logic, or error tolerance. By selecting the subquery 'Sony', the user could include in his top level query a number of different ways to reference Sony Corporation and its products.

Third, we implemented an English to Japanese transliteration scheme for proper nouns. This was a major undertaking. While the Japanese "spellings" for common foreign loan words or Western public figures tend to become quickly standardized (and thus could be included in the thesaurus), company names, new product names, and non-public figures are not likely to be represented in consistent Katakana across sources. Particular difficulties arise when the foreign words contain sounds or patterns of sounds (i.e. two consonants in a row) not used in Japanese. In these cases there are a number of different ways the foreign proper noun might be expressed in Katakana.

Our transliteration algorithm maps an English word to its most likely Katakana possibilities. The basic idea is to break the word apart phonetically and then substitute as many of the possible ways the sounds might be heard by a Japanese speaking person as alternatives. Figure 1 shows three simple examples[1].

```
ronald reagan  => [ロ|ロー|ル][ナ|ネー]ル[ド|ッド] [リ|リー|レ|レー][ガ|ゲー]ン
bill clinton   => [ビ|バ|バイ|ブ]ル ク[リ|ラ|ライ|ル]ントン
mary brown     => [マ|メー]リー ブ[ラウ|ロウ|ロー]ン
```

Figure 1: Sample English to Japanese Transliterations

We believe that this style of transliteration, with the FDF hardware available to execute the resultant expressions easily, is one of the JFDF's most effective and novel features. We reviewed the performance of this algorithm on a sample list of 150 English last names and tallied that the program was picking up the academically correct variation 80-90% of the time.

Designing the retrieval system around the Fast Data Finder is desirable for two reasons. First, the FDF can cost effectively evaluate complex query patterns required to achieve high recall and precision. In searching for Western proper nouns for example, we made heavy use of the FDF's ability to scan complex phonetically equivalent alternative Katakana representations of each word. Second, since the FDF can process data with no word segmentation, data preprocessing, or index construction, it is ideal for real-time dissemination systems. Thousands of detailed user profiles can be evaluated against large batches of documents within minutes of their arrival.

1.2. Processing flow

The JFDF can be divided into four modules that are interconnected as shown below.

1. JFDF - The main JFDF application / user interface program
2. Transliteration - The English to Katakana phonetic transliteration algorithm
3. MLT - The Multi-lingual toolkit from ILA.
4. FDF-3 - The Fast Data Finder system from Paracel (hardware + software)

Modules (1), (2), and (3) compile into a single executable program that is the JFDF. Module (4) is the hardware and control software for the Fast Data Finder. Modules (1) and (2) were developed on the JFDF contract. Modules (3) and (4) are commercial products. The user interacts with the displays of the JFDF main module. The JFDF main program makes client requested to the FDF-3 acting as a search server. The FDF-3 many be hosted on another workstation if desired.

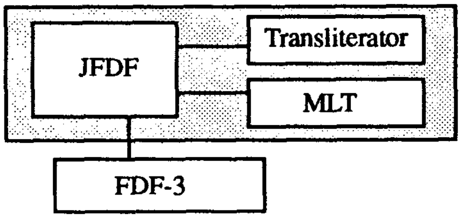

Figure 2: JFDF Modules

1.3. Description of Key Modules/Stages

The main JFDF module is the application program that provides a Japanese Language interface to the FDF. It uses the MLT functions to convert between Japanese text representations (i.e. EUC and Unicode), display Japanese on the screen, and accept entry of Kana or Kanji characters from the keyboard. The JFDFuses the FDF search engine to evaluate database documents against the user entered queries. The JFDF also provides a series of control options, help screens, and thesaurus lookups. The transliteration function is accessed when the user requests a term to be expanded by transliteration.

The transliteration module takes an English proper noun or phrase and converts it to a series of phonetically equivalent Katakana possibilities. The algorithm works by successive passes of string substitution on the input ASCII string. The substitutions are arranged in precedence order. The output is a single Katakana expression, which includes alternations to cover the variety of ways a particular syllable or letter might be pronounced, suitable for input to the FDF.

[1] The notation "a [b | c] d" means an "a" followed by a "b" or "c" followed by a "d". Thus the user entered term of 'Reagan' will match on any of

リガン or リーゲーン or レガン or レーガン

The Multi-Lingual Tooklit (MLT) contains a variety of tools and software routines to facilitate the development of multi-lingual applications. Specifically, the JFDF uses MLT functions to convert between different representations of Japanese text (i.e. EUC and Unicode), convert Romaji to Kana or Kanji, and render Japanese on the X-windows display.

The Fast Data Finder (FDF) is a hardware text search device developed by TRW and sold as a commercial product by Paracel Inc. in Pasadena California. It attaches to the SCSI bus of a host Unix workstation. It comes with several server or daemon processes that control the FDF hardware, manage the FDF's filesystem, and talk with client (user) processes. It also comes with a complete Application Programmers Interface (API).

1.4. Hardware/Software Requirements

The JFDF prototype is available to run on Sun Microsystems workstations running under Sun's Japanese Language Environment (JLE). The man-machine interface uses X-windows and Sun's Open Look / Xview. The text search operations are performed in hardware by the TRW/Paracel Fast Data Finder (FDF-3). The FDF-3 acts like a peripheral device to the host workstation and attaches to the host workstations SCSI device bus. The software is designed to the client-server model so that one FDF-3, located anywhere on the network, can provide search services to numerous user workstations running JFDF. The FDF-3 and associated FDF Executive Software may be purchased as a commercial product from Paracel Inc. of Pasadena California.

In addition to a Sun workstation running JLE and an FDF installed on the network, the JFDF uses a third party Multi-Lingual Toolkit (MLT) to provide character conversions, display, and appropriate fonts. This toolkit may be purchased as a commercial product from International Lisp Associates of Cambridge MA.

Finally, the JFDF utilizes a 28,000 entry public domain thesaurus, which can be supplied with the JFDF software from TRW.

1.5. Efficiency/Speed/Throughput Statistics

Speed and throughput of the JFDF system are a function of the FDF hardware search engine. During development of the JFDF we used an FDF2000 system that searched at 10 MB/s (5 million Kanji / second) and evaluated between 10 and 20 average queries simultaneously. Using a single commercial FDF-3 system, a search rate of around 3.5 MB/s (1.75 million Kanji / second) could be obtained while searching 20 to 40 average queries simultaneously.

1.6. Key Innovations of Final System

The JFDF prototype contains a number of novel features, including:

- Successful adaptation of a commercial high performance 8-bit hardware search engine to searching mixed 8-bit and 16-bit datastreams,
- Development of an automatic English to Katakana transliteration algorithm for proper nouns and phrases, and
- Adapting the man-machine interface to provide tools (such as an extensive English-Japanese thesaurus) to aid the formulation of Japanese queries by non-native Japanese speakers.

2. Original Project/System Goals

Primary goals for the Japanese Fast Data Finder (JFDF) were to build a system that could load electronic Japanese texts in various formats, allow the user to build queries in Japanese, search the queries against the text, and display the retrieved documents to the user with key terms highlighted.

To provide a scalable high performance architecture, the TRW/Paracel Fast Data Finder (FDF) was chosen as the search engine. The FDF also provides a wide variety of text search operators such as proximity searching, character alternation, term counting, term weighting, and fuzzy term matching.

To aide the non-native Japanese speaker in formulating queries, we also aimed at providing three query construction techniques:

- For common nouns, provide an English-Japanese Thesaurus to help select the proper terms,
- For proper nouns, provide an English to Katakana phonetic transliteration capability, and
- For specific topics, provide a menu of predefined subqueries for the user to include in queries as needed.

3. Evolution of System Over 6 Months

The JFDF prototype was developed over a six month period from Jan-93 to Jun-93. We delivered an early version to government evaluators at one of the sponsoring agencies in April-93. They suggested a number of improvements in the man-machine interface design and requested several additional features. We were able to accommodate some of these in the final version completed in June.

4. Accomplishments

Judging from the enthusiastic reception of the government evaluators, the JFDF prototype seems to be a success. It has been demonstrated to hundreds of people and we have received dozens of useful comments and suggestions on how its features might be enhanced or adapted to various operational scenarios. Some of the features that seem to have stimulated potential user interest include:

- Easy to use and expandable query entry form,
- The automatic English to Katakana transliteration for proper nouns,
- Incorporation of a comprehensive public domain thesaurus,
- Ability to copy terms from the browser window and paste them into the query,
- User-friendly X-windows interface,
- Ability to handle both Japanese and English/ASCII terms simultaneously in the same queries and databases, and
- Completion of the prototype on budget and on schedule.

5. Evaluation Summary

5.1. Official Results

Due to time constraints, government relevance assessments were only available for 3 of the 7 Japanese test topics. Nevertheless, the results seemed encouraging

Topic	Rel	Rel Ret	P@100
1	47	31	0.29
5	33	21	0.21

Table 1: Government Relevance Assessment for 3 Japanese Topics

For the three topics taken together the JFDF's recall was about 60% and precision about 20%. The 11pt average for these three topics was 0.1877.

5.2. Unofficial Results

The ARPA test topics did not evaluate some of the JFDF's novel features, especially the automatic transliteration. We performed a self-assessment on the transliteration algorithm by the following method.

- Select 150 western names beginning with different letters of the alphabet and containing various phonetic sounds,
- Run the algorithm to generate the possible Katakana transliterations,
- Have a native Japanese speaker spell out the most likely combination, and then
- Score 1 if the possible transliterations included the one the native speaker generated, otherwise 0.

From this test it appears that we were getting between 80-90% coverage. In other words, if the 150 names had been represented in the text as Katakana, we would have found over 120 of them. We did not perform this test on names already in the sample database, since we used the database during development of the algorithm.

5.3. Explanation/Interpretation of Results

From the point of view of correctly locating regions of text that matched the user's query in the test database, the JFDF performed flawlessly. We were also very pleased with achieving an accuracy of 80-90% on our rule-based transliteration algorithm. While the Japanese evaluation data is fairly sparse, the 0.1877 "11pt average" number for Japanese is a reasonable number and is comparable to what similar query formulation techniques might be expected to achieve in English.