Plaesarn: Machine-Aided Translation Tool for English-to-Thai

Prachya Boonkwan and Asanee Kawtrakul
Specialty Research Unit of Natural Language Processing
and Intelligent Information System Technology
Department of Computer Engineering
Kasetsart University
Email: ak@vivaldi.cpe.ku.ac.th

Abstract

English-Thai MT systems are nowadays restricted by incomplete vocabularies and translation knowledge. Users must consequently accept only one translation result that is sometimes semantically divergent or ungrammatical. With the according reason, we propose novel Internet-based translation assistant software in order to facilitate document translation from English to Thai. In this project, we utilize the *structural transfer model* as the mechanism. This project differs from current English-Thai MT systems in the aspects that it empowers the users to manually select the most appropriate translation from every possibility and to manually train new translation rules to the system if it is necessary. With the applied model, we overcome four translation problems—lexicon rearrangement, structural ambiguity, phrase translation, and classifier generation. Finally, we started the system evaluation with 322 randomly selected sentences on the *Future Magazine* bilingual corpus and the system yielded 59.87% and 83.08% translation accuracy for the best case and the worse case based on 90.1% average precision of the parser.

Introduction

Information comprehension of Thai people should not be only limited in Thai; in contrast, it should also include a considerably large amount of information sources from foreign countries. Insufficient basic language knowledge, a result of inadequate distribution in the past, conversely, is the major obstruction for information comprehension. There are presently several English-Thai MT systems—for instance, *Parsit* (Sornlertlamvanich, 2000), *Plae Thai*, and *AgentDict*. The first one applies *semantic transfer model* via the methodology similar to the *lexical functional grammar* (Kaplan et al., 1989) and it is develop with the intention of public use. The latter two implicitly apply the *direct transfer model* with the purpose of commercial use. Nonetheless, by limited vocabularies and translation rules, the users must accept the only one translation result that is occasionally semantically divergent or ungrammatical.

Due to the according reason, we initiated this project in order to relieve language problem of Thai people. In this project, we develop a *semi-automatic translation* system to assist them to translate English documents into Thai. For this paper, the term *semi-automatic translation* means the sentence translation with user interaction to manually resolve structural and semantic ambiguities during translation period. Despite manual disambiguation, we provided a simple statistical disambiguation in order to pre-select the most possible translation for each source language sentence, though. The automatic semantic disambiguation can be thus excluded with this approach.

1 Translation Approaches

We can classify current translation approaches into three major models as follows—structural transfer, semantic transfer, and lexical transfer (Trujillo, 1999).

- **Structural transfer**: this methodology heavily depends on syntactic analysis (say, grammar). Translation transfers the source language structures into the target language. This method is established by the assumption that *every language in the world uses syntactic structure in order to represent the meaning of sentences.*

- **Semantic transfer**: this methodology heavily depends on semantic analysis (say, meaning). This model applies syntactic analysis as well. On the contrary to the structural transfer, a source language sentence is not immediately translated into the target language, but it is first translated into semantic representation (Interlingua is mostly referred), and afterwards into the target language. This method is established by the assumption that *every language in the world describes the same world; hence, there exists the semantic representation for every language.*

- **Lexical transfer**: this methodology heavily depends on lexicon ordering patterns. The translation occurs at the level of morpheme. The translation process transfers a set of morpheme in the source language into that of the target language.

In this project, we decided to utilize the structural transfer approach, since it is more appropriate for rapid development. In addition, semantic representation that covers every language is now still under research.

2 Relevant Problems and Their Solutions

2.1 Structural Ambiguity

By the reason of the ambiguities of natural languages, a sentence may be translated or interpreted into many senses. An example of structural ambiguity is "I saw a girl in the park with a telescope." This sentence can be grammatically interpreted into four senses as follows.

- I saw a girl, to whom a telescope belonged, who was in the park.
- I used a telescope to see a girl, who was in the park.
- I was in the park and seeing a girl, to whom a telescope belonged.
- I was in the park and using a telescope to see a girl.

Furthermore, an example of word-sense ambiguity is "I live near the bank." The noun *bank* can be semantically interpreted into at least two senses as follows.

- *n.* a financial institution that accepts deposits and channels the money into lending activities
- *n.* sloping land (especially the slope beside a body of water)

In order to resolve structural ambiguity, we apply the concept of the *statistical machine translation approach* (Brown et al., 1990). We apply the *Maximum-Entropy-Inspired Parser* (Charniak, 1999) (so-called Charniak Parser) to analyze and determine the appropriate grammatical structure of an English sentence. From (Charniak, 1999), Charniak presented that the parser uses the *Penn Tree Bank* tag set (Marcus et al., 1994) (or PTB in abbreviation) as a grammatical structure representation, and it yielded 90.1% average precision for sentences of length

40 or less, and 89.5% for sentences of length 100 and less. Moreover, with the intention to resolve word-sense ambiguity, we embedded a numerical statistic value with each translation rule (including lexical transfer rule) with the major aim of assisting to select the best translation parse tree from every possibility (Charniak, 1997). Section 3.4 will describe the method and the tool to do so.

2.2 Phrase Translation

Phrase is a word-ordering pattern that cannot be separately translated. An example is the translation of the verb *to be*. The translation of that depends on the context—for instance, *to be* succeeding with noun phrase is translated to เป็น $/pen^m/$, succeeding with prepositional phrase to อยู่ $/yuu^l/$, in progressive tenses to กำลัง $/kam^m lang^m/$, in passive voice to ถูก $/thuuk^l/$, and succeeding with adjectival phrase to translation omission. Another example is the verbal phrase *to look for* ***something***. It must be translated to มองหา $/m\omega\omega ng^m haa^r/$ not to มองสำหรับ $/m\omega\omega ng^m\ sam^r rab^l/$. The word *look* is translated to มอง $/m\omega\omega ng^m/$, and *for* to สำหรับ $/sam^r rab^l/$.

From empirical observation, we found that the PTB tag set is rather problematical to translate into Thai. We hence implement the parse tree modification process in order to relieve the complexity of transformation process (Trujillo, 1999). In this process, the heads of the tree are recursively modified so as to facilitate phrase translation. A portion of parse tree modification rules shown on Table 1 is described in parenthesis format.

Obviously, from Table 1, we can more easily compose the rules in Table 2 to translate the verb *to be* and the phrasal verb *look for* ***something***.

2.3 Lexicon Rearrangement

In English, we can normally modify a certain core noun with modifiers in two ways—putting them in front of or behind it. We will focus the first case in this paper. The problem occurs as soon as we would like to translate a sequence of nouns and a sequence of adjectives. The first case is translated backwards, while the second forwards. An example for this problem is that "she is a beautiful diligent slim laboratory member" is translated to เธอเป็นสมาชิกแล็บที่สวยขยันผอม $/th\epsilon\epsilon^m\ pen^m\ sa^l maa^m chik^h\ thii^f\ suay^r\ kha^l yan^r\ ph\omega\omega m^r/$. The word *she* is translated to เธอ, *is* to เป็น, *member* to สมาชิก, *laboratory* to แล็บ, *beautiful* to สวย, *diligent* to ขยัน, and *slim* to ผอม.

With the purpose to solve this problem, we first group nouns and adjectives into groups—NNS and ADJS—and we apply a number of structural transfer rules. Table 3 shows a por-

Table 1: Rules for the parse tree modification process

Original PTB	**Modified**
(VP (AUX (be)) (NP))	(VP (be) (NP))
(VP (AUX (be)) (PP))	(VP (be) (PP))
(VP (AUX (be)) (VP (VBG) *))	(VP (be) (VBG) *)
(VP (AUX (be)) (VP (VBN) *))	(VP (be) (VBN) *)
(VP (AUX (be)) (ADJP))	(VP (be) (ADJP))
(VP (VBP (look)) (PP (IN (for)) (NP)))	(VP (look) (for) (NP))

Table 2: Rules to translate the verb *to be* and the verbal phrase *look for* ***something***

English Rules	**Thai Rules**
VP → *be* NP	VP → เป็น NP
VP → *be* PP	VP → อยู่ PP
VP → *be* VBG	VP → กำลัง VBG
VP → *be* VBN	VP → ถูก VBN
VP → *be* ADJP	VP → ADJP
VP → *look up* NP	VP → มองหา NP

tion of transfer rules.

Table 3: A portion of structural transfer rules to solve the lexicon reordering

English Rules	Thai Rules
NP → ADJS NNS	NP → NNS ADJS
ADJS → *adj*	ADJS → *adj*
ADJS → *adj* ADJS	ADJS → *adj* ADJS
NNS → *nn*	NNS → *nn*
NNS → *nn* NNS	NNS → NNS *nn*

2.4 Classifier Generation

The vital linguistic divergence between English and Thai is head-noun-corresponding classifiers (Lamduan, 1983). In English, classifiers are never used in order to identify the numeric number of a noun or definiteness. On the contrary, classifiers are generally used in Thai—for example, in English, a number precedes a noun phrase; but in contrast, a classifier together with the number succeeds in Thai.

In order to generate a classifier, we develop the *classifier matching algorithm*. By empirical observation, it is noticeable that the head noun in the noun phrase always indicates the classifier. For example, supposing the rules in Table 4 are amassed in the linguistic knowledge base.

Table 4: An example of rules for classifier generation

Head Noun	Classifier
รถ $/roth^h/$	คัน $/khan^m/$
รถไฟ $/roth^h fai^m/$	ขบวน $/kha^l buan^m/$

Thus, we can revise "รถไฟเหาะตีลังกา $/roth^h fai^m \ h\omega^l \ tii^m lang^m kaa^m/$ 3 <cl>" and "รถยนต์ $/roth^h yon^m/$ 4 <cl>" can be respectively revised to "รถไฟเหาะตีลังกา 3 ขบวน" (three roller coasters) and "รถยนต์ 4 คัน" (four automobiles). If there is no rule that can match the noun phrase, its head noun is used as the classifier (Lamduan, 1983)—for example, ประเทศ

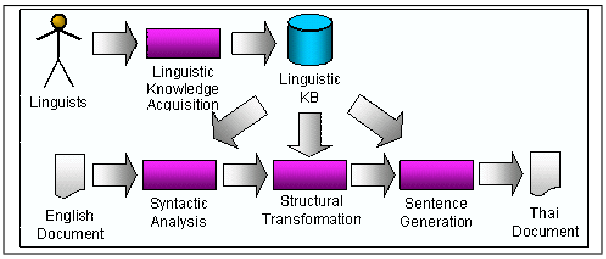

Figure 1: System Overview

$/pra^l thees^f/$ is a Thai word that there is, in fact, no corresponding classifier. As soon as we would like to specify, as the latter example, the numeric number, we say "ประเทศพัฒนาแล้ว 3 ประเทศ" $/pra^l thees^f \ phath^h tha^h naa^m \ l\alpha\alpha w^h/$ (three developed countries).

3 System Overview

As illustrated in the Figure 1, the system comprises of four principle components—syntactic analysis, structural transformation, sentence generation, and linguistic knowledge acquisition.

3.1 Syntactic Analysis and Parse-Tree Modification

In this process, we analyze each sentence of the source documents with the Charniak Parser and afterwards transform each of which into a parse tree.

The first process that we have to accomplish first is the sentence boundary identification. In this step, we require users to manually prepare sentence boundaries by inserting a new-line character among sentences.

The next step is the sentence-parsing process. We analyze the surface structure of a sentence with the Charniak Parser. In this case, the original Charniak Parser nevertheless spends long time for self-initiation to load its considerably huge database. Consequently, we patched it to be a client-server program so as to eliminate such time.

As stated earlier, in the view of the fact that

parse trees generated by the Charniak Parser are quite complicated to translate into Thai, we therefore implement the parse tree modification process (see Section 2.2).

3.2 Structural Transformation

This process performs recursive transformation from the source language parse trees into a set of corresponding Thai translation parse trees with their probabilities. As stated earlier, there are some complexity in order to transfer a PTB-formatted parse tree into Thai, we thus implemented the parse tree modification process (see Section 2.2) before performing transformation. The transformation relies on the transformation rules from the linguistic knowledge base.

A single step of transformation process matches the root node and single-depth child nodes with the transformation rules and afterwards returns a set of transformation productions. As stated earlier, we embedded the probability of each rule. The probability of a parse tree π is given by the equation

$$P(\pi) = \delta(c_\pi, c_{\pi_1}, c_{\pi_2}, c_{\pi_3}, \ldots, c_{\pi_n}) \prod_{k=1}^{n} P(\pi_k)$$

where π_k is the k-th subtree of the parse tree π whose number of member subtrees is n, c_π represents the constituent of the tree π, and δ is a probability relation that maps the constituents of the root and its single-depth children to the probability value.

3.3 Sentence Generation

This process generates a target language sentence from the parse tree. This stage also relies on the linguistic knowledge base. The additional process is the noun classifier. We apply the methodology defined in *classifier matching algorithm* (see Section 2.4). Finally, the system will show the translations of the most possibility and let the users change each solution if they would like to do so.

3.4 Linguistic Knowledge Acquisition

We provided an advantageous tool so as to manually train new translation knowledge. Currently, it comprises of the translation rule learner and the English-Thai unknown word aligner (Kampanya et al., 2002).

In this module, the translation rule learner obtains document and analyzes that into a set of parse trees. Afterwards, the users manually teach it the rules to grammatically translate a certain tree from the source language into the target language with the rules following to the Backus-Naur Form (Lewis and Paradimitriou, 1998) (or BNF in abbreviation). This module will determine whether the rule is re-trained. If so, the module will raise the probability of that rule up. If not, it will add that rule to the knowledge base.

Moreover, the aligner is utilized to automatically update the bilingual dictionary. For our future work, we intend to develop a system to automatically learn new translation rules from our corpora.

4 Evaluation

We established the system evaluation on the *Future Magazine* bilingual corpus. We categorized the evaluation into two environments—under restricted knowledge base and under increasing knowledge base. Each of which is also categorized into two environments—with parsing errors and without parsing errors.

In the evaluation, we randomly selected 322 sentences from the corpus. In order to have a manageable task and facilitate performance measurement, we classify translation result into the following three categories—*exact* (the same as in the corpus), *moderate* (understandable result), and *incomprehensible* (obviously non-understandable result). Table 5 shows the evaluation results.

In this evaluation, we consider the results in the exact and moderate categories as reason-

Table 5: Evaluation Results (in percentages)
The column **A** represents evaluation with restricted knowledge base and with parsing errors, **B** as with restricted knowledge base but without parsing errors, **C** as with increasing knowledge base but with parsing errors, and **D** as with increasing knowledge base and without parsing errors.

Categories	**A**	**B**	**C**	**D**
Exact	3.97	4.41	4.97	5.52
Moderate	55.90	62.04	69.88	77.56
Incomprehensible	40.13	33.55	25.15	16.92
Accuracy	59.87	66.45	74.85	83.08

able translations. Moreover, we also consider that the evaluation with restricted knowledge base and with parsing errors is the worst case performance, and the evaluation with increasing knowledge base and without parsing errors is the best case performance.

From the constraints we established, we found that the system yielded the translation accuracy for 59.87% for the worst case and 83.08% for the best case.

5 Conclusions

In this paper, we propose novel Internet-based translation assistant software in order to facilitate document translation from English to Thai. We utilize the structural transfer model as the translation mechanism. This project differs from the current MT systems in the point that the users have a capability to manually select the most appropriate translation, and they can, in addition, teach new translation knowledge if it is necessary.

The four translation problems—Lexicon Rearrangement, Structural Ambiguity, Phrase Translation, and Classifier Generation—are accomplished with various methodologies. To resolve the lexicon rearrangement problem, we compose a number of structural transfer rules. For the structural ambiguity, we apply the statistical method by embedding probability values to each transfer rules. In order to relieve the complexity of the phrase translation, we develop the parse tree modification process to modify some tree structure so as to more easily compose translation rules. Finally, with the purpose of resolving the classifier generation problem, we define the classifier matching algorithm which matches the longest head noun to the appropriate classifier.

In the evaluation, we established the system experiment on the *Future Magazine* bilingual corpus and we categorized the evaluation into two environments—under restricted knowledge base and under increasing knowledge base. From the evaluation, the system yielded the translation accuracy for 59.87% for the worst case and 83.08% for the best case.

References

Peter F. Brown, John Cocke, Stephen Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

Eugene Charniak. 1997. Statistical parsing with a context-free grammar and word statistics. In *AAAI/IAAI*, pages 598–603.

Eugene Charniak. 1999. A maximum-entropy-inspired parser. Technical Report CS-99-12, Brown Laboratory for Natural Language Processing.

Nithiwat Kampanya, Prachya Boonkwan, and Asanee Kawtrakul. 2002. Bilingual unknown word alignment tool for english-thai. In *Proceedings of the SNLP-Oriental COCOSDA*, Specialty Research Unit of Natural Language Processing and Intelligent Information System Technology, Kasetsart University, Bangkok, Thailand.

Ronald M. Kaplan, Klaus Netter, Jürgen Wedekind, and Annie Zaenen. 1989. Translation by structural correspondences. In *Pro-*

ceedings of the 4th. Annual Meeting of the European Chapter of the Association for Computational Linguistics, pages 272–281, UMIST, Manchester, England.

Somchai Lamduan, 1983. *Thai Grammar (in Thai)*, chapter 4: Parts of Speech, pages 128–131. Odeon Store Publisher.

Harry R. Lewis and Christos H. Paradimitriou, 1998. *Elements of the Theory of Computation*, chapter 3: Context-Free Grammar. Prentice-Hall International Inc.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1994. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330.

Virach Sornlertlamvanich. 2000. The state of the art in thai language processing.

Arturo Trujillo, 1999. *Translation Engines: Techniques for Machine Translation*, chapter 6: Transfer MT, pages 121–166. Springer-Verlag (London Limited).