NEC Corporation and University of Sheffield: "Description of NEC/Sheffield System Used For MET Japanese"

Yoshikazu Takemoto‡ Takahiro Wakao†
Hiroshi Yamada‡ Robert Gaizauskas† Yorick Wilks†
University of Sheffield Computer Science Department†
NEC Corporation, Information Technology Research Laboratories‡
{*takemoto, wakao, h-yamada*}*@hum.cl.nec.co.jp*
{*R.Gaizauskas, Y.Wilks*}*@dcs.shef.ac.uk*

1 Introduction

Recognition of proper nouns in Japanese text has been studied as a part of the more general problem of morphological analysis in Japanese text processing ([1] [2]). It has also been studied in the framework of Japanese information extraction ([3]) in recent years.

Our approach to the Multi-lingual Evaluation Task (MET) for Japanese text is to consider the given task as a morphological analysis problem in Japanese. Our morphological analyzer has done all the necessary work for the recognition and classification of proper names, numerical and temporal expressions, i.e. Named Entity (NE) items in the Japanese text.

The analyzer is called "Amorph". Amorph recognizes NE items in two stages: dictionary lookup and rule application. First, it uses several kinds of dictionaries to segment and tag Japanese character strings. Second, based on the information resulting from the dictionary lookup stage, a set of rules is applied to the segmented strings in order to identify NE items. When a segment is found to be an NE item, this information is added to the segment and it is used to generate the final output[1].

2 System Description

2.1 System Overview

The main processes in the Amorph system are as follows:

1. A basic morphological analysis is carried out on the input text using a simple dictionary. Part-of-speech information is added to the text. Since proper noun is not one of the parts of speech, all the nouns are tagged simply as noun.
2. Three proper noun dictionaries (organization, person, location) are consulted for each noun in the text. If there is an exact match between a noun in the text and an item in one of these dictionaries, the noun is tagged as such.
3. A "co-occurrence" (kyoukigo) dictionary which is a collection of key words is consulted. These key words are words which indicate that the segment is a part of a Named Entity item, or the segment is immediately before or after an NE item. When such a key word is found, it is marked as such.
4. A set of rules for each type of NE items (organization, person, location, date, time, money, percent) is applied to the result so far. Some organization names are recorded temporarily in the system to identify the names which appear somewhere else in the text (name learning). If a segment (or segments) is judged to be an NE item by some rule or name learning, then it is marked so.
5. Finally the output generation module produces the final output, marking NE items in the text with the specified SGML markup.

The use of dictionaries (or lists) of proper names and key words, and the application of rules based on the dictionary lookup have also been adopted in the Sheffield English MUC-6 system ([4]).

Figure 1 shows the overview of the Amorph system illustrating how the input text is processed.

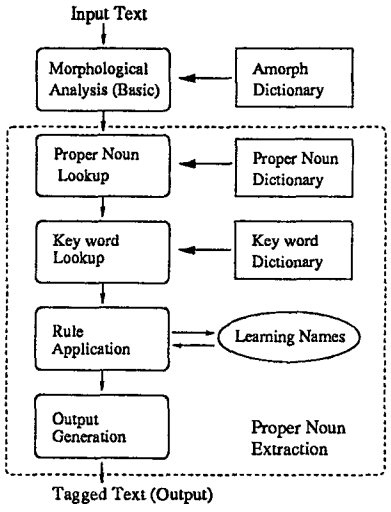

Figure 1: Amorph System Overview

[1] Wakao is a visiting researcher at NEC Information Technology Research Labs in 1996

3 Dictionaries

The Amorph system uses three proper noun dictionaries and a key word dictionary. Table 1 shows the size of the proper noun dictionaries.

Type	Size
Location	10438
Organization	2544
Person	8223
Total	21205

Table 1: Size of Proper Noun Dictionaries

The key word dictionary has 785 items in total. For example, "市" (city) is a key word for a location name, "大統領" (President) for a person name, and "社" (Company) for an organization name. It has 61 key words for location, 324 for organization, 351 for person, and 49 for number and time expressions.

Table 2 shows the sizes of key words for different types of the NE items.

Type	Size
Location	61
Organization	324
Person	351
Number and Time	49

Table 2: Details of Key Word Dictionary

4 Extraction Rules

A set of rules is applied to the text after the dictionary lookup is done. These rules have been produced manually in order to capture different types of proper names as well as number and temporal expressions in the text.

The Japanese language has four different kinds of characters: Hiragana, Katakana, Kanji (Chinese characters), and alphabetic characters, besides numerical and punctuation symbols. Some of the rules below use this feature, i.e. what kind of characters the given string is made of.

Here are some rules for person and organization names.

1. if the segment is positioned immediately before or after a person key word, and it is marked as either location or ambiguous between location and person name, then it is recognized as person.

   e.g. "千葉氏": "千葉" (Chiba) is ambiguous between location and person name, but since it is followed by a person key word, "氏" (Mr. in this case), "千葉" is recognized as person.

2. the segments are a Kanji-, Katakana-, or alphabet-character-only string and followed by a specific expression such as " (本社" (headquarters), they are judged to be an organization name.

   e.g. "ＮＥＣ (本社東京" (NEC, headquarters in Tokyo): "ＮＥＣ" is an alphabet-character-only string and recognized as an organization name.

5 Name learning

When a segment (or a set of segments) is recognized as organization name and at the same time it is an alphabet-character-only or katakana-character-only string, the system records it temporarily and use it to recognize appearances of the name in other places in the text.

For example, a katakana-character-only string, メルセデス・ベンツ (Mercedes Benz) is once recognized as organization name, both segments (メルセデス and ベンツ) are remembered as organization name to capture later appearances of the name (メルセデス in this case) in the text.

This mechanism of learning names is useful for recognizing organization names in the headline since organization names tend to be in shortened forms in the headline and their full names appear in the body of the text.

6 Acknowledgments

We would like to thank Mr. Miyabe of NEC for his help in data analysis. We also thank Mr. Muraki of NEC for encouraging us to carry out the task.

References

[1] Miyazaki, M., "Automatic Segmentation Method for Compound Words Using Semantic Dependent Relationships between Words" (「係り受け解析を用いた複合名詞の自動分割法」), *Jyouhou Shori Gakkai Ronbunshi*, Vol 25 (6), 1984. (In Japanese)

[2] Kitani, T. and T. Mitamura, "An Accurate Morphological Analysis and Proper Name Identification for Japanese Texts Processing", *Journal of Information Processing Society of Japan*, Vol 35 (3), 1994.

[3] Muraki K, S. Doi, and S. Ando, "NEC : Description Of the Veniex System As Used For MUC-5", In *Proceedings of the Fifth Massage Understanding Conference*, Morgan Kaufmann Publishers, 1993.

[4] Gaizauskas R., T. Wakao, K. Humphreys, H. Cunningham, and Y.Wilks, "University of Sheffield : Description Of the LaSIE System As Used For MUC-6", In *Proceedings of the Sixth Massage Understanding Conference*, Morgan Kaufmann Publishers, 1996.