MET Name Recognition with Japanese FASTUS*

Megumi Kameyama
Artificial Intelligence Center
SRI International
333 Ravenswood Ave., Menlo Park, CA 94025, U.S.A.
megumi@ai.sri.com

1 Introduction

SRI's Japanese FASTUS used in the Multilingual Entity Task (MET) evaluation is the initial Japanese system based on the FastSpec pattern specification language. We describe its system architecture, strengths, weaknesses, and its contribution to the prospects of a full information extraction system.

2 Japanese FASTUS: History

The first Japanese FASTUS was the MUC-5 Joint Venture system developed in 1993. Both the English and Japanese MUC-5 FASTUS systems used a graphical user interface called Grasper for rule definition, and recognized tie-up relationships among company organizations [1]. The performance of the Japanese FASTUS, together with other Japanese systems, demonstrated that the basic information extraction (IE) technology was portable to a language very different from English. The MUC-5 Japanese FASTUS gave us experience with 2-byte character input and JUMAN, a morphological analyzer developed at Kyoto University.

The second Japanese FASTUS, called MIMI (for "ears" in Japanese), summarized spontaneous human-human dialogues, and was developed during 1993-1995. MIMI was also Grasper-based, but its input was ASCII character "romaji" with spaces between words, and it had a 3,000-word dictionary in the domain of conference room scheduling [4, 5, 6].

During 1994-1995, the English FASTUS infrastructure underwent a number of changes, the most significant of which was the transition from Grasper to a declarative pattern specification language called FastSpec. FastSpec enables a fast cycle of rule specification, compilation, and testing during development [2].

*MET FASTUS was developed under SRI IR&D support.

3 MET Japanese FASTUS

The first implementation of the FastSpec-based Japanese FASTUS is the MET system . It was developed from scratch in a 4-staff-month effort on internal IR&D funding. In addition to the MUC-6 FASTUS infrastructure, past MUC-5 and MIMI experiences in general rule organization provided leverage. The MUC-5 experience in the use of JUMAN was also helpful.

New FASTUS developments in the MET system include new Japanese grammars in FastSpec, new JUMAN (version 2), customized JUMAN dictionary, 2-byte adaptation of FastSpec-based FASTUS infrastructure, and an SGML-handler phase specified in FastSpec.[1]

3.1 System Overview

FASTUS's basic architecture, shown in Figure 1, is unchanged [3]. The SGML-tagged input document is first tokenized. ASCII characters are sent to the ASCII Tokenizer, and 2-byte characters are sent to JUMAN. The ASCII Tokenizer is identical to the English FASTUS Tokenizer, which recognizes alphabetic, alphanumeric, numeric, and separator tokens as well as SGML tag tokens. JUMAN analyzes the input Japanese string into a single best sequence of morphemes with morphological attributes. These JUMAN morphemes are turned into FASTUS Lexical Item objects with slots for literal string, normalized string, lexical category, inflection type, and so forth.

The mixed sequence of ASCII and JUMAN tokens is then input into the SGML Handler, which recognizes the document structure based on SGML tags, and outputs a FASTUS Document object with slots for the headline, text, and other SGML fields. The headline slot has a sequence of sentences. The

[1] The last two were in collaboration with Mabry Tyson.

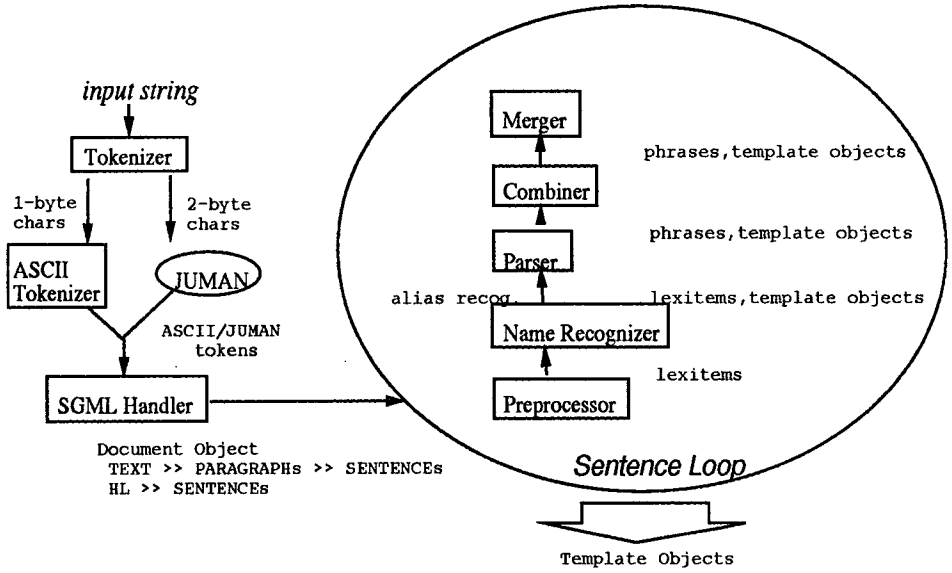

Figure 1: System Overview

text slot has a sequence of paragraphs, each of which contains a sequence of sentences. The SGML Handler is written in FastSpec, so it can be easily adapted to text tagging formats other than SGML, as well as to more complex text structures containing sections, subsections, and tables.

The Document object is input into the Sentence Loop consisting of a sequence of finite–state transducers, namely the Preprocessor, Name Recognizer, Parser, and Combiner. (The MET system did not have the last Domain Event phase that recognizes sentence patterns.) These linguistic phases recognize increasingly complex expressions in the sentence, recording syntactic and semantic attributes and producing template objects. At the end of each sentence loop, the Merger merges the new and existing template objects produced from the document so far. Document processing outputs a set of template objects that represent extracted information.

To recognize name strings for Organizations, Persons, Locations, Dates, Times, Money, and Percents, the MET Japanese FASTUS produces a template entity for each. The name slot of a template entity has a name string value with its start and end positions in the document. Name tagging in the output uses these text position values. Most of the names are recognized in the Name Recognizer phase based on internal patterns. After the Name Recognizer phase, the Alias Recognition routine recognizes some of the unknown words as aliases of the organization names recognized earlier in the same document. The Parser and Combiner phases recognize a name's surrounding linguistic contexts, sometimes converting a phrase of one type into a phrase of another type.

3.2 Strengths and Weaknesses

The system's strengths derive from the FastSpec–based FASTUS infrastructure, and the weaknesses are problems in Japanese name recognition that any system must cope with.

3.2.1 Strengths

The following are the main strengths of the system:

FastSpec enables transparent rule definition of a complex finite–state transducer. The optimizing compiler constructs an efficient finite–state machine, allowing a rapid specify–compile–test cycle.

Name recognition is completely integrated in template entity extraction, so the system is ready for further incremental augmentation toward a full–scale IE system.

Named entities can be recognized based on linguistic contexts in complex phrase patterns. For example, in "zidousya seizou gaisya no papiyon" (Papillon, an automaker), the word "papiyon" (Papillon) may be unknown, but the immediate linguistic context makes it a company name.

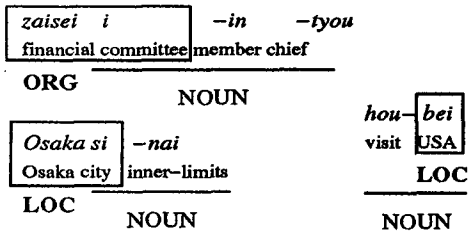

Figure 2: Name–Noun Overlap Examples

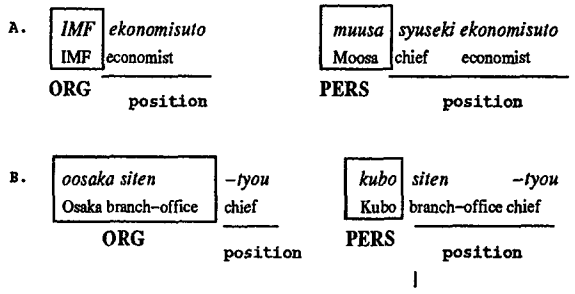

Figure 3: Ambiguous Context Examples

3.2.2 Weaknesses

IE–Customization of Dictionary. FASTUS uses dictionary entries as the smallest linguistic units that are combined to create more complex patterns. There were numerous cases in MET, however, where dictionary entries cut across name boundaries. Examples are shown in Figure 2.

These common nouns are complex morphemes, parts of which can simultaneously belong to organization or location names. These examples indicate the fact that IE requires substantial sublexical analysis in Japanese. There are essentially two methods for coping with this. One method, taken by the NTT DATA system in MET,[2] is to first tokenize with normal dictionary entries, and then later to extract sublexical parts during IE. This approach enables a single dictionary to be used for IE and non–IE purposes. The other method, taken by the SRI system in MET, is to remove these complex morphemes from the dictionary, and combine sublexical items with rules. This approach makes IE dictionaries diverge from the off–the–shelf ones.

Rule–Dictionary Trade–offs. Organization, Person, and Location names comprise a majority of the names to be recognized, and a special difficulty arises when they occur in similar linguistic contexts. A prime example of such overlapping contexts is the positions held by persons within organizations, as shown in Figure 3.

In Example A, the same "economist" position type acts as the context for Organization or Person names. In Example B, the same "branch–office chief" position phrase provides part of an organization name in one example, while it acts as a context for a person name in the other.

The difficulty is that even known organization, person, and location names are often ambiguous. For instance, "murayama" (Murayama) can be a person's last name or a city name, and "foodo" (Ford) can be a person's last name or company name. Disambiguation relies on the linguistic context. Moreover, most organization, person, and location names are unknown to the system. Their recognition relies on both internal name patterns and linguistic contexts. The system must hit the right balance between the size of the dictionary of known names and the complexity of the name–context patterns. When the list is too large, disambiguation requires almost as much effort as if there were no list; but certain names elude predictable internal name patterns, so need to be known a priori. This rule–dictionary trade–off must be fully explored to increase name recognition accuracy.

[2] Yoshio Eriguchi and Tsuyoshi Kitani, personal communication

4 Future

The MET Japanese FASTUS is ready for further development and augmentation toward a full information extraction system. We plan to fully customize the dictionary for IE purposes, and augment the system with coreference resolution and compile–time transformation capabilities demonstrated in the English MUC–6 FASTUS [2].

We also plan to make this Japanese IE system accessible to English–speaking analysts. This will be possible by combining the IE technology with suitably constrained applications of machine translation technology.

References

[1] Appelt, Douglas, Jerry Hobbs, John Bear, David Israel, Megumi Kameyama, and Mabry Tyson. SRI: Description of the JV-FASTUS System Used for MUC-5. In Sundheim, Beth, ed., the *Proc. of the 5th Message Understanding Conference*, ARPA, 1993b.

[2] Appelt, Douglas, Jerry Hobbs, John Bear, David Israel, Megumi Kameyama, Andy Kehler, David Martin, Karen Myers, and Mabry Tyson. 1995. SRI Inter-

national FASTUS System MUC–6 Test Results and Analysis. In the *Proc. of the 6th Message Understanding Conference*, ARPA.

[3] Hobbs, Jerry, Douglas Appelt, John Bear, David Israel, Megumi Kameyama, Mark Stickel, and Mabry Tyson, 1996. FASTUS: A Cascaded Finite–State Transducer for Extracting Information from Natural–Language Text. In E. Roche and Y. Schabes, eds., *Finite State Devices for Natural Language Processing*, MIT Press, Cambridge, Massachusetts.

[4] Kameyama, Megumi and Isao Arima, 1993. A Minimalist Approach to Information Extraction from Spoken Dialogues. In *Proc. International Symposium on Spoken Dialogue (ISSD–93)*, Waseda University, Tokyo, Japan, 137–140.

[5] Kameyama, Megumi and Isao Arima, 1994. Coping with Aboutness Complexity in Information Extraction from Spoken Dialogues. In *Proc. International Conference on Spoken Language Processing (ICSLP–94)*, Yokohama, Japan, 87–90.

[6] Kameyama, Megumi, in preparation. Information Extraction from Spontaneous Spoken Dialogues. Manuscript. SRI International Artificial Intelligence Center.