International Journal of Engineering, Mathematical and Physical Sciences
ISSN: 2517-9934
Vol:3, No:12, 2009

A Thai to English Machine Translation System using Thai LFG tree structure as Interlingua

Tawee Chimsuk and Surapong Auwatanamongkol

***Abstract*—**Machine Translation (MT) between the Thai and English languages has been a challenging research topic in natural language processing. Most research has been done on English to Thai machine translation, but not the other way around. This paper presents a Thai to English Machine Translation System that translates a Thai sentence into interlingua of a Thai LFG tree using LFG grammar and a bottom up parser. The Thai LFG tree is then transformed into the corresponding English LFG tree by pattern matching and node transformation. Finally, an equivalent English sentence is created using structural information prescribed by the English LFG tree. Based on results of experiments designed to evaluate the performance of the proposed system, it can be stated that the system has been proven to be effective in providing a useful translation from Thai to English.

***Keywords*—**Interlingua, LFG grammar, Machine translation, Pattern matching.

I. Introduction

THERE are many languages in our world that may be dissimilar or different in their word order and phrasal structures but still share the same functional vocabulary such as subject, object and so on. The basic sentences of all languages consist of Subject (S), Verb (V) and Object (O). The syntax patterns of a sentence can be classified into six types, i.e. SOV, SVO, VSO, VOS, OVS and OSV. Both the Thai and English languages fall into the SVO category.

The properties of the Thai language are as follows [10]. First, it is a single word language. Second, each word has a complete, self-contained meaning and some may have several meanings. To verify the meaning of a word its position in a sentence must be considered. Third, the word order is very important as it indicates the part of speech of the word. Lastly, a word may be augmented or modified by another word indicating gender, number of nouns, tense or mood of verb. LFG (Lexical functional grammar) [6] was developed by Joan Bresnan and Ronald Kaplan to study several aspects of linguistic structure and how they interrelate. Our machine translation system uses only two LFG-structures, namely Functional structure (f-structure) and Constituent structure(c-structure). F-structure represents the grammatical functional aspects of a sentence e.g. subject, object and abstract features such as tense and case. C-structure represents the word order and phrasal grouping by means of phrase structure trees. There were a number of research works covering machine translation systems based on LFG grammar. A bottom-up parser using LFG grammar was used to translate the Turkish language [8]. LFG grammar was also applied to the Chinese language [13]. Furthermore, LFG was used in syntactic analysis and a conceptual graph was used in semantic analysis [2]. There are several machine translation systems in many languages which successfully use Interlingua as an intermediate representation for machine translation, for instance, Government and Binding (GB) [4], semantic frame [11], recursive framework [18], translation templates [9], translation pattern based on Context-free Grammar [20], a case frame [19] and semantic tree [25]. This paper presents a Thai to English Machine Translation System that uses LFG tree structure as Interlingua. The system can translate both phrases and simple sentences (no serial verb construction) chosen from Orchid Corpus [16].

The paper is structured as follows: Section 2 presents the design and process of the system; Section 3 gives the experimental results designed to evaluate the performance of the system, with Section 4 providing the conclusion.

II. The Design and Process of the System

The system is designed into four phases, as depicted in Fig.1

A. First phase: Thai syntactic analysis

This phase performs a syntactic analysis on a given Thai sentence using LFG grammar. GFU-LAB [7], which is a software package developed for syntactic analysis using lexical functional grammar, was modified to support Thai language, from nested list structure. The Left-corner bottom up parsing technique is used in the analysis. The analysis requires two sets of information as follows,

1) Thai LFG grammar rules

To develop Thai LFG grammar rules [1]-[14]-[15], about 200 Noun phrases, verb phrases, preposition phrases and simple sentences from Orchid Corpus were analyzed using IC (Immediate constituent) theory. Each phrase or sentence is repeatedly divided into successive constituents until the smallest constituents - consisting of only a word or meaningful part of a word - are reached. Each step of the constituent division can then be translated into a Thai grammar rule in a context-free form. Finally, all generated Thai grammar rules are transformed into a c-structure format of LFG grammar.

2) Thai LFG lexicon

Tawee Chimsuk, Ph.D.student is with of Applied Statistics, National Institute of Development Administration (NIDA), Bangkok, Thailand (e-mail: tvchim@gmail.com).

Surapong Auwatanamongkol is an Associate Professor with School of Applied Statistics, National Institue of Development Administration (NIDA), Bangkok, Thailand (e-mail: surapong@as.nida.ac.th)

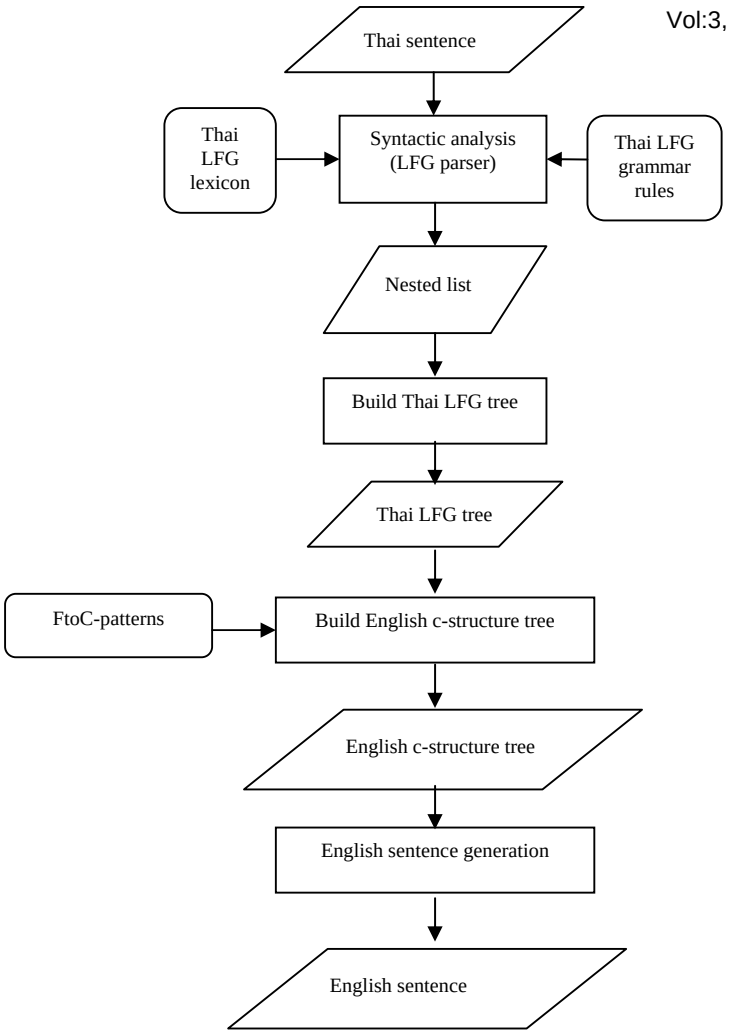

Fig.1 The Architecture of the system

From the contrastive analysis of English and Thai [24], English is an inflectional language. Morphemes can be changed inside or affixed so that information - such as Gender, Number, Tense and so on - can be expressed inside the morphemes, for example:

'books' Number = plural.
'went' Tense = past simple.

Unlike English, Thai is not an inflectional language. Morphemes can not be changed inside. To express information - such as Number, Tense and so on - other morphemes must accompany with them, thereby forming a compound word. Each word in Thai LFG lexicon [3]-[5]-[12]-[17]-[21]-[22]-[23] consists of information about the word, for example:

ระบบ (system) = NOUN
head = ระบบ
pred = system
type = NCMN
ontology = equipment.

The output of the syntactic analysis of a Thai sentence is in the form of a nested list structure of Prolog. The list consists of two structures of LFG grammar, namely c-structure and f-structure. For example, given a Thai phrase **" คณะวิศวกรรมศาสตร์ ",** the output of the syntactic analysis is as fig. 2

```
[[modi=[wordtype=CLASSIFIER, head=คณะ, sem=[pred=faculty,
type=CLTV, num=pl, ontology=community]], wordtype=NOUN,
head=วิศวกรรมศาสตร์, sem=[type=NPRP, pred=engineering,
ontology=applied_science, modifier=[pred=faculty, type=CLTV,
num=pl, ontology=community]]]]
```

Fig. 2 An example of the output resulting from syntactic analysis

B. Second Phase: Building a Thai LFG tree

In this phase the output from the first phase in nested list format is converted into a general tree structure. For example, the output of Fig.2 is converted into a corresponding Thai LFG tree as shown in Fig. 3

C. Third Phase: Building English c-structure

In this phase, an English c-structure is created, which is equivalent to the Thai LFG tree derived from phase 2. There are two steps in this phase. The first step matches the child nodes of the root representing the sentence level of the Thai LFG-tree against a set of predefined patterns. If a match is found, the child nodes are reordered and converted into the equivalent nodes of English c-structure at the sentence level. The predefined patterns are so-called FtoC-patterns. The second step is done recursively at each of the lower levels or the phrase levels. For each subtree it performs recursive phase-level transformation at each child of the given subtree. The phrase-level transformation involves matching the top level of the child against the predefined set of phrase level FtoC-patterns. If a match is found, the necessary reordering and conversion takes place. The output from this phase is an equivalent English c-structure tree for the given Thai sentence.

Each of the FtoC-patterns consists of a sequence of constitutes at the right hand side of a Thai LFG grammar rule. The examples of FtoC-patterns at the sentence level are shown in table I while the examples of FtoC-patterns at the phrase level are shown in Table II.

The abbreviations of symbols used in the table are as follows: V_1, V_2, and V_s are finite verbs while V_3 denotes a participle. NP is a noun phrase and PP is a preposition phrase.

For the completion of the translation, the default values are defined e.g. the tense is simple, the voice is active and the article accompanying the noun is "a". Furthermore, the system can perform subject-verb matching, changing from singular noun to plural noun, changing verb to noun, etc. The examples of FtoC-patterns shown above demonstrate that the proposed system must handle tense, voice, article, and subject-verb matching during the translation.

D. Fourth Phase: Generating an English sentence

To generate an English sentence, the system traverses an English c-structure tree from left to right and prints out all English words of the leaf nodes. The output from this fourth phase will be the English phrases/sentences corresponding with – or matching - the given Thai sentence.

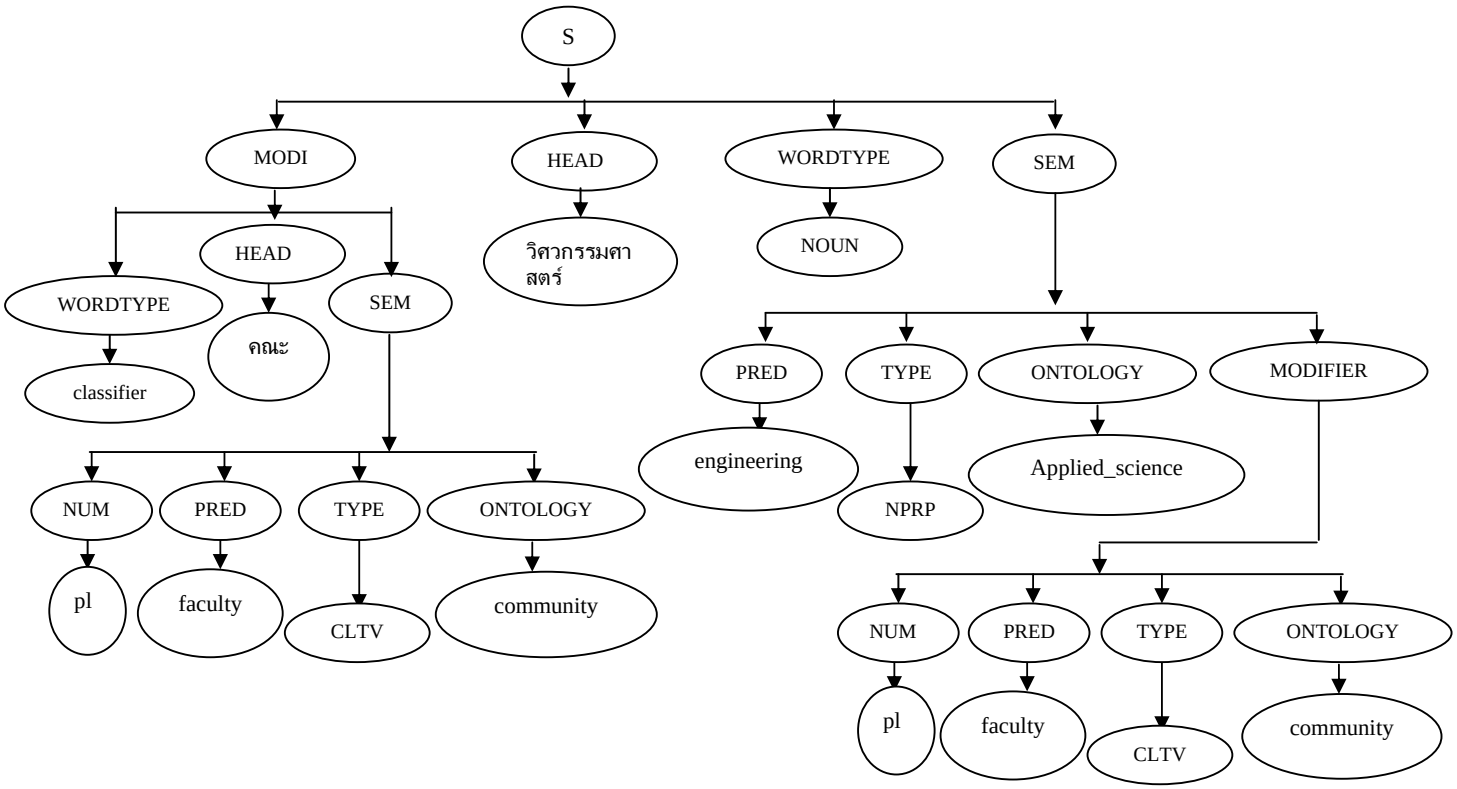

Fig. 3 An Example of a Thai LFG tree (Interlingua)

TABLE I
EXAMPLES OF SENTENCE LEVEL FTOC-PATTERNS

	Thai Patterns				English Patterns			
PASSIVE VOICE	SUBJECT (NP)	VERB	OBLIQUE PHRASE (PP)		SUBJECT (NP)	VERB (is + V_3) (are + V_3)	OBLIQUE PHRASE (PP)	
FUTURE	SUBJECT (NP)	VERB (ADJUNCT + V)	OBJECT (NP)		SUBJECT (NP)	VERB (will + V_1) (shall + V_1)	OBJECT (NP)	
PAST	SUBJECT (NP)	VERB (ADJUNCT + V)	OBJECT 1 (NP)	OBJECT 2 (NP)	SUBJECT (NP)	VERB (V_2)	OBJECT 2 (NP)	OBJECT 1 (NP)
PRESENT	SUBJECT (NP singular)	VERB (V + Modifier)	OBJECT (NP)		SUBJECT (NP singular)	VERB (Vs + Modifier)	OBJECT (NP)	
	SUBJECT (NP plural)	VERB (V + Modifier)	OBJECT (NP)		SUBJECT (NP plural)	VERB (V_1 + Modifier)	OBJECT (NP)	

TABLE II
EXAMPLES OF PHRASE LEVEL FTOC-PATTERNS

	Thai Patterns				English Patterns			
Noun Phrase	MODIFIER (Classifier)	NOUN (Proper Noun)		The	MODIFIER (Classifier)	of	NOUN (Proper Noun)	
	NOUN (Common Noun)	MODIFIER (Ordinal Number)		The	MODIFIER (Ordinal Number)	NOUN (Common Noun)		
	MODIFIER (Common Noun)	NOUN (Proper Noun)		The	MODIFIER (Common Noun)	of	NOUN (Proper Noun)	
	MODIFIER1 (Common Noun)	MODIFIER2 (Common Noun)	NOUN (Proper Noun)	The	MODIFIER2 (Common Noun)	MODIFIER1 (Common Noun)	of	NOUN (Proper Noun)
	NOUN (Common Noun)	MODIFIER (Common Noun)			MODIFIER (Common Noun)	NOUN (Common Noun)		
Preposition Phrase	PREPOSITION	OBJECT (Noun phrase)			PREPOSITION	OBJECT (Noun phrase)		
Verb phrase	NEGATOR	VERB	OBJECT (Noun phrase)		DO + NOT	V_1	OBJECT (Noun phrase)	

III. Experimental results

Two hundred Thai phrases and simple sentences (no serial verb construct) were selected from the Orchid Corpus and used as test samples for the machine translation. Each of the sentences or phrases was segmented into words. To evaluate the quality of the output from the machine translation, three senior master students of the School of Language and Communication of NIDA – who had each taken courses in the theory of translation - were presented with the input samples as well as the results of the machine translations from the system. The students were asked to assess and classify the quality of translation for each of the samples into three levels. The first level - with a score of 3 - was given to a translation that was acceptable. This meant that the output sentence could be understood and had the same meaning as the source sentence. The second level - with a score of 2 - was given to a translation that was moderately acceptable. This meant that the output sentence might have small errors but could still be understood and carried the same meaning as the original source sentence. The third level - with a score of 1- was given to a translation that was not acceptable or had to be rejected outright. This meant that the output sentence could not be understood. The scores given by the three students are summarized in Table III.

TABLE III
THE SUMMARY OF SCORES GIVEN BY THE THREE STUDENTS

	Acceptable (3) (%)	Moderate (2) (%)	Rejected (1) (%)
First student	22.0	57.0	21.0
Second student	31.5	48.0	20.5
Third student	32.5	37.5	30.0
Average percentages	28.7	47.5	23.8
Average Score		2.05	

Table III shows that most of the outputs (about 76.2 %) from the system are either acceptable or moderately acceptable. The remainder - about 23.8 % - was rejected. The average score given by the three students was 2.05.

These figures demonstrate that the proposed system can be an effective tool for Thai to English machine translation, although there are still improvements required to be made in the future.

IV. Conclusions and Future Work

This paper introduces a new framework for Thai to English machine translation. The new framework uses the LFG tree as Interlingua for the translation. The system achieves a certain degree of success in terms of the quality of the translation, making it worthy to be further pursued and improved in the future. Such work would and should include the handling of more complex sentences, tenses, voices and articles. Plans and projections for future improvements include the development of automatic tools to learn Thai grammar rules from corpus so that more complex phrases and sentences can be translated correctly.

Acknowledgment

We are very grateful to Asst. Prof. Dr. Wirot Aroonmanakun, Department of Linguistics, Faculty of Arts, Chulalongkorn University, Bangkok for his advice and comments leading to the successful implementation of the proposed system.

References

[1] A. Prasithrathsint, Y. Hoonchamlong, and S. Savetamalya. Grammatical Theories. 2nd ed. Chulalongkorn University, Bangkok, 2003.
[2] X. Briffault, K. Chibout, G. Sabah, and J. Vapillon. “A Linguistic Engineering Environment using LFG (Lexical Functional Grammar) and CG (Conceptual Graphs),” in Proc. of the LFG97 Conf. University of California, San Diego, 1997 CSLI Publications, pp. 1-14.
[3] W. A. Cook, and S.J., Case Grammar Theory.Georgetown University, Washington, D.C, 1989.
[4] [4] B. Dorr, “UNITRAN: An Interlingual Approach to Machine Translation,” in Proc. AAAI-87 Sixth National Conf. on Artificial Intelligence, Seattle, WA.,1987. pp.534-539.
[5] D. Garden and S. Wannapok, Thai-English dictionary, 3rd ed., B.E. 2549. matichonbook, Bangkok.
[6] Y. N. Falk, Lexical-Functional Grammar : An Introduction to Parallel Constraint-Based Syntax. Center for the study of Language and Informantion Stanford, California, 2001.
[7] J. C. R. Anton. (2004, Jan, 23). GFU-LAB. Available: http://www3.uji.es/~ruiz/gfulab/.
[8] Z. Güngördü and K. Of Lazer, “Parsing TURKISH using the Lexical Functional Grammar Formalism.” in Proc of the. 15th Conf. Computational linguistics, Kyoto, Japan, 1994, pp. 494 – 500.
[9] H. A. Güvenir, and I. Cicekli, “Learning Translation Templates from Biligual Texts,” in Proc. AIMSA'96, Sozopol, Bulgaria, 1996, pp. 3-11.
[10] Khun B. Bandhumedha , Structure of Thai, 14th ed.,B.E. 2545, Ramkhamhaeng University, Bangkok, 2002.
[11] Y. S. Lee, W. S. Yi, S. Seneff, and C. J. Weinstein, “Interlingua-Based Broad-Coverage Korean-to-English Translation in CCLINC,” in Proc.of the first international conf.on Human language technology research, San Diego, 2001, pp. 1 – 6.
[12] Human language Technology Laboratory. LEXITRON [online], Available:http://lexitron.nectec.or.th/
[13] S. Maosong, “LFG for Chinese: issues of Representation and Computation,” in Proc.of the LFG01 Conf, University of Hong Kong, Hong Kong, 2001.
[14] N. Bandhumedha, Thai Grammar. 3rd ed., Chulalongkorn University, Bangkok,2006.
[15] N. Kanchanawan, Analysis of Thai Structure, 2nd ed., Ramkhamhaeng University, Bangkok, 1999.
[16] V. Sornlertlamvanish, T. Charoenporn, and H. Isahara, “Orchid Corpus”,Available: http://www.hlt.nectec.or.th/orchid/corpus/corpus.html.
[17] P. Sungkhavon, “Semantic relationships between nouns and verbs in Thai sentences,” Master of Arts Program, Chulalongkorn Univ., Thailand, 1984.
[18] Q. Xuan, Z. Huiping, and C. Huowang, “An Interlingua-Based Chinese-English MT System,” Journal of Computer Science and Technology. Vol. 17, Issue 4, July, 2002, pp. 464 – 472.
[19] C. K. Turhan, “An English to Turkish Machine Translation System Using Structural Mapping,” in Proc.of the fifth Conf. Applied natural language processing, Washington, DC.,1997, pp. 320-323.
[20] K. Takeda, “Pattern-Based Machine Translation,” in Proc.of the 16th conf. on Computational linguistics, Copenhagen, Denmar, 1996, pp. 1155 – 1158.
[21] The Royal Institute Dictionary, 2003. The Royal Institute Dictionary B.E. 2542 1st ed. nanmeebooks, Bangkok.
[22] V. Sornlertlamvanish and M. Boriboon, “Thai Concept Classification (Technical Report),” Bangkok, Available: www.links.nectec.or.th/pub/tech/akotech.ps.g

[23] V. Sornlertlamvanish and W. Phantachat, "Interlingual Expression for Thai Language (Technical Report)," Bangkok, Available: www.links.nectec.or.th/pub/tech/iltr.ps.gz
[24] W. Nathong, Contrastive Analysis of English and Thai, 13th ed., Ramkhamhaeng University, Bangkok, 2003.
[25] T. Modhiran, K. Kosawat, S. Klaithin, M. Boriboon, and T. Supnithi, "PARSITTE: Online Thai-English Machine Translation," MT Summit X, September 2005, The Tenth Machine Translation Summit, Proc. of Conf., Phuket, Thailand, 2005.

Tawee Chimsuk received a B.Sc. degree (Computer Science) from Ramkhamhaeng University in 1994 and M.S. degree (Computer Science) in Applied Statistics from School of Applied Statistics, National Institute of Development Adiministration (NIDA), Bangkok, Thailand in 1998.
He is currently pursuing the Ph.D. degree in Computer Science at NIDA, Bangkok, Thailand. His research interests include Natural language processing and Machine Learning.

Surapong Auwatanamongkol received a B.Eng.(Electrical Engineering) from Chulalongkorn University, Thailand in 1978 and M.S.(Computer Science) from Georgia Institute of Technology, U.S.A. in 1982 and Ph.D.(Computer Science) from Southern Methodist University, U.S.A.in 1991.
Currently, he is an Associate Professor in Computer Science at the School of Applied Statistics, National Institute of Development Administration (NIDA), Bangkok, Thailand. He research interests include Evolutionary Computation, Pattern Recognition and Natural Language Processing.