Oki Electric Industry : Description of the Oki System as Used for MET-2

J. Fukumoto, M. Shimohata, F. Masui, 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,simohata,masui,sasaki}@kansai.oki.co.jp

INTRODUCTION

This paper describes the Oki Information Extraction system as used for MET-2 evaluation (Japanese task) [1][2]. System architecture of Oki MET system is basically the same as the one for the NE system for MUC-7 evaluation although recognition rules of both systems are different because of language difference (Japanese and English).

Oki MET system [3][4] is based on term recognition rules of surface linguistic expressions and parse trees which is generated by the parsing module of MT system. The MET system firstly identifies term boundaries by Japanese character types. Named Entities are recognized using a Japanese suffix list and a word list for each NE type. After recognition of surface patterns, each sentence of a text is parsed and NE elements are recognized by structural pattern rules. The structural pattern recognition rules are described in GDL and are executed on the GDL system.

SYSTEM DESCRIPTION

Oki MET system consists of a surface pattern recognition module, a SGML tag processing module, a structural pattern recognition module and filtering program to convert internal expression of parser to surface expression. Architecture of the system is shown in Figure 1.

In the surface pattern recognition module, the system firstly identifies term boundaries by Japanese character types such as Hiragana, Kanji, Number and so on. Then, NE elements are recognized using a Japanese suffix list and a word list for each NE type. The detected NE elements in a text are SGML-tagged. In the SGML tag processor, these tags and original SGML tags in a text are embedded in their adjacent words for morphological and syntax analysis. Each tag-processed sentence is parsed by the morphological and syntax analyzer which is originally used in the MT system. After parsing, structural patterns are recognized by structural pattern recognition rules, although the number of rules is not much in the current implementation. These rules are described in GDL and are executed on the GDL system after parsing. The recognized patterns are expressed in node attributes of a parse tree. In order to obtain NE results, node information of NE tags is extracted and embedded in a text as SGML tags by the NE filter.

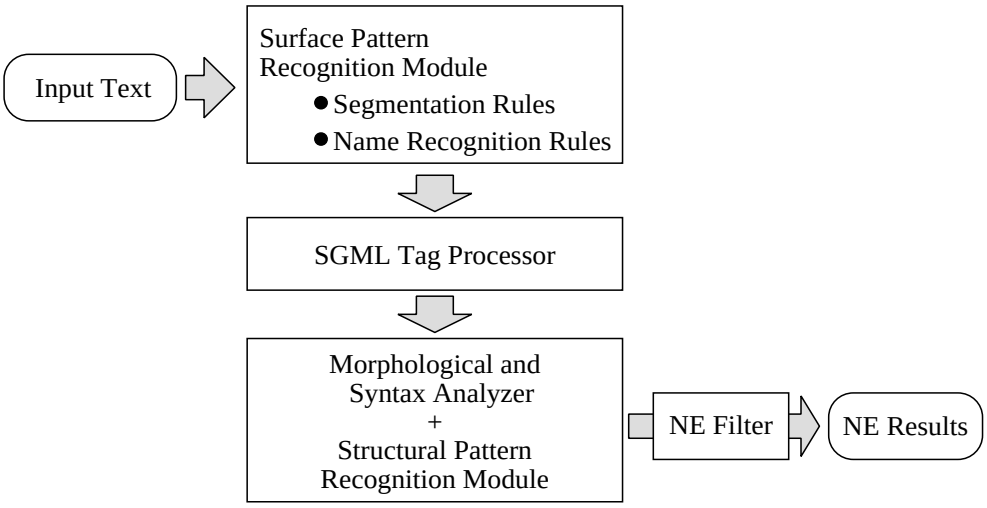

Figure 1: Overview of the Oki MET system

RECOGNITION OF NE ELEMENTS

Segmentation

At surface level processing, term boundaries of an input text is identified by Japanese character types which are Hiragana and some symbols such as brackets and Japanese punctuation. The system firstly extracts a sequence of Japanese characters, which are not Hiragana characters, Japanese comma "、" and Japanese period "。", from a text. Some sequences of Hiragana characters registered in a list of functional Hiragana words and a list of proper nouns are merged with their adjacent terms.

In the following example, ≪原山行雄さん ≫ is extracted by merging the Kanji sequence ≪原山行雄≫ and the Hiragana sequence ≪さん≫ which is registered in a list of functional Hiragana words.

> ≪日本人観光客１７人≫の乗った≪マイクロバス≫が≪トラック≫と≪衝突≫、≪静岡県沼津市≫の≪原山行雄さん≫（≪５９≫）が≪死亡≫、≪５人≫が≪重傷≫、≪３人≫が≪軽傷≫を負った。[1][2]

As for a list of proper nouns, ≪東京・霞が関≫ is generated by merging the divided terms, ≪東京・霞≫, が and ≪関≫ by information of the location name ≪霞が関≫ registered in a list of proper nouns. The term ≪和歌山県伊都郡かつらぎ町中飯降≫ is also recognized by the location noun ≪かつらぎ町≫.

[1] Extracted terms are marked by brackets "≪" and "≫".

[2] One Kanji character will be handled by some heuristics which are discussed later.

Name Recognition

In order to identify scope and type of an NE element, segmented terms are divided into basic terms according to term lists as shown in Table .

Name of list	number of elements	Examples
person name list (pn)	58	高橋, 斉藤
organization name list (on)	255	住友, 三菱
location name list (ln)	10083	大阪, 山口, ワシントン
person name suffix list (ps)	90	さん, 被告, 社長
organization name suffix list (os)	155	社, 株式会社
location name suffix list (ls)	66	市, 町, 村
organization name modifier list (onm)	124	銀行, 電気, 産業, 工業
stop word list of person name (swp)	48	農業, 大学生
stop word list of organization name (swo)	37	大手, 主要

Table 1: Term Lists

The divided basic terms are merged by concatenation rules and then NE elements and their type such as a person, an organization and a location is recognized. Figure 2 shows concatenation rules which are applied in this order.

(1)	$pn + ps$	$\rightarrow$	Person name (p) + ps
(2)	$on + os$	$\rightarrow$	Organization name (o)
(3)	$ln + ls$	$\rightarrow$	Location name (l)
(4)	$ln + ps$	$\rightarrow$	Person name (p) + ps
(5)	$pn + onm^*$	$\rightarrow$	on
(6)	$on + onm^*$	$\rightarrow$	on
(7)	$ln + onm^*$	$\rightarrow$	on
(8)	on	$\rightarrow$	Organization name (o)
(9)	ln	$\rightarrow$	Location name (l)

Figure 2: Sample Concatenation Rules

According to the rules (1), (2) and (3), each type of an NE element is identified by a sequence of a basic name and a suffix. A sequence of a location name and a person name suffix is recognized as a person name by rule (4) because a location name is often used as a person name. Rule (5), (6) and (7) show that a person / organization / location name and a sequence of an organization name modifier are important cues to identify an organization name. Rules (8) describes that an organization name is identified as an NE element when it has no suffix information. Rules (9) describes the case of a location name. Stop word lists are used to remove an element from candidates for NE elements. When a candidate for NE elements contains an element of a stop word list, it might be a general noun, not a proper noun.

Heuristics of Name Recognition

After recognition of NE elements at surface level, some heuristics are applied to each sentence to identify more NE elements. Some of them are as follows:

- The words "東", "西", "南" and "北" are merged with a location name.
- If Japanese characters "月" and "日" are independent element after segmentation, they are recognized as a location name.
- If one character country names such as "日", "米" and "英" appear in a sequence, they are recognized as a location name.
- A person name without a person name suffix is not recognized as an NE element because most of person names accompany with a suffix.

Table 2 shows examples of Name recognition from divided form. Each tag name in this table indicates that the tagged element is an element of Name of list presented in Table . For example, xxx of the tagged element "< *pn* xxx>" shows is an element of a person name list.

Original form	Divided form	Results
静岡県沼津市	< *ln* 静岡>< *ls* 県>< *ln* 沼津>< *ls* 市>	< *l*静岡県>< *l*沼津市>
大阪府出身	< *ln* 大阪>< *ls* 府>出身	< *l*大阪府>出身
北京	< *ln* 北京>	< *l*北京>
広瀬さん	< *pn* 広瀬>< *ps* さん>	< *p* 広瀬>< *ps* さん>
越田典子さん	越田典子< *ps* さん>	< *p* 越田典子>< *ps* さん>
原山行雄さん	< *ln* 原山>行雄< *ps* さん>	< *p* 原山行雄>< *ps* さん>
日本大使館	< *ln* 日本>< *os* 大使館>	< *o* 日本大使館>
セイコー電子工業	< *on* セイコー>< *onm* 電子>< *onm* 工業>	< *o* セイコー電子工業>
欧州連合	< *ln* 欧州>< *onm* 連合>	< *o* 欧州連合>
花巻空港	< *ln* 花巻>< *os* 空港>	< *o* 花巻空港>
交通事故	< *onm* 交通>< *onm* 事故>	交通事故
交通安全委員会	< *onm* 交通>< *onm* 安全>< *os* 委員会>	< *o* 交通安全委員会>
米道路交通安全局	米< *ls* 道路>< *onm* 交通>< *onm* 安全>< *os* 局>	< *o* 米道路交通安全局>

Table 2: Term Lists

Tag-processing for Parsing

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

Name Recognition from Parse Tree

After parsing[3], structural pattern rules are applied to a parse tree in order to recognize NE elements. Semantic information utilized in parsing rules is used to correct the scope of NE elements. For example, if there is an unknown word between a person name and a word which has a meaning of human, the sequence of words are recognized as a person name. In the following sentence, "中村" is recognized as a person name in surface name recognition and "副操縦士" has a meaning of human in a parse tree level, then "中村貴洋副操縦士" is recognized as a person name.

中村貴洋副操縦士（３０）を起訴猶予処分にした。

Post-processing

After structural pattern recognition, information of the NE tags is extracted from the parse tree and added to the pattern tagged text.

DISCUSSION AND FUTURE DIRECTIONS

We took the same kind of approach to develop the Oki MET-2 system as the NE system for MUC-7 although target languages are different. That is, both systems consist of a surface linguistic pattern recognition module and a structural pattern recognition module using parsing module of MT system. However, performance of the Japanese NE system is higher than that of the English NE system. It is because of the difference of uniformity of expressions in texts. Most of the person names in Japanese newspaper articles have a term of respect or title. On the other hand, in English ones, the first NE element has a term of respect or title but its repeated ones are represented by the name only.

The parsing module of the Japanese-English MT system has been originally developed for language transformation, therefore, a part of tree structure of the original language is sometimes converted to structure of the target language during parsing. That is, some parts of tree structure express sentence structure of the target language and others express sentence structure of the original language. This caused some problems of pattern matching of Name recognition on a parse tree.

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 good 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).

[3] In order to parse a SGML-tagged text, tag information has to be concealed like the Tag-processing of the MUC-7 NE system.

[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/