

# Textual Information Extraction in the face of Information Deluge

**Li Kwang Angela Wee,  
Loong Cheong Tong**  
Kent Ridge Digital Labs  
21, Heng Mui Keng Terrace  
Singapore 119613

**Chew Lim Tan**  
School of Computing  
National University of Singapore  
Kent Ridge  
Singapore 119260

## Abstract

An information extraction system has been developed. The system processes the input text to turn it into a directed graph structure by consulting grammar rules and a domain ontology. The directed graph enables a template filling process to allow extraction of salient information through a frame structure. The system has found three applications, namely, Personalized News Dissemination, Message Formatting Expert, Recruitment Processing System.

## 1 Introduction

Large organizations nowadays process thousands of free-form text messages each day, including application forms, faxes, telexes, memos, and reports. The ability to use computers to extract relevant information quickly from texts is becoming crucial to the productivity and competitiveness of the individual and the organization. To achieve this, we have developed an information extraction system that allows extraction of relevant information from free form text based on a set of pre-defined template of what information is to be extracted.

## 2 Knowledge Representations

The system uses three knowledge representation structures: Message Intermediate Representation (MIR), a domain ontology, and Frame for Extracting Information from Messages (FEIM).

MIR is a directed graph whose nodes represent concepts that occur in the text. The nodes are inter-connected through labeled directed links which represent the relationships between these concepts. The domain ontology represents the domain knowledge in object-centered hierarchies with inheritance, with the objects representing the terms and concepts in the domain, and the relationships of these concepts through the hierarchies.

FEIM employs templates for the specific purpose of extracting information from input messages. It collects all related attributes of the target information in a single source, such as the value, functions required to augment the value and actions that may be triggered by the target information. Furthermore, it handles operational issues such as the derivation of information that is a prerequisite of the target information as well as cross-checking of the target information and the derived information.

## 3 Information Extraction Process

The system first performs text pre-processing by decomposing the input text into useful text segments, such as addresses, organization names, and other layout structures i.e. sections, paragraphs, sentences and tabulated data. These text segments are next subjected to Surface Text Analysis to turn into a node-list containing basic word constituents with appropriate potential parts of speech and other linguistic information, through a Morphological Analysis. This is followed by a Word Pattern Analysis to form lexical patterns from each base word with its surrounding contextual words. A Deep Text Analysis step is next employed to discover more complex concepts through a structural analysis followed by syntax and semantic normalization. Grammar rules and the domain ontology are consulted in the above processes, the final output of which is the MIR representing the meaning of the input text.

FEIM initiates a series of template filling processes by accessing the MIR through a graph manipulation language in order to extract relevant information to fill the appropriate slots in the respective templates. These slot values form the target information to be extracted which may be placed in the database as appropriate. Figure 1 depicts the entire information extraction process.

## 4 Applications

The system has been customized for three different information extraction applications, namely, DeNews, Message

Formatting Expert (MFE), and Recruitment Processing System (REP).

DeNews is a personalized news dissemination. It filters relevant news articles from selected multilingual sources (English, Malay, Chinese) on the Internet, and categorize them to user-defined classes. Following this, more advanced processing is performed on the news items, including summarizing the news by extracting relevant sentences from them, or translating them into another language.

The MFE system processes electronic applications of Letter of Credit (LC) and its accompanying annexes and extracts relevant information for update into the database as well as generate the letter of offer. The MFE LC application is a useful application to reduce the data entry work as well as to automate the generation of the first draft of letter of offer to the customers, thus reducing turnaround time for LC application.

The REP prototype processes electronic letters of applications and accompanying resumes from job applicants and then extracts relevant information about the applicants to be updated into the REP job application database. Examples of the information extracted pertaining to the candidate include name, age, qualification, majors, working experiences, strengths. The REP is a useful application for the human resource division to screen incoming applications and short-list relevant candidates for interview or even for future reference.

More details about DeNews and MFE may be found in [Wee, 1997, 1999]. Figure 2 gives an overview of REP.

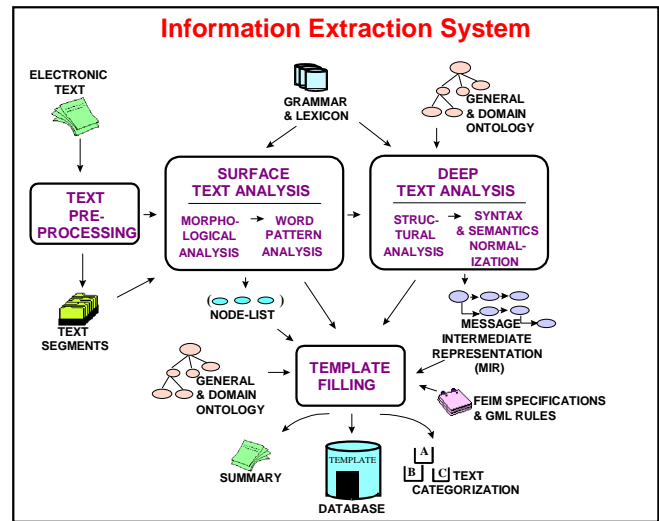


Figure 1. Entire Information Extraction Process

## References

[Wee, 1997] Li Kwang Angela Wee, Loong Cheong Tong, Tiak Jung Chng (1997). "DeNews - A Personalized News System." Journal of Expert Systems with Applications, Vol. 13, No. 4, 249-257, 1997, Elsevier Science Ltd.

[Wee, 1999] Li Kwang Angela Wee, Loong Cheong Tong, Chew Lim Tan (1999). "A Generic Information Extraction Architecture for Financial Applications", to appear in Journal of Expert Systems with Applications, Vol. 16, Elsevier Science Ltd.

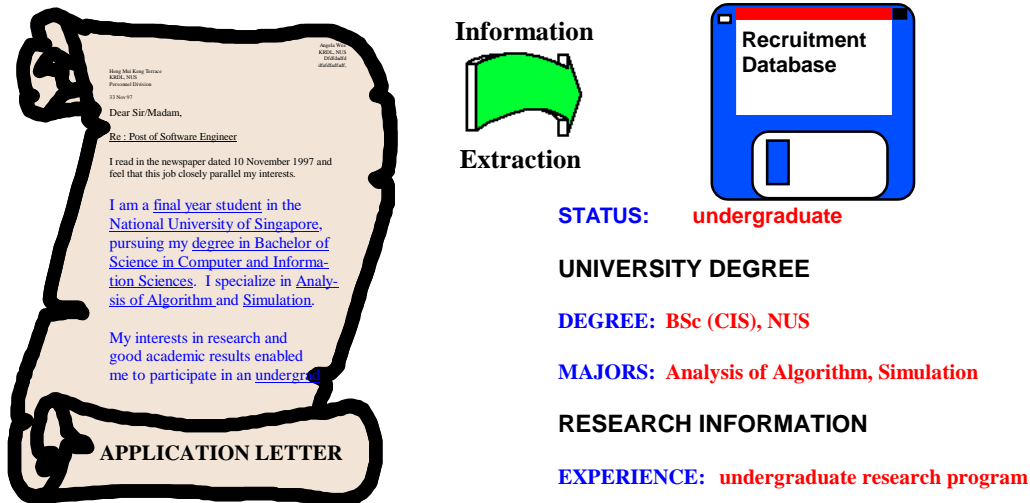


Figure 2. Recruitment Processing System