An English Analysis System of KMITT's MT Project

Ms. Monthika Boriboon
Mr. Boonchareon Sirinaokul
Asst.Prof.Woranut Kerdsinchai
Asst.Prof. Nuantip Tantisawetrat
Artificial Intelligence Center
King Mongkut's Institute of Technology Thonburi
Bangkok 10140, Thailand

1 Introduction

One important area of natural language processing which was a dream of human in 1950s is machine translation. At present, several universities and software houses have established natural language processing laboratories to develop machine translation programs. Some of them are available in the market and are widely used in many private and government sectors in many parts of the world.

In Thailand, research and development on machine translation is still in the early stage. In 1988, National Electronics and Computer Technology of Thailand (NECTEC) and Center of International Cooperation for Computerization of Japan (CICC) have signed an agreement to work on an international cooperative project on multilingual machine translation system. Other countries joining in this project are The People's Republic of China, Malaysia and Indonesia. The project will be ended in 1994.

The KMITT's MT project reported in this paper is one of the products of NECTEC-CICC Machine Translation Project.The background knowledge of the working process of the multilingual machine translation system is developed in order to respond to the aim of NECTEC that the machine translation technology should be expanded to increase the capability of researchers in computer field.

This project is only one part of a machine translation research in KMITT. It aims to build an English analysis system for analyzing the meaning of source language text input and represent its meaning in the form of internal structure, called interlingua. Some ideas on analysis dictionary, parsing techniques and internal semantic representation are adopted from PIVOT machine translation approach. The result of this study shows the possibility of using the artificial intelligence technique in constructing the practical machine

aids translation system in the future.

2 Analysis Grammar

The grammar used in this analysis system includes:

2.1 Morphology

The morphology is needed in morphological analysis. In general, the scheme of English word is in the form

$$WORD = PREFIX^{*} + ROOT + SUFFIX^{*}$$

(* known as Kleen star which indicates that the item can be repeated zero or more times)

This given rule generates quite a number of pattern, such as

$$\emptyset + ROOT + SUFFIX$$
$$\emptyset + ROOT + 0$$
$$\emptyset + ROOT + SUFFIX + SUFFIX$$

and so on (The zero indicates a null items). Consequently, the morphology used in morphological analysis is grammar on suffixes and prefixes such as

-Plural Ending : -s, -es, -ies, ives, ss
-Tense Ending : -d, -ed, -ing, -en
-Comparative Ending : -er, -est, -ier
-Adverbial Ending : -ly

2.2 Phrase Structure Grammar

The phrase structure grammar consists of a list of possible construction stated in terms of the constituents that a sentence can have. The phrase structure rules used in syntactic-semantic parsing process are as follows:

Rule	1.	S	-->	NP VP NP
Rule	2.	S	-->	NP VP (PP)*
Rule	3.	NP	-->	(<DET>) N ¦PRON
Rule	4.	NP	-->	(<DET>) (ADJP) NP
Rule	5.	NP	-->	NP PP
Rule	6.	PP	-->	<PREP> NP*
Rule	7.	VP	-->	(<AUX>) (<NEG>) VP1
Rule	8.	VP1	-->	(ADJP) V (ADV)
Rule	9.	ADJP	-->	(ADV) ADJ

* means that the constituent can occur recursively.
<> means that the inside element is treated as a feature of the head of the phrase.
() means that the constituent is optional.

2.3 Dependency Grammar

The dependency grammar is used in the construc-

tion of syntactic-semantic representaion to specify the controller term, called head and the controlled term, called depender. This grammar can express as the following expressions :

1.^(V)	verb can appear with no controller thus verb can locate at the top of the hierarchy.
2. V(NP,^,NP)	verb controls the left and right noun phrase of the verb.
3. N(<DET>,^,PP)	noun controls the left determiner and the following preposition phrase of noun by keeping the left determiner as a feature of the noun.
4. N(ADJ,^)	noun controls the left adjective of noun.
5. V(<AUX>,^)	verb controls the left auxiliary verb by keeping it as a feature of the verb.
6. V(^,PP)	verb controls the following prepositional phrase.
7. V(ADV,^ADV)	verb controls the left and right adverb of the verb.
8. V(ADJ,^)	verb controls the adjective phrase in front of the verb.
9. ADJ(ADV,^)	adjective controls left adverb.

2.4 Case Grammar

A case expresses the semantic role that a noun phrase plays with respect to the verbs, adjectives or other nouns around it. In this paper, a list of case has been defined for expressing the semantic relation between words such as AGENT, OBJECT, CAPACITY, NUMBER, LOCATION etc. Base on the Case Grammar, a main verb is treated to be the main focus of the other phrases.

3 Parsing Technique

This syntactic-semantic parser is designed to process a sentence input by using both deterministic top-down and bottom-up parsing. The deterministic bottom-up parsing is used in the constructions of noun phrases, adjective phrases and verb phrases. The constructions is started by considering the surface words and then grouping them into phrase structures. On the other hand, the deterministic top-down parsing is used in the sentence construction by starting from the main verb and find its surrounding arguments. Both cases are carried out with no backtracking since the the parsing is deterministic.

4 Procedural Components and System Design

The procedural components of the system are morphological analyzer, syntactic-semantic parser and conceptual interpreter which function respectively.

The system design is based on an artificial intelligence technique. The processor of the system is divided into two main parts which are knowledge base and inference engine. The knowledge base describes the operations corresponding to the associative subsystem. The inference engine controls the knowledge based processing. This engine composes of two parts. The first is knowledge based compiler which functions as rule command interpreter and the second is rule inferencing system which controls the sequence of rule usages. The advantage of this system design approach is that it is easy to maintenance the the knowledge.

Before the operation in the system is carried out, the knowledge engineer has to put the linguistic knowledge, written in a designed language, into the knowledge base. This knowledge is compiled by the knowledge based compiler and stored in the linked list structure.

In addition to the knowledge base, the analysis system also utilizes the knowledge contained in analysis dictionary. The information from the analysis dictionary will be loaded and attached to each entry of the linear structure in morphological stage.

The block diagram of the designed system is shown below:

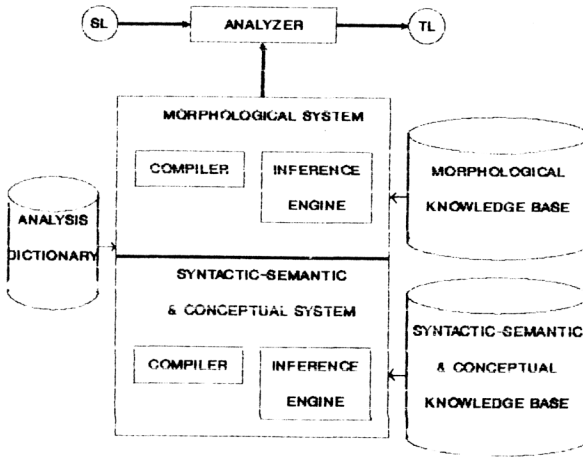

Fig 1. The block diagram of the designed system