A Generation System of KMITT's MT Project

Ms. Kanlaya Narue-domkul
Mr. Booncharoen Sirinaovakul
Asst.Prof. Nuantip Tantisawetrat
Machine Translation Laboratory
King Mongkut's Institute of Techonology Thonburi

Abstract This paper provides an overview of a part of the machine translation project which has been studying in Machine Translation Laboratory at King Mongkut's Institute of Technology Thonburi. The project aims at developing a generation system which is to generate the Thai language as a target language in the machine translation process. In this study, the generation system produces Thai sentence from the interlingua which represents the meaning of the source language. This interlingua expresses conceptual relations by using semantic cases resulted from the analysis system. There are three main steps involves in generation system, first, syntactic generation which creates the syntactic structure of the target language by using the generation grammars; second, words selection which selects the most appropriate Thai word for each concept in the interlingua basing on word category, subcategory and other related issues; third, words ordering which orders the words according to the patterns of the Thai sentences.

The most important aspect for the generation process is the dictionary. The generation dictionary contains syntactic information, verb patterns and mapping patterns required for semantic-syntax mapping.

This system is developed and tested with 50 simple sentences. There are 189 words in the dictionary. The study has indicated that the target language (Thai) developed is accurate and reliable in representing the source language in the translation process.

1. **Introduction.** One of the famous strategies in Machine Translation(MT) system is the Interlingua MT strategy. The main idea of this strategy is that the source language and the target language never contact directly. The meaning of the source language sentence is represented in an artificial language, called 'INTERLINGUA'. The process of this translation begins with analyzing the source language sentence. The output of the analyzed sentence is represented by the interlingua which is dependent from any form of a particular lan-

guage. This interlingua represents the semantic structure of the input sentence. The target language, then, is generated directly from this interlingua. One of the advantages of using this process is that it reduces a lot of redundant information. For instance, n different languages, n concept dictionaries and n sets of *grammar rules are needed for n(n-1) translations.*

In this paper, the proposed method uses the interlingua as a strategy of generation and artificial intelligence as a technique for problem solving. The reasons are described as follows:

First, the development of the dictionary and the grammar rules of any language are related with the interlingua, not only for analyzer but also for generator. The analysis module and the generation module are connected by interlingua. Therefore, the dictionary and the grammar rules for each module can be developed separately.

Second, by using artificial intelligence technique, the grammar rules in the knowledge base can be developed independently from computer programming. Therefore, the created rules are maintainable. The knowledge developer can develop and edit the grammar rules without touching the computer program. It is convenient for the developer who is scare of computer language to develop grammar rules for the system.

This paper provides an overview of a generation system which has been studied in Machine Translation Laboratory at King Mongkut's Institute of Technology Thonburi. The generation process, the designed system and the results and conclusion showing examples of generation process are discussed in details.

2. **Generation Process.** The project aims at developing a generation system which is to generate the Thai language as a target language of the machine translation. The interlingua resulted from the analysis system is used as an input. The developed prototype is limited to a simple sentence, and each sentence is considered independently. The databases used in the process are a generation dictionary (IL-TL dictionary) and other tables which are related to the process. In this study, the generation system produces Thai sentence from the interlingua which represents the meaning of the source language. This interlingua, resulted from the analysis system, expresses the conceptual relations by using semantic cases.

The three main steps in generation process are Syntactic Generation, Words selection and Words ordering. Syntactic Generation creates the syntactic structure of the target language by using the generation grammar. Words Selection selects the most appropriate Thai word for each concept in the interlingua. Words Ordering orders the generated words.

2.1. Syntactic Generation. The interlingua, represented in a semantic tree structure, is the input of the generation system. The syntactic generation procedure is the fist procedure to process the interlingua. It creates the syntactic structure for the Thai sentence by mapping the interlingua's semantic relation with the syntactic relation. The procedure consists of dictionary loading, syntactic mapping and subject selection.

Dictionary Loading process searches the information for all conceptual primitive (CP) from generation dictionary. The process is done by #LDICT command in the knowledge base and each CP-name is used as a keyword. In searching for CP's information, if there are some CPs that have more than one set of information, the appropriate information will be selected by Syntactic Mapping procedure.

Fig.1 is the example of dictionary information for CP-name "TALK" that has two sets of information.

CP-name : TALK

CONCEPTUAL	*ENTRY*	*TCAT*	*TSUBCAT*	*TMAPS*	*TVP*	*AKO*
TALK	คุย	*V*	*V*	*SUB=AGT,COMP=OBJ*	*3*	*2111*
TALK	คุย	*V*	*V*	*SUB=AGT*	*1*	*2111*

Fig.1 Dictionary information

Syntactic Mapping procedure maps the semantic relations (only the relations between root node and its daughters) with TMAPS of the root node. At this state, the most appropriate information of the root node is selected and the syntactic cases are also mapped. For the interlingua which its relations are both obligatory and free cases, the procedure has to cut free cases by comparing them with the free-case table before selecting the syntactic cases.

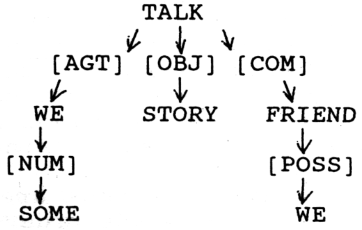

Fig.2a

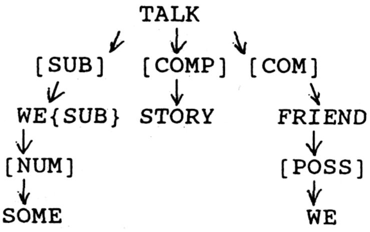

Fig.2b

Subject Selection process searches for the branch whose syntactic case is 'SUB' and moves it to the left most branch. Then, the procedure assigns value 'SUB' to the feature of its daughter.

Fig.2a is the interlingua. It composes of two obligatory cases, AGT and OBJ, and one free case, COM. Fig.2b shows the output of the syntactic generation procedure. The value 'SUB' is assigned to Node WE.

2.2. Words Selection. Words selection is the main task of the generation system. It selects the most appropriate Thai words and maps them onto the CPs of interlingua. In this project, the process of words selection are as follows.

2.2.1. Thai word generation. The process maps Thai words onto the root node. This is done by retrieving Thai words from the "ENTRY" field of the information, selected by syntactic mapping procedure. For daughter node, The system uses the syntactic case of the syntactic structure as an information for selecting the Thai word for the CP. This syntactic case indicates category or subcategory of the daughter node. The system compares this syntactic case with the category or subcategory of the information loaded from dictionary. The Thai word that has the same category or subcategory as indicated by syntactic case will be selected.

For other nodes, or leaf nodes which their parent node is not root node, their TSUBCATs are selected by using NMAPS table. NMAPS is defined by considering the relation of meanings between noun and its modifiers. The example of NMAPS is shown in fig.3.

Semantic Case	TSUBCAT
NUM	DDBQ, DIAN, JNRN, JNRP, NCNM
CAP	NCMN
POSS	PPRS, NCMN
...	....

Fig.3. NMAPS table

From fig.3., the process compares the leaf node's case with the semantic case of NMAPS table. Then, the set of TSUBCATs of the matched semantic case is loaded. These TSUBCATs are compared with the TSUBCATs loaded from dictionary and the intersected TSUBCATs are selected.

2.2.2. Classifier Generation. In Thai, a noun has a classifier when it is numbered. The classifier node