# Information-based Language Analysis for Thai

Virach Sornlertlamvanich
Researcher
Center of International Cooperation for Computerization
12-35, Shibaura 4-Chome, Minato-ku, Tokyo, Japan 108

Wantanee Phantachat
Researcher
National Electronics and Computer Technology Center
Rama 6 Rd., Phayatai, Bangkok, Thailand 10400

## ABSTRACT

Thai language is recognized as an isolative language having neither lexicon inflection due to word agreement and tense (as in English) nor obvious syntactic case marker (as in Japanese). The position of word in a sentence is the important superficial syntactic information for recognizing the meaning and the syntactic role. In this paper, we are going to describe the methodology and algorithm to cope with the mentioned Thai specific language phenomena. The rather fixed relation of word position and its syntactic role gives a well-formed pattern of a phrase. Therefore, localization of pattern analysis helps much in phrasal recognition and works well in lexicon disambiguition. In a sentence, the relatively less ambiguous in concept of words (variety of concepts) are consecutively determinated to make up a bunch of concept. Then, according to the information retrieved from dictionary, subcategorization employed by the verb of the sentence will finally create the relation between those groups of concepts to build up a dependency structure to represent the meaning of the sentence. Besides the lexicon information from the dictionary, the grammatical rules are employed to identify the appropriate semantic relation between concepts with lexicon functional reasoning in the pair of provides and requires attribute.

Keywords: subcategorization, dependency structure,
functional reasoning, interlingua

## 1 INTRODUCTION

This paper presents a methodology and algorithm of the parser in Thai sentence analysis in multilingual machine translation system. The language analysis methodology is partially based on dependency grammar, representing the meaning of a sentence in an interlingual representation--in other word is called a conceptual dependency structure. As Thai language is an isolative language with richness of lexical ambiguities, the methodology is constructed to extract those ambiguities by interpreting both syntactic and semantic roles of the language. The presenting methodology is mainly considered in two approaches. Firstly, it concerns the subcategorization of verb and its arguments. The verb pattern table is created as the information based knowledge. Secondly, the provides and requires attributes are considered to define the semantic relation of two concepts by using lexicon *functional reasoning which is implemented in the rule* base.

The system is authorized in the name of the Machine Translation Project for Asian Languages, supported by the Ministry of Trade and Industry (MITI) of Japan, conducted by the Center of Cooperation for Computerization (CICC) cooperating with other four governments of the People's Republic of China, the Republic of Indonesia, Malaysia and Thailand. The parser, mentioned hereinafter, is an out-come of the co-research between CICC and NECTEC (National Electronics and Computer Technology Center, Ministry of Science, Technology and Energy of Thailand).

## 2 DIFFICULTIES IN THAI SENTENCE ANALYSIS

Isolative and mono-syllable characteristics in Thai sentence leave us so many levels of problems to solve in the computer system. One surface word usually has more than one meaning and/or more than one syntactic category. In the information preparation step, we have tried to identify the grammatical role of words in each sentential form. As the result, we realized that besides the meaning of the word itself, the word position is properly notified to be the grammatical role for itself. After testing the words with any arbitrary position in a sentence form, we grouped up a set of word category with the consideration of the implementation of grammatical rule when applying to the organization. The inventory of word category employed in this analysis system was presented in Computer Processing of Asian Languages '89 at AIT.

The difficulties in Thai sentence analysis, from language computing standpoint, may be raised in this prototyping analysis system to a summary of such:-

(1) ***Polysemy phenomenon*** which occurs in most of Thai single word. The more frequently the word appears, the more meaning derivation it has. This is the nature of the easy-to-use words. So that, formulating the

constrains for their usage distinction is needed. The constrains which is taken into considered can be its grammatical role (Word category; CAT, SUBCAT) or syntactic usage pattern (Verb pattern; VP) or the information of neighboring words in the sentence (in pragmatic rules). For instance, the word "/caak/" has at least three meanings as follows:

L1; /caak/ : #CAT.{V}, #CP.{LEAVE}
#CAT.{PREP}, #CP.{FROM}
#CAT.{N}, #CP.{NIPA^PALM}

(2) *Appropriate word, as well as sentence, boundary assignment.* Thai language has a nature of being written in a string of characters with no any remarkable word boundary or sentential marker. This really causes the difficulties in analysis as it must have been segmented into sentences or words. In addition, Thai language has neither punctuation marker to mark the clause boundary. To separate the clause, space between string of characters is proposed to be the marker determining the boundary of the clause or the sentence. But the word segmentation is still the problem in analyzing as how precise the word segmentation is. As the word formation in Thai language is formed by attaching each words together to form the new word, so the problem is how to keep the word in dictionary, single word or compound word. For instance, "/kaanplxxphaasaaduaikhoomphiuter/" is composed of 5 single words as "/kaan/", "/plxx/", "/phaasaa/", "/duai/","/khoomphiuter/". This word can be interpreted as follows:

L2; /kaanplxxphaasaaduaikhoomphiuter/
can be segmented in 4 different ways:-
5 words as /kaan/, /plxx/, /phaasaa/, /duai/, /khoomphiuter/
4 words as /kaanplxx/, /phaasaa/, /duai/, /khoomphiuter/
3 words as /kaanplxxphaasaa/, /duai/,/khoomphiuter/
1 words as /kaanplxxphaasaaduaikhoomphiuter/

(3) *No inflection, no verb agreement.* Thai language is an isolative and monosyllable language. There is no inflection to mark morphology of the language like English or Japanese. On the other hand, those morphology is designated by the lexical item. For example, the passive voice is indicated by a lexical item in the position of pre-verb.

S1; /nakrian/ /thuuk/ /khruu/ /longthoot/
student passive marker teacher punish
"The student is punished by the teacher."

Like the passive voice, Thai language expresses tense, aspect, modality in lexical items modifying verb in pre- or post-position.

(4) *Tense point of view*. In (3), we have mentione that tense in Thai is expressed overtly by a lexica item as auxiliary category. Only one lexical item o "/ca/=will", is a marker expressing that the event i not yet occurred. So it can be summarized that Tha recognizes only two tenses:-

(a) Irrealis tense expresses that the even is not yet occurred, corresponding to future tense.

(b) Realis tense expresses that the event ha already been occurred, corresponding to present an past tense which is not distinctive.

The difficulty appears in how to assign th universal tenses of present, past or future to th interlingual representation of Thai sentence.

## 3 ANALYSIS ARCHITECTURE

The target of this analysis system is to produce a *interlingual representation (dependency tree structure)* from a linear sequence of Thai character string. Th output interlingual representation will then b transferred to sentence generation system to generat the any other specified languages. Therefore, th interlingua must be exhaustive to represent all th meaning units of the source language. The research o interlingua is carried on in other framework of th project. Here the detail of interlingual representatio and the generation part will not be discussed.

Designing this analysis system scoped to process syntactic sophisticated structure of Thai languag needs a lot of tactics in the rule implementing an rather flexible parser with ample functions for dat manipulating. The parser itself will be discussed i the next section of this paper. The followings are th postulates for system construction. These are realize in both of the parser capability and methodolog implemented in the rules.

(1) *Sentence analysis*

This is a restriction narrowing the possibl information which can be taken into account in th parse time. But, this restriction protects us fro unpredictable calculation time and misinterpretation Especially for the Thai language, there is no an sentential marker preferable whether it is a comm between phrases or a full stop at the end of sentence. Nevertheless, discourse analysis is believe to be another precise method to accurate th translation. The idea of discourse analysis is also i our extension plan.

(2) *Lookahead in parsing*

This thought positively supports the idea o using all the available information in parsing. Th