ANOTHER DECADE OF THAI LANGUAGE PROCESSING RESEARCH

Virach Sornlertlamvanich
Chief
Linguistics and Knowledge Science Laboratory
National Electronics and Computer Technology Center
THAILAND

ABSTRACT

This paper describes lexicon and corpus base construction that supports the development of multi-lingual machine translation system. The interlingua method yields some particular problems that obstruct the improvement of translation result. It is because we have to define a set of meaning to be an intermediate representation for matching words between languages. The difficulties occur in cross-linguistic divergences and mismatches. Here, we will focus on the problems that occur in defining unit of meaning and surface form selection in generation process. The first case is that the coverage of a unit of meaning for different language is not necessary to be the same. One word in a language may be defined to have a wider sense than in the other. In this case, we propose a set of operation on concept to realize concept composition and decomposition procedure. The latter case occurs because an interlingua for a sentence represents the meaning that it would be extracted. Surface information is translated into concepts, relations or the structure. Therefore, appropriate surface word selection is another important task in the process of sentence generation. A lot of words are culturally or conventionally used. In case of Thai classifier usage, it is hard to explain why a word is more desirable than the other when there are more than one candidate. Hence, we propose a statistical method for classifier assignment.

LEXICON

Lexicon is created to support multilingual machine translation system based on the interlingual approach. There are 2 bases of lexicon constructed according to the area of word utilization. There are 50,000 entries for basic term and another 25,000 entries for technical term (information processing area). The lexicon is designed to be accessible through more than one index type because most of the word information can be shared in both analysis and generation modules. The advantages of having only one lexicon base that can serve both sentence analysis and generation modules are memory minimality and maintainability.

Generation module generates a sentence in the reverse way that the analysis module does. Case-mapping provides the information for mapping of syntactic and semantic case structurally [3]. Generation module assigns word positions corresponding to the syntactic case when the case is clarified through the part of speech disambiguation process. Once, the word information is updated, it gives effect to both the way analysis and generation does. The counter checking of the rules of both sides always keep the consistency of word information in the lexicon.

Each record is defined to include the information as shown in figure 1.

MORPH	::= <thai>,<concept-ID>,<english>
SYNTAX	::= <word-type>,<pos>,<vp>,<classifier>, <case-mapping>
SEMANTIC	::= <ako>,<similarity>
OPERATION	::= <constraintp>,<constraintc>,<prune>, <pruneall>,<replace>

Figure 1. Record information

The lexicon is designed not only for the use in machine translation, but we are also concerned in the maintenance to keep it updated and to widen its use. Therefore, a record of word is composed of morphological information (MORPH) that provides a word form of Thai and a concept ID. Every word sense has a unique concept ID and description according to EDR's definition [6]. The concept ID provides a sense definition to link the word forms of languages. Syntactic information (SYNTAX) includes <pos> (part of speech), <vp> (verb pattern for a verb), <classifier> (co-occurring classifier for a noun) and <case-mapping>. The relation of the meaning of word is defined in a hierarchy of concept, <ako> (a-kind-of) [1].

Operation on Concept

Quality of translation can generally be improved by implementing the following major functions [5]:

1. selection of equivalents for word
2. reordering of words, and
3. improvement of sentence styles.

Besides the improvement of grammar rules, the linking between lexicon is another effective theme to improve the translation result. The interlingual approach allows various representations for a sentence. It is theoretically true because a unit of meaning representation for a language is not always the same as the other. The defined concept is not necessary to be a primitive concept. A concept may be composed of more than one conceptual unit. This property of a concept allows a word of a language to occupy a larger unit of concept than the other.

E: It rains.

T: /fon/ /tok/

(rain) (drop)

E: to whiten *st.*

T: /tham/ /hai/ *st.* /kwau/

(cause) (white)

Some compound concepts have complicated structures. They need restructuring operator to change the original structure to a structure that contains the concepts corresponding to the target language. The operators for concept composition and decomposition are prepared as follows:

<constraintp>	: constraint on parent concept
<constraintc>	: constraint on child concept
<prune>	: drop the determined child branch
<pruneall>	: drop all the child branch

<replace>	: compose with the determined child concept

Figure 2. Concept operator

Parent and child concept are the structural constraint. A concept within a structure is specified by its parent and/or child concept together with the relation. In English, “whiten” means to cause to become more white, in Thai we have to express in different concept structure. The concept of “whiten” is decomposed to “cause” and “white”.

E: Toothpaste whitens the teeth.

(c#toothpaste <-agt- c#whiten –obj-> c#tooth)

T: Toothpaste causes the teeth to be white.

(c#toothpaste <-agt- c#cause –obj-> (c#white –a-obj-> c#tooth))

In this case, c#whiten is replaced with (c#cause –obj-> c#white) by the <replace> operation. Then the structure is reconstructed according to the case requirement of c#white.

TAGGED CORPUS BASE

There are very few electronic text data available for Thai language study. Many laboratories make their own text base for a specific use. The text bases then are very small and lack of information. To have a large enough data for system evaluation and statistically study, we started collecting Thai text data and then marked them up with a structure that stores some addition information for language study. At present, the structure of text is defined locally. It will be extended to some standard in the future.

Text Base Structure

We classified text data into 15 categories such as article, book, proceedings, etc. The area of data is limited to computer science so that we do not need to include the field information of the data in the present text base. The bibliographic data is marked according to the syntax shown in figure 3.

```
@entry_type{key,
        <Required fields>
                field_name = {field text}, field_name = {field text}, .....
        <Optional fields>
                field_name = {field text}, field_name = {field text}, .....
        <Ignored fields>
                field_name = {field text}, field_name = {field text}, .....
        }
```

Figure 3. General syntax for entry type

We marked up the text data with a stochastic part-of-speech tagger. In general Thai text, there is no explicit marker for word boundary as well as for sentence boundary. Words are written continuously. Spacing is occasionally used to break a listing of word or between phrases, clauses or sentences to make ease the reading in some cases.

Therefore, we have to preprocess the text with word segmentation procedure. The word segmentation program employs heuristic rules of longest matching and least word count incorporated with character combining rules [2]. The accuracy is as high as 95.91% based on

the tested result of 8,428 words in text. Part of speech tagging is done semi-automatically with the part of speech selecting program. The part of speech is selected from the lexicon base.

Extraction of Noun·Classifier Collocation

In Thai language, there are several usages of noun classifier. The classifier plays an important role in construction with noun to express ordinal, pronoun, for instance. The classifier phrase is syntactically generated according to a specific pattern such as, N-CL-DET for referential expression, CL-N for noun modification expression, etc. [4]. We extract the collocation of noun and classifier to make a table called Noun Classifier Association (NCA) as shown in figure 4.

(คณะกรรมการ_111, คณะ_2, 11)	(ทหาร_111, นาย_1, 9)
(คณะกรรมการ_111, กลุ่ม_2, 5)	(ทหาร_111, ฝ่าย_2, 1)
(คณะกรรมการ_111, คน_1, 6)	(คนงาน_111, คน_1, 6)
(นก_13111, ตัว_1, 9)	(ส้ม_13114, ลูก_1, 12)
(นก_13111, ฝูง_2, 4)	(ส้ม_13114, ผล_1, 3)
(ไก่_13111, ตัว_1, 10)	(แตงโม_13114, ลูก_1, 8)
(ไก่_13111, เล้า_2, 3)	(ทุเรียน_13114, ลูก_1, 9)
(นกกระจอก_13111, ตัว_1, 7)	(โค_13111, ตัว_1, 7)
(คน_111, คน_1, 67)	(หมา_13111, ตัว_1, 13)
(คน_111, กลุ่ม_2, 1)	(หมู_13111, ตัว_1, 5)
(ทหาร_111, คน_1, 17)	(ช้าง_13111, เชือก_1, 3)

Figure 4. Table of Noun Classifier Association (NCA)

The associations above are useful for determining a proper classifier for a given noun. For a noun occurring in the corpus, alternative determination is accomplished in a straightforward manner by using its associated representative classifier which occurs in the corpus more frequently than any other classifiers. In the other case where the given noun does not exist in the corpus, the determination is done by using the representative classifier of its class in the concept hierarchy.

Semantic class	Unit classifier	Collective classifier
animal	ตัว_1	ฝูง_2
human	คน_1	คณะ_2
plant	ต้น_1	-
fruit	ลูก_1	-

Figure 5. NCA for representative classifier

Unit classifier	Collective classifier
(1) นักเรียน คน ที่ สี่ /nakrian kon tii sii/	(3) คณะกรรมการ คณะ นั้น /kanagammagarn kana nan/

student <student> number four	committee group that
(2) แอปเปิ้ล ลูก ไหน	(4) กางเขน ฝูง นั้น
/appern luuk nai/	/gangken fuung nan/
apple <apple> which	magpie group that

(1) and (3) show the case of nouns appearing in the corpus, while (2) and (4) show a different scenario In (2), the unit classifier of /appern/ is obtained by using the representative unit classifier of its class `fruit' which is ลูก_1 /luuk/ according to figure 5. Similarly, in (4), the collective classifier of /gangken/ is determined by the representative collective classifier of its class `animal' which is ฝูง_2 /fuung/.

DISCUSSION

Lexicon and text data base will be another great resources. Many language phenomena of Thai are still obscure. Many questions are left unanswered; What is a word? What is a sentence? Shall we use punctuation marks in Thai sentence? These cause difficulties in processing Thai electronically as well as in human communication, especially in the technical field or high-end research. We need new regulation that may result from the study of practical language data. Stochastic method based on a large scale data base is another solution for the new era of Thai language processing. We plan to increase the amount of the data base and also prepare the data in the most accessible way. In this paper, we depicted the effectiveness in using NCA to assign the appropriate classifier. This is just one of the valuable utilization of the large language database.

ACKNOWLEDGMENT

We wish to thank the National Electronics and Computer Technology Center (NECTEC) and Center of the International Cooperation for Computerization (CICC) who facilitated the research during the whole project. Many participated researchers from Chulalongkorn University, Kasetsart University, King Mongut's Institute of Technology Ladkrabang and Thonburi provided the great contribution. The project will never be successful if it is without their exertion to the utmost and the unlimited support from the colleagues of LINKS laboratory.

REFERENCES

[1] Muraki, K., Sornlertlamvanich, V., et al. (1989), 'Thai Dictionary for Multi-lingual Machine Translation System' **Computer Processing of Asian Languages (CPAL)**, 211-220.

[2] Sornlertlamvanich, V. (1993), 'Word Segmentation for Thai in Machine Translation System', **Machine Translation**, National Electronics and Computer Technology Center, (in Thai).

[3] Sornlertlamvanich, V., Phantachat, W. (1993), 'Information-based LanguageAnalysis for Thai', **Asean Journal on Science & Technology for Development**, Vol.10 No.2, 181-196.

[4] Sornlertlamvanich, V., Phantachat, W., Meknavin S. (1994), 'Classifier Assignment by Corpus-based Approach', **Proceedings of COLING 94**, 556-561.

[5] Tsutsumi, T. (1990), 'Wide-Range Restructuring of Intermediate Representation in Machine Translation', **Computational Linguistics** Vol. 16 No.2, 79-88.
[6] Yasuhara, H. (1993), 'Text Compiler and Concept-Tagged Corpus', **Proceedings of KB & KS 93**, 251-256.