pre-CODIE – Crosslingual On-Demand Information Extraction

Proceedings of HLT-NAACL 2003
Demonstrations , pp. 25-26
Edmonton, May-June 2003

Kiyoshi Sudo **Satoshi Sekine** **Ralph Grishman**
Computer Science Department
New York University
715 Broadway, 7th Floor,
New York, NY
{sudo,sekine,grishman}@cs.nyu.edu

1 Introduction

Our research addresses two central issues of information extraction — portability and multilinguality. We are creating information extraction (IE) systems that take foreign-language input and generate English tables of extracted information, and that can be easily adapted to new extraction tasks. We want to minimize the human intervention required for customization to a new scenario (type of facts or events of interest), and allow the user to interact with the system entirely in English. As a prototype, we have developed the pre-CODIE system, an experimental *C*rosss-lingual *O*n-*D*emand *I*nformation *E*xtraction system that extracts facts or events of interest from Japanese source text without requiring user knowledge of Japanese.

2 Overview

To minimize the customization of the IE system across scenarios, the extraction procedure of pre-CODIE is driven by the query from the user. The user starts the procedure by specifying the type of facts or events of interest in the form of a narrative description, and then pre-CODIE customizes itself to the topic by acquiring extraction patterns based on the user's description. Pre-CODIE, as an early attempt at a fully-automated system, still needs user interaction for template definition and slot assignment; automating these steps is left as future work.

Pre-CODIE interacts with its user entirely in English; even for slot assignment of the extraction patterns, the system translates the Japanese extraction patterns, which are based on subtrees of a dependency tree (Sudo et al., 2001), by word-to-word translation of each lexical item in the patterns. For ease of use, the Japanese extraction patterns are not only translated into English, but also shown with translated example sentences which match the pattern.

3 System Architecture

Pre-CODIE is implemented as an integration of several modules, as shown in Figure 1: translation, information retrieval, pattern acquisition, and extraction.

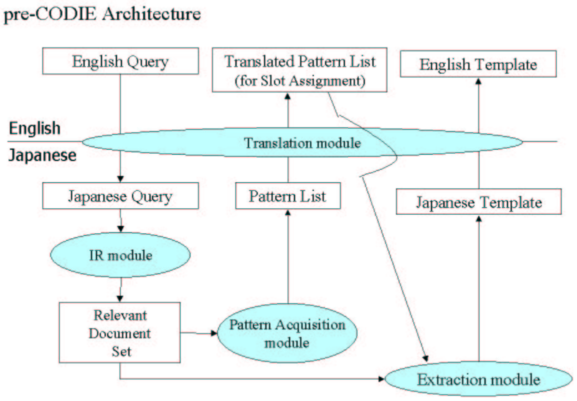


Figure 1: system architecture

First, the system takes the narrative description of the scenario of interest as an English query, and the Translation module (off-the-shelf IBM *King of Translation* system) generates a Japanese query. The IR module retrieves a set of *relevant documents* from Japanese Mainichi Newspaper from 1995. Then, the Pattern Acquisition module produces a list of extraction patterns from the *relevant document set* sorted by their relevance to the scenario (Sudo et al., 2001). Pre-CODIE asks the user to assign each placeholder in the patterns to one of the slots in the template. Finally, the Extraction module performs the pattern matching with slot-assigned patterns to each text in the *relevant document set* and generates a filled Japanese template, which is translated slot-by-slot into English for the user.

4 An Example procedure: Management Succession

From the user's point of view, pre-CODIE works as follows with the screenshots in Figure 2.

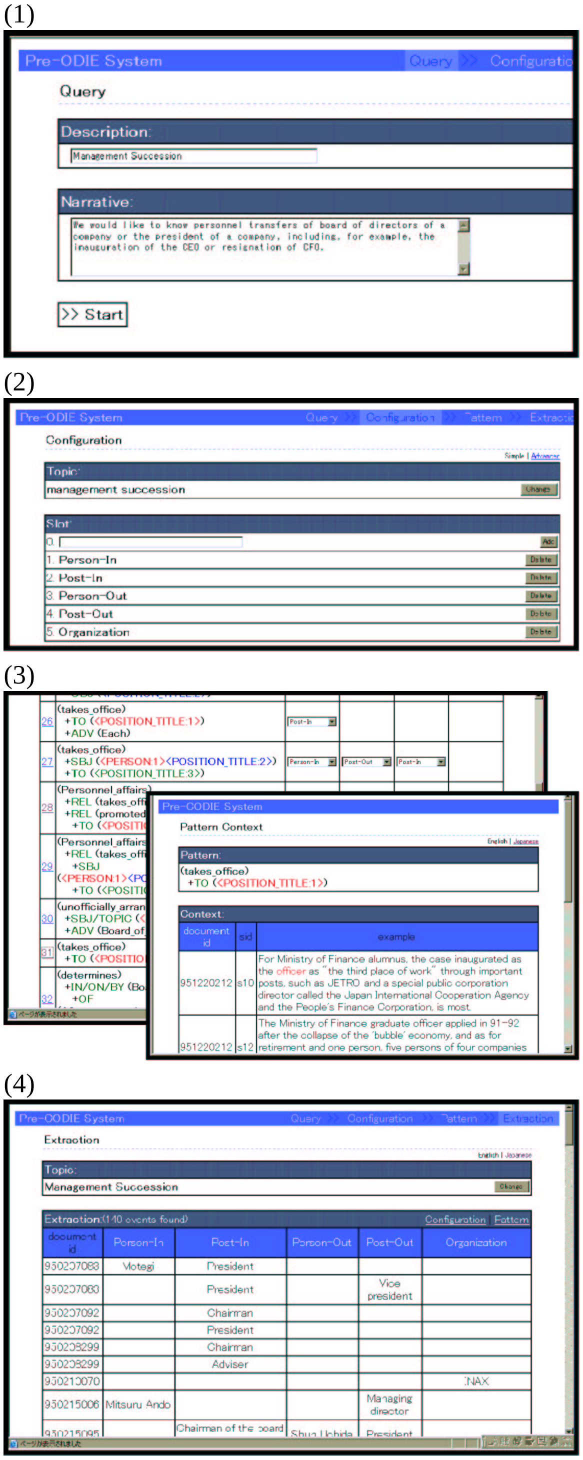
1. **Query:** The user types in the narrative description of the scenario of interest, one phrase in "description" text-box and more detail optionally given in "narrative" text-box: *"Management Succession: ..."*.

2. **Configuration:** The user adds and/or deletes the slots in the template; Add *"Person-In"*, *"Person-Out"*, *"Post-In"*, *"Post-Out"*, and *"Organization"*.

3. **Slot Assignment:** The user assigns a slot to each placeholder in the pattern by choosing one of the slots defined in step 2; Assign *"(be-promoted (PERSON-SBJ))"* to *"Person-In"*.

   Also, the user can see the example sentences with the match of the pattern highlighted. This will make it easier for the user to understand what each pattern aims to extract.

4. **Extraction:** The user gets the extracted template and repeats this procedure until the user gets the right template by going back to step 3 to change and/or add slot assignments, and by going back to step 2 to delete and/or add slots in the template.



Query

Description: Management Succession

Narrative: We would like to know personnel transfers of board of directors of a company or the president of a company, including, for example, the inauguration of the CEO or resignation of CFO.

Configuration

Topic: management succession

Slot:

0.

1. Person-In
2. Post-In
3. Person-Out
4. Post-Out
5. Organization

```
26 (takes_office)
     +TO (<POSITION_TITLE:1>)
     +ADV (Each)                                  Post-In
27 (takes_office)
     +SBJ (<PERSON:1><POSITION_TITLE:2>)
     +TO (<POSITION_TITLE:3>)                     Person-In   Post-Out   Post-In
28 (Personnel_affairs)
     +REL (takes_offi
     +REL (promoted
       +TO (<POSITI
29 (Personnel_affairs
     +REL (takes_offi
       +SBJ
   (<PERSON:1><PC
       +TO (<POSITI
30 (unofficially_arran
     +SBJ/TOPIC (<
     +ADV (Board_of
31 (takes_office)
     +TO (<POSITIO
32 (determines)
     +IN/ON/BY (Bo
       +OF
ページが表示されました
```

Pre-CODIE System

Pattern Context

Pattern: (takes_office) +TO (<POSITION_TITLE:1>)

Context:

document id	sid	example
951220212	s10	For Ministry of Finance alumnus, the case inaugurated as the officer as "the third place of work" through important posts, such as JETRO and a special public corporation director called the Japan International Cooperation Agency and the People's Finance Corporation, is most.
951220212	s12	The Ministry of Finance graduate officer applied in 91-92 after the collapse of the 'bubble' economy, and as for retirement and one person, five persons of four companies

Pre-CODIE System

Extraction

Topic: Management Succession

Extraction:(140 events found) Configuration | Pattern

document id	Person-In	Post-In	Person-Out	Post-Out	Organization
950207083	Motegi	President			
950207083		President		Vice president	
950207092		Chairman			
950207092		President			
950208299		Chairman			
950208299		Adviser			
950210070					INAX
950215006	Mitsuru Ando			Managing director	
950215095		Chairman of the board	Shun Uchida	President	

ページが表示されました

Figure 2: Screenshots of an Example procedure: Each image corresponds to the procedure in Section 4.

Acknowledgments

This research is supported by the Defense Advanced Research Projects Agency under Grant N66001-00-1-8917 from the Space and Naval Warfare Systems Center San Diego. This paper does not necessarily reflect the position or the policy of the U.S. Government.

References

Kiyoshi Sudo, Satoshi Sekine, and Ralph Grishman. 2001. Automatic pattern acquisition for japanese information extraction. In *Proceedings of the Human Language Technology Conference (HLT2001)*, San Diego, California.