Oki Electric Industry : Description of the Oki System as Used for MUC-7

J. Fukumoto, F. Masui, M. Shimohata, M. Sasaki
Kansai Lab., R&D group
Oki Electric Industry Co., Ltd.
Crystal Tower 1-2-27 Shiromi, Chuo-ku, Osaka 540-6025 JAPAN
{fukumoto,masui,simohata,sasaki}@kansai.oki.co.jp

INTRODUCTION

This paper describes the Oki Information Extraction system as used for MUC-7 evaluation [1][2]. The tasks we have conducted are Named Entity, Co-reference, Template Element and Template Relation. Each module is implemented using MT system modules and pattern recognition modules. Our purposes to participate MUC-7 evaluation are to evaluate how MT system modules are effective for other application such as IE system and to develop our information extraction technology based on pattern recognition.

Oki's MT system, PENSÉE [3][4], is a commercial system and is one of major MT systems in Japan. Translation mechanism of PENSÉE system basically applies transfer method. There are English-to-Japanese (EJ) and Japanese-to-English (JE) systems and both are based on the same software systems, that is, both MT systems are using the same Grammar Description Language (GDL), and GDL rule translator and interpreter (GDL system). Oki has already spent more than ten years for development and improvement of the MT system.

The Oki IE system used for MUC-7 is composed of a surface pattern recognition module and a structural pattern recognition module. The surface pattern recognition module traces a text at surface linguistic level and detects NE elements and co-referred elements without any language analysis system such as lexical analysis and syntax analysis. The structural pattern recognition module traces parse trees of a text, which is generated by parser of MT system. Syntactic and semantic information embedded in the parse tree are used to detect NE elements, co-referred elements and so on. The structural pattern recognition rules are described in GDL and are executed on the GDL system which is based on pattern matching mechanism of tree data. Detected elements of tree data are marked and are extracted after execution of rules.

BACKGROUND

Oki has submitted two systems: the English system for MUC-7 (NE, CO, TE and TR) and the Japanese system for MET-2 (NE). This is our first participation in MUC and MET evaluation. In order to develop the systems in a short period, we utilized parsing module of the MT system

for a sentence analyzer. The English system is developed using the English sentence analyzer of the PENSÉE-EJ and the Japanese system is developed using the Japanese sentence analyzer of PENSÉE-JE. We also utilized the GDL system for description of extraction rules (only for description of structural patterns). The GDL system has powerful rule debugging environment, which is very helpful to develop an extraction system in a short period.

OKI MT SYSTEM : PENSÉE

In the MT system, PENSÉE, translation rules are executed on the GDL system. A GDL rule consists of pattern matching part and action part. Pattern matching part describes conditions to specify a part of tree data and action part states changes of the specified tree structure and/or modification of node information of the tree. Sample transformation of tree data is shown in Figure 1.

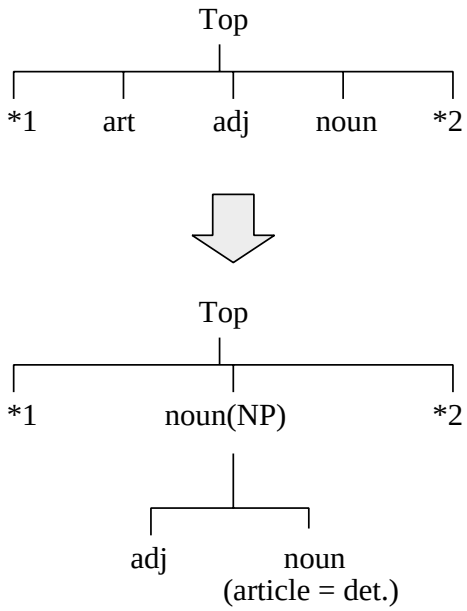

Figure 1: Transformation of tree data in the GDL system

Pattern matching specifies a sequence of article (art), adjective (adj) and noun and both sides of the sequence are arbitrary number of nodes ("*1" and "*2"). After transformation, node information of article is embedded into the noun node, and the noun and adjective nodes are moved under the noun phrase node (NP). In our current implementation, a lower node modified its upper node.

OVERVIEW OF THE OKI IE SYSTEM

Oki's IE system consists of surface pattern recognition modules, structural pattern recognition modules, filtering programs to convert internal expression of parser to surface expression, and

SGML tag processing modules. Architecture of the system is shown in Figure 2.

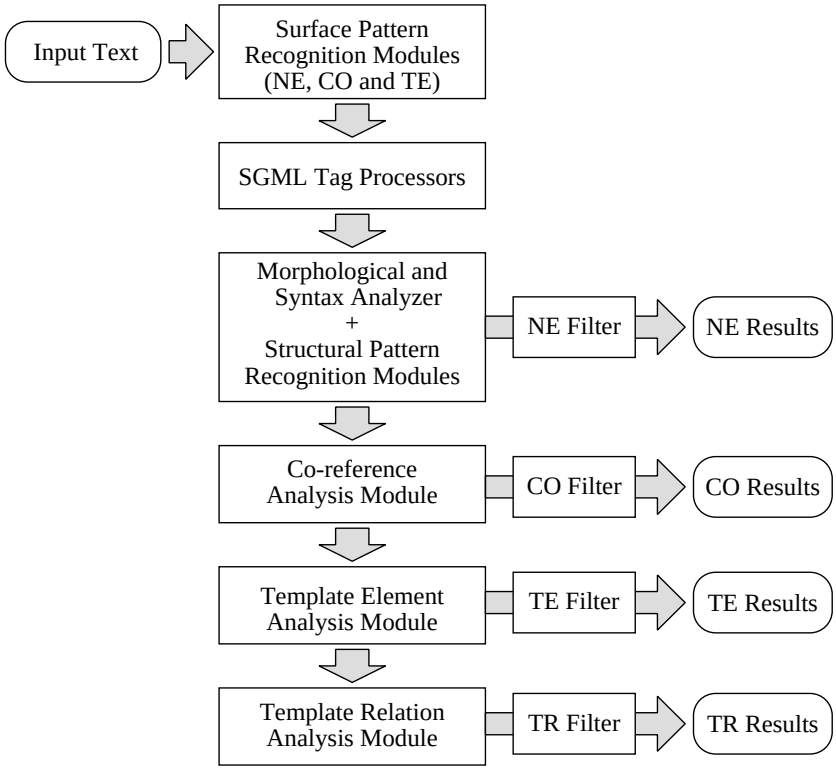

Figure 2: Architecture of the Oki MUC system

The system firstly recognizes surface level patterns in a text and adds SGML tags for each task. In this analysis, NE elements, CO elements and sub-category of NE elements for the TE task are extracted. In the SGML tag processing module, these tags and original SGML tags are embedded in their adjacent words for morphological and syntax analysis. Each sentence of the tag-processed text is parsed in the morphological and syntax analyzer which is originally used in the MT system. Structural pattern analysis rules for the NE task are implemented in the syntax analysis rules. The recognized patterns are expressed in node attributes in a parse tree. In order to obtain NE results, information of node attributes is extracted and embedded in a text as SGML tags by the NE filter.

For the CO task, all parse trees of sentences are converted into one tree structure each child node of which is a parse tree of a sentence. In recognition of the CO task, anaphoric elements are extracted and their antecedents are detected by traversing the tree structure. Information of CO elements is extracted from the tree structure and embedded in the original text as SGML tags by the CO filter. CO-reference numbers are also set by the CO filter. For the TE and TR tasks, a text tree structure is used for identifying TE and TR elements. The TE and TR filters generate TE and TR information in BNF format from the tree structure of a text.

The NE system

The NE system consists of a surface pattern recognition module, a structural pattern recognition module and some filtering programs for tag-processing for parsing and post-processing.

Surface Pattern Recognition

Surface pattern recognition is processed in the following sub-processes.

- Recognition of capitalized area

  Sequential capitalized words are recognized as NE candidates.

- Head word processing

  A head word of a sentence is removed from NE candidates when it is registered in Non-element word list.

- Merging capitalized area

  Some functional words and prepositions, "for" and "of", are utilized for recognition of NE candidates. For example, the functional word "Bank" and the preposition "of" are used for recognition of the NE element, "Bank of Tokyo". The functional word "University" and the preposition "of" are used for recognition of the NE element, "University of Tokyo".

- Type recognition of NE elements

  Type information of NE candidates is recognized by functional words such as "Mr." of a person name, "Bank" of a organization name, "City" of a location name and so on.

- Dictionary look up

  The rests of NE candidates are checked with word list of each NE type. The lists are manually extracted from newspaper articles, index information of a world map and company names from stock market news, etc.

In the recognition of NE elements, identified elements are utilized for recognition of their abbreviation which is repeated in a text. For example, when "Mr. John Doe" in a text is recognized as a person name, the words "John" and "Doe" in the same text are also recognized as a person name. Moreover, the abbreviation "FAA" in a text is used for checking NE candidates after recognition of "Federal Aviation Administration" in the same text.

Tag-processing for Parsing

An original text is SGML-tagged one and NE elements in the text are also SGML-tagged after the surface pattern recognition. In order to parse such a SGML-tagged text, these tags have to be concealed. In a parse tree, tag information is expressed in node attributes of a parse tree, therefore, the system can handles information obtained at a surface level pattern recognition during parsing.

Structural Pattern Recognition

After parsing, several structural pattern rules are applied to a parse tree in order to recognize NE elements. Some of them are as follows:

- The subject element of some types of verb such as "say", "die", "play" and so on is recognized as a person name.
- The noun phrase in front of relative pronoun "who" is recognized as a person name.
- The noun phrase, whose appositive phrase is a person name, is recognized as a person name, vice versa.
- The noun phrase followed by "employee", "spokesman" and so on is recognized as an organization name.
- The noun phrase with preposition "in", "at", "near" or "over", whose appositive phrase is an organization name, is recognized as an organization name.

Post-processing

After the structural pattern recognition, NE tag information is extracted from a parse tree and added to the pattern tagged text. All the NE tag information is utilized for tagging to non-body part of the text.

The CO System

The CO system also consists of a surface pattern recognition module, a structural pattern recognition module and some filtering programs for tag-processing for parsing and post-processing.

Surface Level Recognition

In the surface pattern recognition of the NE system, abbreviation and repeated one are utilized for recognition of NE elements. This mechanism is also used for surface level recognition of CO elements. For example, when the person name "Mr. John Doe" appears in a text, the words "John" and "Doe" are utilized as abbreviation of the person name.

Tag-processing for Parsing

SGML-tags and CO tags obtained from the surface pattern analysis are concealed for parsing as well as the NE system.

Structural Level Recognition

This is the main module of CO-reference analysis. Firstly, the CO system recognizes co-reference expression of appositions and the expression of "A is B". Then, the CO system extracts anaphoric expressions of pronouns and noun phrases with definite articles, and traverses the tree structure from bottom to top and left to right way to detect its antecedent.

Post-processing

Information of CO tags is extracted from the tree structure of a text and is added to the pattern tagged text. All the CO tag information is utilized for tagging to non-body part of a text in the same way of the NE system. Moreover, all the co-referred elements have the same co-reference number, therefore, renumbering according to the CO task definition is also done in this module.

The TE System

The TE system consists of the following sub-processes.

1. Recognition of Entities

   Entities which are recognized in the NE system are selected. Capitalized elements which are not identified in the NE system will be handled as candidates for TE elements. These elements will be identified as TE elements, if they are related to some descriptor.

2. Recognition of Descriptors

   Noun phrases and prepositional phrases which have functional words of descriptors such as "Comdr.", "President", "agency" and so on are extracted. Semantic information of Descriptors which is used in parsing will be utilized for recognition of type information of TE entities in the next process.

3. Recognition Relation between Entities and Descriptor

   Relation between entities and descriptors is recognized by structural pattern rules for the TE task. For example, an entity and a descriptor are related on some verb such as "called", "named" and "be", they are recognized as a TE pattern.

4. Merging TE Patterns

   Some TE patterns are merged using Co-reference information, that is, the same entities (co-referred elements) are recognized as one entity.

The TR System

The TR system identifies relation between TE elements. In our current simple implementation, if a person and an organization are related on some predicate, they are recognized as having "employee-of" relation. Moreover, if an organization and a location are related on some predicate, they are recognized as having "locate-of" relation. In case of an artifact and an organization, they will have "product-of" relation.

WORKTHROUGH ARTICLE

Table shows results of the workthrough article. As for the TE task, we mainly implemented rules for `entity`, therefore, the TE score for `entity` is at an average level according to the results of

MUC-6. However, the score for `location` is very low. It is because the TE system has a few location rules and has no dictionaries to identify the detailed information of a location.

Task	Recall		Precision		F-measure
NE (nyt9602140704)	94/132	71.1	94/114	82.5	76.4
TE (nyt9602140509)	52/186	30.0	52/62	83.9	41.9
TR (nyt9602140509)	45/168	26.8	45/62	72.6	39.1
CO (nyt9609100378)	26/79	32.9	26/47	55.3	41.3

Table 1: Results of Workthrough Article

CONCLUSION AND FUTURE DIRECTION

This participation to the MUC-7 and MET-2 tasks was our first trial and we had to develop our IE systems for the tasks in a short period. Our purpose to participate MUC-7 tasks is to evaluate how MT parsing modules are effective for development of an IE system in a short period and how they work. We found that it is useful to utilize the MT system modules for the other application such as IE system because our MT system has practically achieved robustness and the rule description system called the GDL system is also very helpful due to its pattern matching mechanism. However, parsing modules of the MT system has been originally developed for language transformation in a transfer method, therefore, tree structure of an original language is sometimes converted to the structure of a target language. These kinds of rules caused some difficulty of pattern matching in the IE tasks.

We have participated the MET-2 task in Japanese which has only the NE task. We are planning to apply our IE technology developed for MUC-7 tasks to a practical Japanese information extraction system. Moreover, it will be useful to participate Japanese CO, TE, TR and ST tasks if they will be defined in the next conference.

References

[1] TIPSTER TEXT PROGRAM Phrase II, *DARPA*, (1996).

[2] Proceedings of 6th Message Understanding Conference (MUC-6), *DARPA*, (1995).

[3] Masui, F., Tsunashima, T., Sugio, T., Tazoe, T. and Shiino, T.: "Analysis of Lengthy Sentences Using an English Comparative Structure Model", *System and Computers in Japan*, pp.40–48, SCRIPTA TECHNICA Inc., (1996).

[4] PENSÉE, http://www.oki.co.jp/OKI/RDG/English/kikaku/vol.1/sugio/main.html http://www.oki.co.jp/OKI/Home/English/Topic/PENSEE/