CLASSIFIER ASSIGNMENT BY CORPUS-BASED APPROACH

Virach Sornlertlamvanich **Wantanee Pantachat** **Surapant Meknavin**

Linguistics and Knowledge Science Laboratory
National Electronics and Computer Technology Center
National Science and Technology Development Agency
Ministry of Science Technology and Environment
22nd Gypsum Metropolitan Tower,
539/2 Sriayudhya Rd., Bangkok 10400, Thailand
{virach,wantanee,surapan}@nwg.nectec.or.th

Abstract

This paper presents an algorithm for selecting an appropriate classifier word for a noun. In Thai language, it frequently happens that there is fluctuation in the choice of classifier for a given concrete noun, both from the point of view of the whole speech community and individual speakers. Basically, there is no exact rule for classifier selection. As far as we can do in the rule-based approach is to give a default rule to pick up a corresponding classifier of each noun. Registration of classifier for each noun is limited to the type of unit classifier because other types are open due to the meaning of representation. We propose a corpus-based method (Biber,1993; Nagao,1993; Smadja,1993) which generates Noun Classifier Associations (NCA) to overcome the problems in classifier assignment and semantic construction of noun phrase. The NCA is created statistically from a large corpus and recomposed under concept hierarchy constraints and frequency of occurrences.

Keywords: Thai language, classifier, corpus-based method, Noun Classifier Associations (NCA)

1. Introduction

A classifier has a significant use in Thai language for construction of noun or verb to express quantity, determination, pronoun, etc. By far the most common use of classifiers, however, is in enumerations, where the classifiers follow numerals and precede demonstratives (Noss,1964). Not all types of classifier have a relationship with noun or verb as a unit classifier does.

A *unit classifier* is any classifier which has a special relationship with one or more concrete nouns. For example, to enumerate members of the class of /rya/ 'boats', the unit classifier /lam/ is selected as in the phrase below:

เรือ หนึ่ง ลำ
/rya nung lam/
boat one <boat>
'one boat'.

Other than the unit classifier, there are collective classifier, metric classifier, frequency classifier and verbal classifier.

A *collective classifier* is any classifier which shows general group or set of mass nouns, นก สอง ฝูง /nok soong fung/ 'two flocks of bird'. A *metric classifier* is any classifier which occurs in enumerations that modify predicates as well as nouns, น้ำ สาม แก้ว /nam saam kaew/ 'three glasses of water'. A *frequency classifier* is any classifier which is used to express the frequency of event that occurs, บิน สี่ รอบ /bin sii roob/ 'fly four rounds'. A *verbal classifier* is any classifier which is derived from a verb and usually used in construction with mass nouns, กระดาษ ห้า ม้วน /kradaad haa muan/ 'five rolls of paper'.

The unit classifier has a special relationship with concrete noun. The member of this class of classifier is closed for each noun. Most of the unit classifiers are used with a great many concrete nouns of very different meaning, but few are restricted to a single noun. Except for the unit classifier, the members of classifier for a noun or predicate are open. Especially for the metric classifier, the number of classifiers for numeral expression of distance, size, weight, container and value is large.

The use of classifier in Thai is not limited to the numeral expression but is extended to other expressions such as ordinal, determination, relative pronoun, pronoun, etc. The detail of each classifier phrase is described in the next section.

In many existing natural language processing systems, the list of available classifiers for each noun is attached to a lexicon base. Rules for classifier selection from the list can somehow provide the

default value but does not guarantee the appropriateness. However, the problems on classifier phrase construction still remain unsolved.

To overcome the problems of using classifiers, we propose a method of classifier phrase extracting from a large corpus. As a result, Noun-Classifier Associations (described in Section 3) is statistically created to define the relationship between a noun and a classifier in a classifier phrase. With the frequency of the occurrence of a classifier in a classifier phrase, we can propose the most appropriate use of a classifier. Furthermore, we introduce a hierarchy of semantic class for the induction of a classifier class when they are employed to construct with nouns which belong to the same class of meaning. Section 3 and Section 4 describe the generation and the implementation of the NCA, respectively.

2. The roles of classifier in Thai language

In Thai language, we use classifiers in various situations. 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. **Fig. 2.1** shows the position of a classifier in each pattern, where N stands for noun, NCNM stands for cardinal number, CL stands for classifier, DET stands for determiner, VATT stands for attributive verb, REL_M stands for relative marker, ITR_M stands for Interrogative marker , DONM stands for ordinal numeral, DDAC stands for definite demonstrative

Study on the use of classifier in each expression mentioned above, we can conclude that the types of classifier are not restricted to any kinds of expression. To consider the semantic representation of each expression, it happens that the unit classifier is not regarded as a conceptual unit in all expressions except in pattern 6, but the other types are. (see examples in a. and b.)

a) ประชาชน สอง คน
/prachachon 2 khon/
(Unit-CL)
people 2 <people>
'2 people'

b) ประชาชน สอง กลุ่ม
/prachachon 2 klum/
(Collective-CL)
people 2 <group>
'2 groups of people'

We encountered to generate the appropriate classifier for noun or verb in a semantic representation. The classifier assignment for non-conceptual representation and the classifier selection of one to many conceptual representation are over handleable by the rule-based approach. The propose on classifier assignment using the corpus-based method is another approach. Based on the collocation of noun and classifier of each pattern shown in **Fig. 2.1**, we decided to construct the Noun Classifier Association table (see Section 3). A stochastic method combined with the concept hierarchy is proposed as a strategy in making the NCA table. The table composes of the information about noun-classifier collocation, statistic occurrences and the representative classifier for each semantic class in the concept hierarchy.

3. Extraction of Noun-Classifier Collocation

In this section, we describe the algorithm used for extraction of Noun Classifier Associations (NCA) from a large corpus. We used a 40 megabyte Thai corpus collected from various areas to create the table. The algorithm is as follows:

Step 1: Word segmentation.
Input: A corpus.
Output: The word-segmented corpus.

In text processing, we often need word boundary information for several purposes. Because Thai has no explicit marker to separate words from one another, we have to preprocess the corpus with word segmentation program. We used the program developed by Sornlertlamvanich (1993) with post-editing to correct fault segmentation. The program employs heuristic rules of longest matching and least word count incorporated with character combining rules for Thai words. Though the accuracy of the word segmentation does not reach 100%, but it is high enough (more than 95%) to reduce the post-editing time.

Step 2: Tagging.
Input: Output of step 1.
Output: The corpus of which each word is tagged with a part of speech and a semantic class.

The word-segmented corpus is then processed with a stochastic part-of-speech tagger. Each word *w* together with its part of speech is then used to retrieve the semantic class of the word from a dictionary. The result yields a data structure of *(w,p,s)*, where *p* denotes the part of speech of *w* and *s* denotes the semantic class of *w*. For example, the data structure of the word นักเรียน /nakrian/ 'student' is (นักเรียน, NCMN, person), where NCMN stands for common noun and person represents นักเรียน in the class of person.

Step 3: Producing concordances.
Input: Output of step 2, a given classifier *cl*.
Output: All the fragments containing *cl*.

Expressions	Patterns	Samples
1. Enumeration	N/V-NCNM-CL	/nakrian 3 khon/ (N) (N) (CL) *student 3 <student>* 'three students'
2. Ordinal	N-CL-/tii/-NCNM	/kaew bai thii4/ (N) (CL) (N) *glass <glass> 4th* 'the fourth glass'
3. Determination -Definite demonstration	a) N-CL-DET	a) /raw chop kruangkhidlek kruang nii/ (N) (CL) (DET) *we like calculator <calculator> this* 'we like this calculator'
-Indefinite demonstration	a) N-CL-DET	a) /phukhawfung khon nung sadaeng (N) (CL) (DET) *participant <participant> one express* khwamhen nai thiiprachum/ *opinion in conference* 'A participant expressed his opinion in the conference.'
	b) N-DET-CL	b) /sunak bang tua/ (N) (DET) (CL) *dog some <dog>* 'some dogs'
-Referential	a) N-CL-DET	a) /kamakan kana nii thukkhon (N) (CL) (DET) *committee <group> this everyone* chuua waa ja thamngan samret/ *believe that will work success* 'It is this committee that everyone believed its mission would be success.'
4. Attributive	N-CL-VATT	/dinsoo theng san/ (N) (CL) (VATT) *pencil <shape> short* 'a short pencil'
5. Noun modifier	CL-N	/kana naktongtiew/ (CL) (N) *group tourist* 'a group of tourist'
6. Pronoun -Relative pronoun	a) CL-REL_M	a) /nakbanchii khon thii thamngan (N) (CL) (REL-M) (V) *accountant who work* thii borisat nii/ *at company this* 'the accountant who works at this company'
-Interrogative pronoun	b) CL-ITR_M	b) /sing nai/ (CL) (ITR-M) <*thing*> *which* 'which one'
-Ordinal pronoun	c) CL-DONM	c) / tua raek/ (CL) (DONM) *one first* 'the first one'
-Pronoun	d) CL-DDAC	d) /khon nii chop bia mak/ (CL) (DDAC) *the one like beer very* 'The one likes beer very much'

Fig. 2.1 Classification of classifier expressions table

(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)

Fig. 3.1 Table of Noun Classifier Associations (NCA)

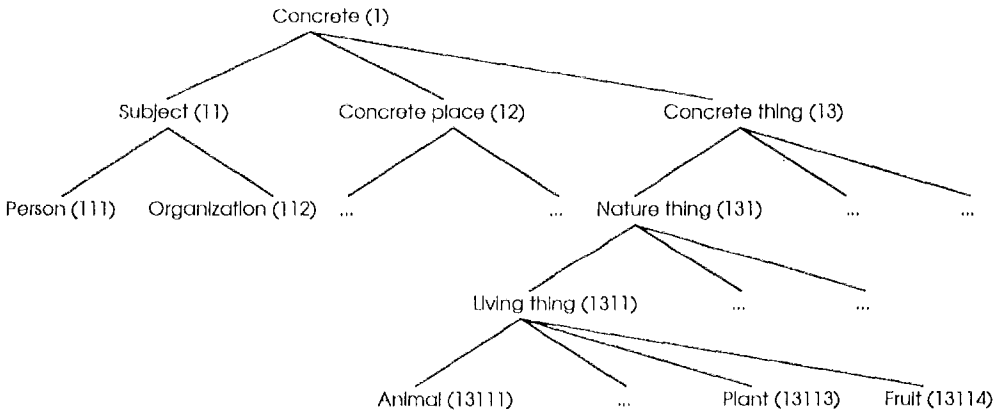

Fig. 3.2 Concept hierarchy

Instead of picking up the data sentence by sentence, we extracted a fragment of data around the *cl*, because there is no explicit marker to indicate sentence boundaries. We used the range of -10 to +2 words around the *cl* in our experiments which appeared to cover most of co-occurrence patterns.

Step 4: Pattern matching
Input: Output of step 3.
Output: A list of nouns-classifiers with frequency information of co-occurrences.

In this step, the tagged corpus is matched with each pattern of classifier occurrences shown below:

N- -NCNM-CL (Enumeration)
N- -CL- ที่ /tii/-NCNM (Ordinal expression)
N- -CL-DET (Referential expression)
N- -DET-CL
(Indefinite demonstration expression)
N- -CL-VATT (Attribute noun phrase)
CL-N (Noun modifier)
N- -CL-{ที่ /tii/, ซึ่ง /sung/, ใน /nai/,..}
(Relative/ Interrogative pronoun)

where N denotes noun, CL denotes classifier, NCNM denotes cardinal number, DET denotes determiner, VATT denotes attributive verb, ที่ /tii/, ซึ่ง /sung/ and ใน /nai/ are specific Thai words, A-B denotes a consecutive pair of A and B, and A--B denotes a possibly separated pair. Actually, A--B can be

separated by several arbitrary words but in our experiments we considered only possible separations by a relative pronoun phrase having no more than 5 words. This is to limit the search space of general cases to a manageable size with some loss of generality.

The pattern matching process was carried out one by one with each pattern. For each pattern of A- -B-C, the matching of B-C pair was simple and was performed at first. Next, the matching of a pair A- -B was done by:

1. searching for the nearest A from B. If found, mark A1.
2. from B within a span of five, searching for the nearest relative pronoun. If found, mark p1 then go to 3. Otherwise, match A1.
3. further searching for the nearest A from p1. If found, mark A2. If A2 is farther from B than A1, match A2. Otherwise, match A1.

At the end of these steps, we obtained a list of nouns *Ni* along with the frequency of *w* in the corpus for each matching pattern (see **Fig. 3.1** for sample outputs). Each entry is of the form (W_N1, CL_N2, Freq) where W denotes a noun, N1 denotes a number representing semantic class of W, CL denotes the associated classifier, N2 is a number indicating whether CL is a unit or collective classifier (1 for unit, 2 for collective) and Freq denotes the frequency of co-occurrence between W and CL. The semantic class is shown in **Fig. 3.2.**

Step 5: Determine representative classifier
Input: A list of noun-classifier with frequency information of co-occurrence.
Output: Representative classifier of each noun and each semantic class of nouns.

As it can be observed in **Fig. 3.1**, each noun may be used with several possible classifiers. In language generation process. However, we have to select only one of them. For each noun we select the classifier with the greatest value of co-occurrence frequency to be the representative classifier for both representative unit classifier and representative collective classifier. The classifier in **Fig. 3.1**, for example, will have คน_1 as the representative unit classifier and have คณะ_2 as the representative collective one for the noun คณะกรรมการ_111. Collective classifiers are used instead of unit classifiers when the notion of `group' is required.

We also find the representative classifier for each semantic class of nouns in the same manner. For each semantic class of nouns (grouped by the semantic class attached with each noun), the classifier with the greatest value of co-occurrence frequency is selected to be the representative. The classifier is used to handle the assignment of classifier to noun which does not exist in the trained corpus. For example, the representative unit classifiers for each semantic class extracted by the pattern (N- -NCNM-CL) are shown in **Fig. 3.3.**

4. Classifier Resolution

The associations as produced in the previous section 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.

Some examples of classifier determining are listed below. (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 **Fig. 3.3**. Similarly, in (4), the collective classifier of /gangken/ is determined by the representative collective classifier of its class `animal' which is ฝูง_2 /fuung/.

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

Fig. 3.3 NCA for representative classifier

Unit classifier

(1) นักเรียน คน ที่ สี่

/nakrian kon tii sii/

student <student> number four

(2) แอปเปิ้ล ลูก ไหน

/appern luuk nai/

apple <apple> which

Collective classifier

(3) คณะกรรมการ คณะ นั้น

/kanagammagarn kana nan/

committee group that

(4) กางเขน ฝูง นั้น

/gangken fuung nan/

magpie group that

5. Conclusion

The proposed approach is a significantly new method to manipulate the classifier phrase in Thai language. The fact that the expression of some syntactic constituents needs a specific classifier to be constructed with and the selection of classifier for each noun or noun phrase depends on the traditional use and the semantic class. The corpus-based approach is quite suitable for detecting the traditional use and searching for the most appropriate one when it does not exist in the corpus yet. Concept hierarchy of noun provides another path for searching when the NCA does not cover the noun in question.

In the future, this NCA will be included in the generation process of Machine Translation to solve the classifier assignment, and incorporated in the analysis process to produce a proper syntactic and semantic structure. The classifier will then be a key for pattern disambiguation when it is fixed to one of the patterns illustrated in **Fig. 2.1**.

Acknowledgement

We wish to thank the National Electronics and Computer Technology Center (NECTEC) and Center of the International Cooperation for Computerization (CICC) who provide facilities and a large corpus base for the research.

References

[1] Biber, Douglas. (1993). "Co-occurrence Patterns among Collocations: A Tool for Corpus-Based Lexical Knowledge Acquisition". Computational Linguistics, Vol. 19, No.3, September 1993.

[2] Nagao, Makato. (1993). "Machine Translation: What Have We to Do". Proceedings of MT Summit IV, June 20-22, 1993, Kobe, Japan.

[3] Noss, Richard B. (1964). Thai Reference Grammar, U.S. Government Printing Office, Washington, DC.

[4] Smadja, Frank. (1993). "Retrieving Collocations from Text: Xtract". Computational Linguistics, Vol. 19, No.1, March 1993.

[5] Sornlertlamvanich, Virach. (1993), "Word Segmentation for Thai in Machine Translation System", Machine Translation, National Electronics and Computer Technology Center, (in Thai).

Corpus-based NLP