PATTERN-BASED MACHINE TRANSLATION FOR ENGLISH-THAI

Kaewchai Chancharoen, Nisanad Tannin and Booncharoen Sirinaovakul

Artificial Intelligence Center, King Mongkut's University of Technology Thonburi
91 Pracha-uthit Thungkhru District, Bangkok 10140, Thailand
Tel & Fax (662) 4284204, E-mail : ikaeroen@cc.kmutt.ac.th

ABSTRACT

This paper proposes a new model of machine translation system in which rule-based and example-based approaches are applied for English-to-Thai sentence translation. The proposed method has 4 steps :1) analyze an English sentence into a string of grammatical nodes, based on Phrase Structure Grammar, 2) map the input pattern with a table of English-Thai sentence patterns, 3) look up the bilingual dictionary for the equivalent Thai words, reorder and then generate output sentences and 4) rank the possible combinations and eliminate the ambiguous output sentences by using a statistical method. The translated sentences will then be stored in a bilingual corpus to serve as a guide or template for imitating the translation, i.e., the example-based approach. The future work will focus on disambiguation by using semantic features to make a more precise translation.

1. INTRODUCTION

Various efforts have been made in developing machine translation (MT) systems for practical use. Historically, there are many approaches on MT research: transfer-based, interlingua-based, and etc. Among these approaches, the most distinctive are rule-based and corpus-based methods. Research on the corpus-based approach has emphasized on the importance of text corpora used as a source for linguistic and knowledge databases. There have been two major approaches among the corpus-based MT known as statistics-based and example-based. It might be said that all approaches have their own pros and cons. Therefore some MT researchers have selected and combined them together for creating a new effective model. We also combine two potential approaches to produce our own strategy; namely, rule-based and example-based.

1.1 Rule-Based and Example-Based Approaches

The rule-based translation mostly consists of (1) a process of analyzing input sentences of a source language morphologically, syntactically and/or semantically and (2) a process of generating output sentences of a target language based on an internal structure or interlingua. Each process is controlled by the dictionary and the rules. Meanwhile, the basic idea of example-based method is to translate a sentence by using translation examples of similar sentences [1]. The primary steps of example-based method are 1) collect examples in a database, 2) given an input, retrieve similar examples from the database, and 3) adapt the results of the similar examples to the current input and obtain the output.

Utsuro et al. [2] propose an example retrieval method for avoiding full retrieval of examples. The proposed method generates retrieval queries from similarities, retrieving examples through the tree structure of a thesaurus and then using binary search along subsumption ordering of retrieval queries. Cranias et al. [1] introduce a matching method that measures similarity according to both surface structure and content. Another contribution involves the use of a clustering procedure to make the best matching

example from the database. This method relies on the segmentation of sentences into coherent segments and their alignment at the sub-sentential level.

1.2 The hybrid translation method

Many researchers apply both the rule-based and example-based methods as their own hybrid methods. Shirai et al.[3] propose a new hybrid translation method that combines a rule-based with an example-based method. An outline of the hybrid algorithm is: 1) find candidate sentences which are similar to the input sentence, 2) select the template: (a) rank the candidates by similarity to the input sentence (b) cluster the translations of the candidate sentences (c) select the highest ranked pair of the best cluster, 3) translate input sentence by analogy to a selected template 4) output the adjusted sentence. For each difference, find it and translate using the rule-based modules.

They point out that this hybrid system is a method selects the strongest features of rule-based and example-based, while avoiding their weaknesses. The strengths of the rule-based method are that the information can be obtained through introspection and analysis, while those of example-based are that correspondences can be found from raw data. The weakness of the rule-based method is that the accuracy of entire process is the product of the accuracy of each sub-stage. The weakness of the example-based method is the difficulty in finding appropriate examples.

They also conclude that a useful example-based system should be able to accept loosely aligned corpora, not those aligned at low levels. Their prototype Japanese-English system tested by translating with a corpus of 5,000 sentences, can use loosely aligned texts. It allows users to take advantage of any aligned text they have by adding it to the set of sentences searched by the system.

Although these combined methods work successfully to a certain extent, it can not be applied directly to the English to Thai sentence translation. Compared with Japanese and English, both of them have sentence markers and each word in a sentence is segmented by a pause or space between the words. Therefore, the task of sentence alignment can be performed efficiently. The translation process based on the principle of analogy with a large corpora of parallel texts as database is quite successful and the resulting translations are reliable. In contrast, one lingering linguistic problem of Thai is word segmentation, due to its run-on sentences that have no boundary marker. It, therefore, causes the difficulty in sentence alignment. As a result we have not a large enough volume of aligned sentence corpus as raw data for example-based methods.

1.3 The proposed method

Nevertheless, there is an interesting method to circumvent this disadvantage. It comes from noticing that the Thai simple sentence patterns are quite similar to English, though Thai and English may be classified as from different language families. Noticeably, we can make use of this point to suggest the new method and system designed for an efficient translation. After structurally studying and comparing both English and Thai, we find that they have syntactical similarity. Therefore we attempt to match their sentence patterns which are equivalent and reconstruct the string of output before generating.

2. STRUCTURE OF ENGLISH AND THAI SIMPLE SENTENCE

Many linguists classify the Thai language is a *topic-prominent language*. The distinctive feature is that a topic or content of any sentence will be emphasized on surface structure. Some linguists classify Thai as a *discourse-oriented language* of which the sentence structure has non-restrictive pattern or form. The others point out that Thai is a language in which the omission of noun phrase is more common than any other language types. On the contrary, the English language is *subject-prominent and sentence-oriented*. The *subject* in English is emphasized more than the *predicate*. In addition, the grammatical pattern of its surface structure is specific or restrictive. However, it is generally accepted among Thai linguists and native speakers that the word ordering in a sentence of Thai is very significant. It is so because the

different order of each word is able to make a difference in the meaning of that entire sentence or make a grammatical error. For example:

(1)	จิม	ตี	จอห์น
	/jim/	/ti:/	/jon/
	Jim	hit	John.
(2)	จอห์น	ตี	จิม
	/jon/	/ti:/	/jim/
	John	hit	Jim.
(3)	*ตี	จิม	จอห์น
	/ti:/	/jim/	/jon/
	hit	Jim	John

The first two sentences are not equal in their meaning, because in (1) "Jim" is the doer and "John" is the recipient of the action "hit", and vice versa in (2). It can be said that these two grammatical strings have the same lexical entries and number of words, but different orders, and, therefore, the meaning. In case (3), this sentence is grammatically incorrect because it violates the syntactic rules of the Thai language and is meaningless.

In conclusion, though English and Thai are identified as being from different language families, there is a certain phenomenon of linguistic similarity between the two. Theoretically, there are at least *7 kernel or basic sentences* in English and Thai which are very similar, especially on word linear ordering based on syntactic rules of each language and the grammatical relationship between words in a certain pattern.

2.1 Similarity of linguistic phenomena

Every language has sentence structure including a Subject, Object and a Verb, though some sentences do not have all the three elements. Languages have been classified according to the basic or unmarked order in which these constituents occur in the language [4]. There are six possible orders : **SOV (Subject, Object, Verb, SVO, VSO, VOS, OVS, OSV-** permitting six possible language types. Coincidentally, English and Thai fall into the same type: **SVO (Subject, Verb, Object.)** The tendency of word order is as follow: *Auxiliary verb* tends to precede *Verb*, *Adverb* tends to follow *Verb*, and *Preposition* tends to precede *Noun*.

Based on the **Phrase Structure Grammar rule**, the sentence and major phrase structure of English can be rewritten as :

S	->	NP VP
NP	->	(Det) (Adj) n (PP)
VP	->	v (NP) (PP)
PP	->	Prep NP

The structure of Thai, examined in the same framework, can also be rewritten as :

S	->	NP VP
NP	->	n ┌ (Adj)(Class) (Det)(PP) └ (Class)(Adj) (Class)(Det)(PP)
VP	->	v (NP)(PP)
PP	->	Prep NP

It is remarkable to witness that the sentence structure and phrase structure of both languages are rather similar. The linguistic study, based on structuralist grammar [5] reveals that (as mentioned before) there are at least **seven basic sentence patterns** in which each grammatical constituent is put in the same ordering and underlies the same deep structure as shown in the table below :

English Sentence Pattern	**Thai Sentence Pattern**
1. NP BE ADJ	1. NP ADJ
2. NP BE ADV	2. NP BE ADV
3. NP BE NP	3. NP BE NP
4. NP V	4. NP V
5. NP V PP	5. NP V PP
6. NP V NP	6. NP V NP
7. NP V NP_1 NP_2	7. NP V NP_2 NP_1

* NP_1 = Indirect object, NP_2 = Direct object

Figure 1: Table of E-T Sentence pattern mapping

English sample sentences

1. Meg is beautiful.
2. The little cat is here.
3. My mother is a nurse.
4. They laugh.
5. She walks in the garden.
6. Pretty girl buys a ring.
7. A kind man gives the girl a dress.

Thai sample sentences

1. เม็ก สวย
   /meg/ /sua:y/
   Meg beautiful
2. แมว ตัว เล็ก อยู่ ที่ นี่
   /meaw/ /tua/ /lek/ /yu:/ /thi/ /ni/
   cat (classifier) small is here
3. แม่ ของ ฉัน เป็น นางพยาบาล
   /mae:/ /khong/ /chan/ /pen/ /nangphaya:ba:n/
   mother of mine is nurse
4. พวกเขา หัวเราะ
   /phuakkhaw/ /?hua:raw/
   they laugh
5. เธอ เดิน ใน สวน
   /theu:/ /deo:n/ /nai/ /sua:n/
   she walk in garden
6. เด็ก ผู้หญิง น่ารัก ซื้อ แหวน
   /dek/ /phu:ying/ /narak/ /seu/ /wae:n/
   child girl pretty buy ring
7. ผู้ชาย ใจดี ให้ เสื้อผ้า เด็กผู้หญิง
   /phu:cha:y/ /jaidi:/ /?hai/ /suapha:/ /dekphu:ying/
   man kind give cloth girl

2.2 Different sections

As mentioned above English and Thai simple sentence patterns are coincidentally more alike than many other languages. There are, nevertheless, some significant differences. The internal structure of a noun phrase and verb phrase are not always identical in the grammatical ordering of each element. An English noun phrase construction may consist of three components: premodifier, head noun and postmodifier. In case of premodifier, each constituent that modifies head noun is put in the left hand side or before it as illustrated below :

a) NP -> Pron
b) NP -> Det NP
c) NP -> Art NP
d) NP -> Poss Pron NP
e) NP -> Adj NP
f) NP -> n
g) NP -> n PP

Meanwhile, in Thai, a head noun is put on the leftmost position and followed by modifiers such as determiner, adjective and, uniquely different from English, **classifier** (which will not occur without determiner) :

a) NP -> Pron
b) NP -> NP (class) Det
c) NP -> NP class num
d) NP -> NP Poss Pron
e) NP -> NP (class) Adj
f) NP -> n
g) NP -> n PP

In the case of verb phrase construction, their structure is different in detail. Each language has its own particular surface structure to illustrate the grammatical feature and represent the deep underlying structure. The grammatical expressions of auxiliary verbs and modal convey a sense of *mood or intention*. The verb and noun arguments express *tense* and *aspect*. The basic verb phrase construction of English and Thai can be shown as :

English VP -> (Aux) V (NP) (PP)

Thai VP -> (Aux) V (NP) (PP)

After studying the similarities in linguistic phenomena and the differences in sections, the mapping of patterns of one language onto the other based on our criteria is experimentally shown. The different phenomena like the ordering of any element in NP construction are performed by the specific syntactic rules. We also create specific rules for adding or deleting any elements in VP constructions.

3. PATTERN-BASED TRANSLATION METHOD

This following method is designed to produce an experimental system in translating English into Thai by using the 7 basic sentence patterns as a template. After that the output sentences will be stored as raw data for further applying an example-based method. The outline of the system is as follows:
1. Morphological analysis
2. Pattern mapping

3. Looking up bilingual dictionary
4. Disambiguating possible combinations

3.1 Morphological Analysis

An input sentence is first segmented into a word, written English sentences are automatically segmented, that is, each word is separated by a pause or space, then analyzed morphologically into a morpheme (in the form of a stem or root) by applying morphological analysis rules :

```
if check_RightPos (1) ="s" then
        if check_RightPos (2) ="es" then
                if check_RightPos (3) {"ies","ves"} then
                        cut_RightPos (3) ;
                        if check_RightPos (3) = "ies" then
                          Add_char ("y") ;
                        else
                          Add_char ("f") ;
                        end if
                        if Search Dic( ) = TRUE then
                                break ;
                        end if
                else
                        cut_RightPos (2) ;
                        if Search Dic( ) = TRUE then
                                break ;
                        end if
                end if
        else
                cut_RightPos (1) ;
                if Search Dic( ) = TRUE then
                        break ;
                end if
        end if
end if
```

Figure 2 : Sample of morphological rules for cutting off the suffixes of English plurality

3.2 Pattern mapping

We make an attempt to map each pair of patterns from the most simple one to the least by using their similarity as the basis. In brief, a pair that can be mapped should be identical both in surface and deep structure. The two syntactic and semantic criterion, based on Phrase Structure Grammar and Case Grammar, respectively, of pattern mapping that we have presumed is:
a) Each entry or word in a pair should have or represent the same syntactic relationship such as "subject", "verb" and "object", lying in linear order from left to right,
b) Each entry should underlie the same semantic relationship such as an "agent" of the action, an "object" or an "experiencer" etc.

Pattern mapping or transfer between the two languages involves a few steps. First, an English input sentence is syntactically analyzed into a series of non-terminal symbols (NP, VI, VT, ADJ, etc.). This string will be checked with the table of E-T sentence pattern mapping (Figure 1). If the pattern of input sentence is identical to any pattern of English, it will be mapped to the Thai sentence pattern that is correspondent. Next, each English lexical entry will be reordered according to word ordering of Thai sentence pattern. If the different sections are found, the rules can be of help before entering the next stage.

3.3 Looking up bilingual dictionary and generating

The bilingual dictionary of 30,000 entries is created in dbase format and looked up for mapping Thai equivalent entries onto the input string. Then a Thai output sentence is generated. Due to multiple meanings of one word, there is a large number of possible combinations produced inevitably by this process. Therefore we plan to use the statistical data to determine what the most likely one should be. At least it can help in reducing the number of candidates.

3.4 Disambiguating possible combinations

In this step the statistical method is used to calculate the probabilities of word that should be translated. In other words, we search through the statistical data stored and pick out the most likely word for our translation. With this method, we can eliminate a large number of possible combinations or candidate sentences. The output sentences that are ambiguous or have nonsensical meaning will be deleted as much as possible. As a result, we can obtain the most accurate and accepted outcome.

4. SYSTEM CONFIGURATION

The system is, at present, an experimental system on the threshold to development. The translation software is written in C++ builder version 3 and runs under Window 9x on a Pentium PC. The size of the bilingual dictionary is approximately 30,000 lexical entries and stored in a text file. The dictionary has only three major fields: English entry, word category and Thai equivalent entry. The sample sentences are between 1,000-1,500 sentences.

5. CONCLUSION

Although English and Thai are classified into different language families, there is significant similarity which can be employed for translation. The proposed method is an attempt to put together two promising approaches that are claimed to be effective. Pattern mapping has resulted in reducing the number of rules and the stored translated sentences can be of great help in retrieving or imitating the translation.

Future work will involve finding a method for more precise disambiguation. In addition, more patterns should be added and mapped. Complex and compound sentence patterns should be covered. Inevitably, the dictionary must be revised to add more needed information such as semantic features and co-occurence.

Though the system is in its preliminary stage, we seek to continually improve our translation model in order to give an output with the minimum error.

6. REFERENCES

[1] Lambros Cranias, Harris Pagageorgiou and Stelios Piperidis. 1995. A matching technique in example-based machine translation. (cmlg/9508005)
[2] Takehito Utsuro. 1994. Thesaurus-based efficient example retrieval by generating retrieval queries from similarities. In *proceedings of COLING-94,* pages 1044-1048, August.
[3] Satoshi Shirai, Francis Bond and Yamato Takahashi. 1997. A hybrid rule and example-based method for machine translation. In p*roceedings of the Natural Language Processing PacificRimSymposium1997,*pages49-54, December.
[4] Fromkin Victoria and Rodman Robert. 1988. *An introduction to language*. Holt, Rinehart and Winston, Inc.,Florida.
[5] Wittaya Nathong. 1988. *Contrastive analysis of English and Thai.* Ramkhamhaeng University Press, Bangkok.
[6] Andrew Radford. 1988. *Transformational Grammar.* Cambridge University Press, Newyork.

[7] Jing-Shin Chang and Keh-Yih Su. 1993. A corpus-based statistics-oriented transfer and generation model for machine translation. In *proceedings of TMI'93*, pages 3-14, July.
[8] Naohiko Uramoto. 1994. A best-match algorithm for broad-coverage example-based disambiguation In *proceedings of COLING-94*, pages 717-721, August.
[9] Quirk Randolph, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik. 1980. *A grammar of contemporary English.* Longman Group Ltd, London.
[10]S. Kereto, C. Wongchaisuwat, Y. Poovarawan. 1993. Machine translation research and development. In *proceedings of the Symposium on Natural Language processing in Thailand,* pages 167-195, March.