KMITL Sci. J. Vol. 5 No. 1 Feb. 2005

ROLE ASSIGNMENT TO THAI WORD USING MULTILAYER PERCEPTRON NETWORK

Ponrudee Netisopakul[(a), (c), *] Saichon Jaiyen[(b), (c)] Peerasak Intarapaiboon[(b), (c)]

[(a)]Faculty of Information Technology
[(b)]Department of Mathematics and Computer Science, Faculty of Science
[(c)]Research Center for Communications and Information Technology
King Mongkut's Institute of Technology Ladkrabang
Bangkok 10520, THAILAND

ABSTRACT

Thai language has some certain characteristics, which lead to a number of interesting research problems in natural language processing. For instance, Thai words in a sentence do not have feature embedded form, i.e., Thai words always have the same form in any sentences. This partially explains the lack of feature information in Thai part of speech (POS). Thai POS tagging only tags word category, such as noun, pronoun, verb, preposition and so on. Category information alone is inadequate for further language processing. This paper proposes the idea of tagging word role in a sentence using multilayer perceptron as a tool. The role assignment information can be used in complimentary to category assignment in Thai language processing.

KEYWORDS: Part of speech tagging, Thai language processing, POS tagging with neural network, multilayer perceptron network, role feature tagging.

1. INTRODUCTION

Natural language processing is one of a computer science research area where one try to train a computer to process human languages. The research in this field is numerous in some popular language such as English. However, among languages diversity, processing Thai Language are very difficult because of some certain characteristics, which are quite different from some popular languages, such as English. First of all, Thai language is written without boundary between words. Therefore, word segmentation problem in processing Thai language is one of pioneer research in Thai language processing. Some work in Thai word segmentation are [1],[2]. Second, there is no mark to denote the end of a sentence in Thai language. The current status of research in Thai sentence segmentation is still an under investigation area. Third, each Thai word, which put together to form Thai sentences, unlike English, has no feature indicator. For instance, to form a sentence in English, one must concern about subject-verb agreement, singular and plural form of nouns, changing of verb form to indicate tenses, and so on. In contrast, Thai words always have the same form in any sentences they constitute. To make matters worse, when it comes to constructing a syntactic tree for a sentence, it is unclear whether there is a complete and consistent Thai grammar that is suitable for parsing Thai sentences.

* Corresponding author. Tel: 02-737-4321 Ext. 530 E-mail: ponrudee@it.kmitl.ac.th

Among the problems mentioned above, one emerging and promising area is to train neural network to recognize part of speech. Part of speech (POS) is a grammatically role of a word in a context. An example of Thai corpus with annotated POS is available from [3]. In this paper, we viewed part-of-speech tagging as a preprocessing step for constructing sentence syntactic tree. The current research in Thai POS tagging reflect the difficulties from four problems mentioned above, especially the third problem. Since Thai words in a sentence are not in a feature embedded form, Thai POS is mainly tag word's category, such as noun, pronoun, verb, preposition, and so on. Grammatically, a word in a sentence has a particular function, i.e., it plays a specific role in the sentence. The example of word's roles such as subject, direct object, indirect object, main verb modifier and so on. This research suggests that, in addition to word category, word role must also be tagged for Thai words in a sentence. The idea is realized by training multilayer perceptron network to do the role assignment to a given word in a given Thai sentence. The system is called Role-POS tagger for Thai. Section two reviews some related work that influent this research. Section three presents the concept and proposed methodology used in this work. The experimental result and discussion is given in section four together with conclusion and future work in section five.

2. RELATED WORK

Neural networks were suggested for NLP application in [4]. Three practical applications suggested are part of speech tagging, error detection in annotated corpus and self-organization of semantic map. The application of multilayer perceptron network for tagging POS was developed in [5]. The input to the network is the fixed length vector of size n where n is a total number of tags. Words are obtained from a corpus and pre-calculated for the probability of being tagged with each tag type. The network learned the word-tag mapping from a large corpus set using back-propagation algorithm. The output of the network is the corresponding tag for a seen word and (hopefully) for an unseen word. Note that the method in [5] relies heavily on the completion of the training corpus. For example, a Thai word "ฉัน" is usually used as a first person pronoun; however, sometimes it can be used to mean "eat" in a monk's context. If the corpus contains no such context, the training network will always tag the word to be a first person pronoun. Another drawback of [5] is that one must tag and calculate in advance the occurrences of word-tag mapping in the training corpus. Again, the reliability of the corpus is imposed on the accuracy of the network. Feature-based approach is used in [1] and [2] for the purpose of segmenting Thai word. It is interesting to note that word's neighborhood concept is used in both [1], [2] and [5], although the calling names are different. In [1] and [2], neighborhood concept is employed in two ways: context words and collocations. Context words are a specific word present nearby a target word. Collocations are patterns of contiguous words around the target word. Both are used as clues to segment the target word. The neighborhood word concept is also employed in [5]. The sentence-level tagging in [5] allows input vector to include l words on the left and r words on the right of the target word. The work in [6] examined POS tagging in Thai language using support vector machine. The problem was viewed as a decision to choose between two categories (of word types), which have equally positive and negative examples in the training set. It is interesting to note that the author of [6] mentioned that this method requires huge machine resources; therefore, it is not suitable for large-scale corpus.

The idea of this research is directly inspired by the problem found in [7]. The goal of the research in [7] is to segmenting a simple Thai sentence using Modified Augmented Transition Network (M-ATN) for Thai grammar proposed by [8]. The overall results are not very impressive. Detail analysis of the results in [7] showed that the lack of role assignment in the M-ATN grammar used is directly responsible for the erroneous cases, i.e., the cases where sentences are segmented incorrectly. In addition, the work in [7] used the corpus in [3] as one of its sample inputs. During the work, the mapping of POS tag in [3] to the word tag in [8] is

relatively straightforward. The details investigation of [3] also showed that the POS tagging in [3] only contain category tag, but does not contain role assignment tag.

3. CONCEPT AND METHODOLOGY

The current status of research in Thai POS tagging mainly tags for word's category, such noun, pronoun, verb, preposition and so on. The category information alone is hardly sufficient for further processing Thai languages. For example, in order to segmenting a sentence, one must know whether a given noun plays a subject role or an object role, whether a given verb requires an object or not. This work suggested that POS tag information should also include the role of words in sentences, in addition to word's category information.

It should be noted that, unlike English, this role information is not easily extracted from Thai words. Some of the reasons are the following. First, Thai word in a sentence does not change form, i.e., it does not encoded any grammatical information at all. Second, Thai sentences in normal usage may completely omit a subject or object and still tolerate some level of ambiguity. Consequently, from grammatical viewpoint, sometimes a transitive verb may look like intransitive verb.

The goal of Role-Pos tagging is to assign a role to a target phrase or word in a given sentence. This work does not consider the segmentation of a sentence into phrases or words. We assume that the phrases or words have been segmented before our Role-POS tagging begins. From now on, the segmented phrase or word is considered as a word unit. The input to the multilayer perceptron network is a category tag of a target word. The output is a role tag of the given word. Let n be the number of words in a sentence and $S = w_1w_2...w_n$ be a sentence where w_i is a target word under tagging. Let the string of POS of a sentence S be denoted by $T=t_1t_2...t_n$ and the string of the role of a word in sentence be denoted by $F=f_1f_2...f_n$ *where* f_i is the role of word w_i. In this paper, a multilayer perceptron network is trained using a back-propagation learning algorithm in order to assign a role POS to the word w_i in a sentence S. In addition to the target words, the network allows k words to the left and the right of the target phrase or word to form an input vector. The input vector x_i is constructed from the sequence of the category POS which is centered on target word w_i. The input vector x_i is denoted as follow:

$$x_i=(t_{i-k}, ..., t_{i-1}, t_i, t_{i+1}, ..., t_{i+k})$$

where k is the number of neighborhoods of t_i.

The experiment uses k = 1. Moreover, we assume that the sentence boundary is known and use a special tag to represent the boundary. The network is trained for 24 sentences excerpt from news in a daily newspaper.

4. EXPERIMENTAL RESULTS AND DISCUSSION

In this experiment, we excerpt 24 sentences from news in a local newspaper. These sentences are divided into two sets: a training set and a test set. The same multilayer perceptron network with one hidden layer and the back-propagation learning algorithm are used in both sets. The method used is detailed in [9]. The network consists of 15 nodes in its input layer, 32 nodes in the hidden layer, and 5 nodes in the output layer. In the training set, we use 2 neighbors of target words to construct the input vector. The network was trained on 103 words from 19 sentences and was tested on 25 words from 5 sentences. Word's category is shown in Table1.

Category POS Tag
Noun (N)
Pronoun (Pron)
Adjective (Adj)
Verb (V)
Adverb (Adv)
Preposition (Prep)
Conjunction (Conj)

Table 1: Details of category POS Tag used in the experiment

The experiment classifies input categories into 7 categories; those are: noun, pronoun, adjective, verb, adverb, preposition and conjunction. This classification is manually done for both the training set and the test set for this experiment.

Role-POS Tag
Subject (S)
Indirect Object (IO)
Direct Object (DO)
Prepositional Object (PO)
Noun Modifier (MN)
Transitive Verb (VT)
Intransitive Verb (VI)
Verb modifier (MV)
Place (PP)
Time (PT)
Possessive (PS)
Coordinate Conjunction (C1)
Subordinate Conjunction (C2)

Table 2: Details of Role POS Tag used in the experiment

The Role POS tags are shown in Table2. The experiment classifies output tags according to their roles in a given sentence. A noun can be classified as subject, direct object, indirect object, prepositional object, and noun modifier. A pronoun can be classified similarly. A verb can be a transitive verb or an intransitive verb. An adverb and an adjective play the role of verb modifier and noun modifier respectively. A preposition can show place, time or possessive. Finally, a conjunction can connect two sentences or connect two clauses.

Category Tag		Role Tag
Category	Training Set	(#count, %)
N	50	S=(23, 46%), DO=(14, 28%), IO=(0, 0%), PO=(9, 18%), MN=(4, 8%)
Pron	0	-
Adj	6	MN=(6, 100%)
V	25	VI=(11, 44%), VT=(14, 56%)
Adv	8	MV=(8, 100%)
Prep	9	PP=(2, 22.5%), PT=(2, 22.5%), PS=(5, 55%)
Conj	5	C1=(0, 0%), C2=(5, 100%)

Table 3: Frequency of category tags and role tags in the training set

Table 3 shows the frequency of category tags and the corresponding role tags in the training set. Note that there is a number of missing tags in the set. For example, there is no pronoun tag as input tag in the training set. For output tags, the training set contains neither an indirect object nor a coordinate conjunction.

		Results (# words)		
Input Tag	Output Tag	Expected	Actual	% correct
Noun (N)	Subject (S)	5	5	100
	Indirect Object (IO)	-	-	-
	Direct Object (DO)	4	4	100
	Prepositional Object (PO)	1	1	100
	MN	1	1	100
Pronoun (Pron)	Subject (S)	-	-	-
	Indirect Object (IO)	-	-	-
	Direct Object (DO)	-	-	-
	Prepositional Object (PO)	1	1	100
Adjective (Adj)	Noun Modifier (MN)	2	2	100
Verb (V)	Transitive Verb (VT)	3	4	100
	Intransitive Verb (VI)	4	3	75
Adverb (Adv)	Verb modifier (MV)	1	1	100
Preposition (Prep)	Place (PP)	1	0	0
	Time (PT)	-	-	-
	Possessive (PS)	1	2	100
Conjunction (Conj)	Coordinate Conjunction (C1)	-	-	-
	Subordinate Conjunction (C2)	1	1	100

Table 4: Result of Role-POS Tagging for 25 words from 5 sentences

The result of Role-Pos tagging in the testing set of 25 words from 5 sentences is shown in Table 4. The expected tag is the tag judging by human, which the network should output. The actual tag is the tag the network actually output. The percent correct is the percentage of the actual output divided by the expected output, but not over a hundred percent. The network classified each input category into a number of Role-POS tags. For instance, the noun tags are correctly classified into subject tag, direct object tag, indirect object tag, prepositional object tag and noun modifier tag. Two incorrect tagging are found in the testing set. Those are intransitive verb, which is incorrectly tagged as transitive verb, and place preposition, which is incorrectly tagged as possessive preposition. The average correctness of Role-POS Tagging for this experiment is 92 percent. Note that the testing set and the training set are different sentences but are excerpted from the same news. This could contribute to the high correctness percentage of Role-POS tagging.

Although the result is promising, the analysis of incorrect output tags could caused by the small training set. The training set used in this experiment is very small comparing to other work which usually used the whole corpus for training. Consequently, there are some tags with zero occurrences. In proposition category, 55 percent are tagged as possessive proposition (PS). This could explain the incorrect tagging from PP to PS, similar to the incorrect tagging from intransitive verb (VI) to transitive verb (VT). However, in order to confirm this result, the larger set of a training set is needed for further experiments.

5. CONCLUSIONS AND FUTURE WORK

This work proposes the concept of tagging a role feature to Thai words. Traditionally; Thai words are tagged with word category only. The role tag is useful for further processing Thai language, such as sentence segmentation and syntactic tree construction. Role assignment is realized by artificial neural network (ANN), specifically, multilayer perceptron network with back-propagation learning algorithm. The experiment is done using Thai sentences from local daily newspaper. The input tags are category tags and the output tags are role tags. Although the percentage of correct tagging from the testing set is high (92%), this could come from the fact that the training set and the testing set are excerpted from the same news, although different sentences. The incorrect tags signify that the training set is still too small. More experiment could be done in the future using a larger training set, possibly from available Thai corpus.

6. REFERENCES

[1] Meknavin, S., Charoenpornsawat, P., and Kijsirikul B., **1997.** Feature-based Thai Word Segmentation. In *Proceedings of the Natural Language Processing Pacific Rim Symposium (NLPRS'97), Phuket, Thailand.*

[2] Chareonpornsawat, P., Kijsirikul, B., Meknavin, S., **1998.** Feature-based Thai unknown word boundary identification using Winnow. *IEEE Asia-Pacific Conference, Nov 1998.* 547-550.

[3] Sornlertlamvanich, V., Charoenporn T., and Isahara, H., **1997.** ORCHID: Thai Part-of-Speech Tagged Corpus. In *Technical Report Orchid Corpus.*

[4] Ma, Q., **2002.** Natural Language Processing with Neural Networks. In *Proceedings of the Language Engineering Conference(LEC'02), 13-15 Dec. 2002, 45-56.*

[5] Ahmed., Raju S.B., Chandrasekhar Pammi V.S., and Prasad M.K., **2002.** Application of Multilayer Perceptron Network for Tagging Parts-of-Speech, *Proceedings of the Language Engineering Conference (LEC'02), 13-15 Dec. 2002, 57-63.*

[6] Murata, M., Ma, Q., and Isahara, H., **2001.** Part of Speech Tagging in Thai Language Using Support Vector Machine. *The Second Workshop on Natural Language Processing and Neural Networks (NLPNN' 2001).*

[7] Netisopakul, P., Keawwan, K., **2004.** Simple Thai Sentence Segmentation Using M-ATN. *Graduate Project Report, Faculty of Information Technology, King Mongkut's Institute of Technology Ladkrabang.*

[8] Varakulsiripunth, R., Junwun, S., and Maneenate, N., **1989.** Thai Syntactical Analysis by M-ATN. *Papers on Natural Language Processing: Multi-lingual Machine Translation and Related Topics (1987-1994), 192-202.*

[9] Haykin, S., **1998.** *Neural Networks: A Comprehensive Foundation (2ed), Prentice Hall.*