Building A Large Thai Text Corpus
- Part-Of-Speech Tagged Corpus: ORCHID -

Thatsansëe Charoenporn\textsuperscript{1,2}

Virach Sornlertlamvanich\textsuperscript{1,3}

Hitoshi Isahara\textsuperscript{4}

\textsuperscript{1}Linguistics and Knowledge Science Lab.,
National Electronics and Computer Technology Center, Thailand.

\textsuperscript{2}Dept. of Linguistics,
Chulalongkorn University, Thailand.

\textsuperscript{3}Dept. of Computer Science
Tokyo Institute of Technology, Japan.

\textsuperscript{4}Intelligent Processing Section
Communications Research Laboratory
Ministry of Posts and Telecommunications, Japan.

E-mail: thatsce@nw.nectec.or.th, virach@cs.titech.ac.jp, isahara@crl.go.jp

Abstract
This paper presents a procedure in building a Thai part-of-speech (POS) tagged corpus named ORCHID. It is a collaboration project between Communications Research Laboratory (CRL) of Japan and National Electronics and Computer Technology Center (NECTEC) of Thailand. We proposed a new tagset based on the previous research on Thai parts-of-speech for using in a multi-lingual machine translation project. We marked the corpus in three levels: paragraph, sentence and word. The corpus keeps text information in text information line and numbering line, which are necessary in retrieving process. Since there are no explicit word/sentence boundary, punctuation and inflection in Thai text, we have to separate a paragraph into sentences before tagging the POS. We applied a probabilistic trigram model for simultaneously word segmenting and POS tagging. Rule for syllable construction is additionally used to reduce the number of candidates for computing the probability. The problems in POS assignment are formalized to reduce the ambiguity occurring in case of the similar POSs.

1. Introduction
Thanks to the present availability of Thai electronic texts from various sources, the collection of the text data can be done in a considerable time. Studying a language with a data oriented approach needs a huge amount of real text data. Idealistically, the collected text data have to be free from any kinds of errors of such, word spelling errors, syntactic errors, semantic errors, etc. depending on the purpose of the designed corpus. It inevitably includes such kinds of errors when dealing with a huge amount of data. In building Thai text corpus, we therefore, need a post-editing process by human after obtaining the output from each module. To tag Thai text with parts-of-speech (POS), we have to separate a paragraph into sentences and then into words before assigning POS to each word. Since there are no punctuation and inflection in general Thai text at all, the tasks of sentence segmentation and word segmentation are not less crucial than word tagging. This paper describes the design and procedure in building Thai POS tagged corpus named ORCHID. Our tagger considers word segmenting and POS tagging together within a probabilistic model.

ORCHID is the code name of the project for building Thai POS tagged corpus initiated by a group of researchers from Communications Research Laboratory (CRL) of Japan and National Electronics and Computer Technology Center (NECTEC) of Thailand. The project started in April 1996 with the purpose of preparing Thai language corpus for language study and application research especially for developing applications for processing Thai language under the computational environment. In the first year of the project, we collected about two million characters of plain texts. It is discussible for defining lexical unit for a Thai text because of the absence of the explicit use of word boundary, displaying the scale of text is suitably done in number of characters rather than words or sentences. The simply counting number of words by considering a space character as a word boundary can mislead the reader in the sense that a string between two space characters is not always identical to a unit of word in case of a compound for example. The collected texts are technical papers appeared in the past six years of the proceedings of the National Electronics and Computer Technology Center (NECTEC) annual conferences. A text is separated into sentences manually guided by our own standard of sentence structuring. Word boundaries are determined by a trigram probabilistic model trained by a set of preliminary POS tagged data \cite{1,2,5}. Since the training set for the probabilistic model is very limited, the accuracy of the output from the automated procedure is still problematic and it requires some human correction.
In this paper, we propose a procedure in building Thai POS tagged corpus with the design of data structure and the POS. We apply our POS developed from the set which had been used in developing the multilingual machine translation system in the co-operation project with Japan and other three Asian countries [3], as the tagset used in tagging words in the text. The original 45 POSs are carefully revised to be able to cover all roles of the words used in the real text. As a result, the new 47 POSs are defined as the tagset. Section 2 discusses the word class for real text. As a result, the new 47 POSs are defined as the tagset. We apply our POS developed Thai POS tagged corpus with the design of data structure and the POS. We revised the Thai language has no inflection and most of the compound words are created from simply combining two or more small units of word. We found that the difficulty in tagging occurs because of the unchanging of lexical form though the word is used in a different position or role in a sentence. We classify some problematic tagging as the guidance for making decision as follows.

2. Word Class

We firstly developed our word class or parts-of-speech to classify words according to their syntactic roles and implemented in a dictionary used in a machine translation system [4]. The parts-of-speech contains 13 categories that are subcategorized into 45 subcategories. They are essentially used in both analysis and generation modules. We revised the original parts-of-speech by observing a lot more text data. As a result, we redefined some parts-of-speech to clarify the ambiguous parts and set up a new set of 14 categories with 47 subcategories included. The significant changes are the subcategories for the classifier (CLAS) and prefix (FIXP). We subcategorized the original CLAS into 5 subcategories and FIXP into 2 subcategories. Classifier plays an important role in constructing phrases in Thai language, see [6] for detail discussion. Therefore, we subcategorized the CLAS to help in disambiguating the structures of the phrases.

Another modification is done on the FIXP, which is our attempt to support the construction of noun phrase and adverb phrase that are ambiguous because of the absence of word inflection in changing the syntactic role.

The sentence (2) is still valid and has the equivalent meaning to the sentence (1) though the underlined FIXN is absent. From the above sentences, we may define “กรุณา” either a word means “taking a rest” or two words of “กรุณา” (a nominal prefix) and “หลับ” (to rest). If we define it as one word, it is unsuitable to accept the sentence (2) by assigning “กรุณา” as a verb paralleling with the noun “การนอนหลับ”.

We used the 47 subcategories as the tagset for POS tagging in ORCHID corpus.

3. Problematic Tagging

Thai language has no inflection and most of the compound words are created from simply combining two or more small units of word. We found that the difficulty in tagging occurs because of the unchanging of lexical form though the word is used in a different position or role in a sentence. We classify some problematic tagging as the guidance for making decision as follows.

3.1 Verb and Preposition

There are a lot of prepositions having the same lexical form as verbs and sometimes hardly making the distinction between them. Followings are the additional guides for making the decision.

a) A preposition cannot be negated, but a verb can.

b) A preposition can be tested by moving the phrase around. A preposition always goes together with the following noun, but not a verb.

For example,

\[
\text{กรุณา} \quad \text{กินอาหาร} \quad \text{ข้าว} \quad \text{ข้าว} \quad \text{เป็น} \quad \text{การ} \quad \text{เตรียม} \quad \text{อาหาร} \quad \text{ข้าว} \quad \text{ข้าว} \quad \text{เป็น} \quad \text{การ} \quad \text{เตรียม} \quad \text{อาหาร} \quad \text{ข้าว}
\]

We use “การ” as a preposition and “เป็น” as a verb.

3.2 Adverb and Preposition

In general, adverbs can be placed more freely than prepositions. There is no any strict rule for discriminating the two categories. But with some noticeable use of a preposition with the following noun, it is recommended to consider for a preposition at first, as to the criteria in 3.1. For example,

\[
\text{กรุณา} \quad \text{กิน} \quad \text{ข้าว} \quad \text{ข้าว} \quad \text{เป็น} \quad \text{การ} \quad \text{เตรียม} \quad \text{อาหาร} \quad \text{ข้าว} \quad \text{ข้าว} \quad \text{เป็น} \quad \text{การ} \quad \text{เตรียม} \quad \text{อาหาร} \quad \text{ข้าว}
\]

We use “กรุณา” as a preposition and “เป็น” as a verb.

3.3 Verb and Verbal Classifier

The classifiers, which are classified into the verbal classifier (CVBL), are the classifiers derived from verbs or having the same lexical form as verbs. Classifiers are used in the very rigid patterns as
discussed in [6]. Most of the classifiers can be determined by checking with the possible patterns which verbs cannot be conformed. For example,

(7) ติดหนังสือกิจกรรมสัปดาห์ CICL = Classifier ที่ หนังสือกิจกรรมสัปดาห์ CICL = Verb

3.4 Verb and Auxiliary
Verbs and auxiliaries can have the same lexical form in many cases. In Thai language, there are mainly two types of auxiliary classified by their positions relating to the verb of the sentence. It is possible to negate both verbs and auxiliaries therefore, it is recommended to tag as verb if there is no other candidate for being the main verb of the sentence. For example,

(8) อาจทำสิ่งดิบๆได้แต่ไม่ได้ทำสิ่งดิบๆ หน้า = Verb แต่ = Auxiliary
(9) ยังไม่ได้ทำสิ่งดิบๆ แต่ = Auxiliary หน้า = Verb

3.5 Verb and Adverb
Verbs or adverbs sometimes make us confused, especially when such verbs and adverbs are in the same lexical form. For example,

(10) เดินทางไปรักษาตัวที่ ต่างประเทศ ที่ = Classifier ต่างประเทศ = Noun
(11) เดินทางไปรักษาตัวที่ ต่างประเทศ ที่ = Classifier ต่างประเทศ = Noun

"ที่" can be either a verb (VACT) or an adverb (ADVN). There is no problem in (14) because there is no other verb in the sentence then "ที่" must be the verb to make a sentence. In (12) and (13), there is a verb "ที่" and "ที่" can better be interpreted as a modifier to the verb to make the meaning concisely. For example, it is more concise to interpret (12) as “He walks straight to school” by considering “ที่” as an adverb rather than “He walks and directs to school” by considering “ที่” as a verb.

3.6 Nominalization
The process of forming a noun or a noun phrase from other POS frequently occurs in Thai language. Word in Thai can be nominalized by adding a prefix “ที่” or “ที่” (FIXN) before a root word. But it is often difficult to differentiate the case whether it is a nominalized noun or a nominalized noun phrase. We, thus, proposed to consider the nominalized noun or noun phrase as a composition of a prefix with the following noun or noun phrase. As a result, the decompositional consideration of the nominalized noun always gives the consistent solution in the interpretation. For example,

(15) เรียนการตั้งตัวแปรในภาษาไทย RSYR = Classifier ภาษาไทย = Noun ที่ = Classifier
(16) เรียนการตั้งตัวแปรในภาษาไทย RSYR = Classifier ภาษาไทย = Noun ที่ = Classifier

3.7 Noun and Classifier
In case that a common noun and its classifier have the same lexical form, we can easily get confused because a noun and a classifier usually occur in the similar pattern. In this case, we use the following testing templates of which a noun and its classifier may occur relating to some types of determiner.

a) Noun Classifier DDAC
b) Noun DDAN
c) Noun DCNM Classifier
d) Classifier DONM
e) X DDAC [X is a classifier if it has a form of classifier else it is a noun.]

For example,

[เรียนการ] is a classifier because it can be either a noun or a classifier.
[เรียนการ] is a noun because it can only be a noun.
[เรียนการ] is a classifier because it can only be a classifier.

3.8 Common Noun (NCMN) and Proper Noun (NPRP)
NCMN is a class of entity but not an individual while NPRP is a particular person, place, organization, institution, painting or unique thing, and usually not to be referred to by its meaning. There is no distinction in its lexical form between NCMN and NPRP, such as beginning with a capital letter for a proper noun in English. We then, add the following remarks for tagging a noun as NPRP.

a) Name of product
b) Abbreviation name
c) Name of person, group of person, company
d) Geographical name, such as name of region, continent, country, province, etc.
e) Astronomical name
f) Chemical name
g) Scientific name
3.9 DCNM, DONM, NLBL, ADV in Ordinal and Quantitative Expression

DCNM and DONM are classified by the following test frames:

a) DCNM X Classifier
b) DCNM Classifier X

If a cardinal number (a figure or a word) occurs between a noun and a classifier, it is assigned as DCNM. If an ordinal number (a word or a figure preceding with “th”) occurs after a classifier, it is assigned as DONM. For example,

\[ \text{C4} \text{ DONM - Classifier X } \text{N} \]
\[ \text{C4} \text{ Classifier X - DONM} \]

It is notable that sometimes the classifier between noun and the ordinal number (DONM) can be omitted if it has the same lexical form as its noun. Besides the ordinal number can be assigned as DONM, there is a set of ordinal expression that is assigned as DONM. The ordinal expressions are ฤๅ (one), ฤๅฤๅ (one), ฤๅฤๅฤๅ (first), ฤๅฤๅฤๅฤๅฤๅ (last), ฤๅฤๅฤๅฤๅฤๅฤๅฤๅ (middle) and ฤๅฤๅฤๅฤๅฤๅฤๅฤๅฤๅ (last). For example,

\[ \text{C5} \text{ DONM - Classifier X} \text{N} \]
\[ \text{C5} \text{ Classifier X - DONM} \]

However, the ordinal expression can function as an adverb since it modifies the verb. The ordinal expressions in the following cases are all assigned as an adverb of the sentence. For example,

\[ \text{C6} \text{ DONM - Classifier X} \text{N} \]
\[ \text{C6} \text{ Classifier X - DONM} \]

3.10 Classifier Expression

Besides the general use of classifier in the construction of quantitative expression, relative pronoun, demonstrative noun, etc.[6], we introduce a classifier to construct some types of verb or noun modifier (adverb or adjective phrase). A classifier preceding a verb or a noun forms an adverb or adjective phrase consequently. Followings are some examples of the construction. For example,

\[ \text{C7} \text{ DCNM - Classifier X} \text{N} \]
\[ \text{C7} \text{ Classifier X - DCNM} \]

4. Conclusions

ORCHID is the very first project to build Thai POS tagged corpus. It is not limited to Thai language and the POS tagged corpus. We plan to extend our technology to other similar languages that share the similar language features and include other information to the corpus such as syntactic tree bracketing, semantic information, etc. Based on the first created corpus, we hope that we can study and know more about Thai language with the corpus based approach.

This paper revised the first version of Thai part-of-speech used in developing the multi-lingual machine translation system and applied it to the wider range of the real Thai text. It is not finalized but it can somehow cover all parts of the text we have in hand at present. It is proved to have the widest coverage for POS assignment. The POS confirmed by the real text is another crucial target of the building of the ORCHID corpus.

References