A Statistical Approach to Thai Morphological Analyzer*

Kawtrakul Asanee , Thumkanon Chalathip

Natural Language Processing and Intelligent Information System Technology Research Laboratory,
Dept. of Computer Engineering,
Kasetsart University, THAILAND
email : ak@nontri.ku.ac.th

Abstract

Three nontrivial problems of Thai morphological processing are word boundary ambiguity, tagging ambiguity and implicit spelling errors. These problems cause a lot of difficulty to the parser due to the alternative or erroneous chain of word. This work attempts to provide a computational solution, called Word Filtering, to those linguistic phenomena. The filtering process calculates the probabilities of all possible chains of tagged words using a Markov Model. The most likely sequence of tagged word is the one that maximizes the chain probabilities. However, it may be an erroneous chain which has an implicit spelling error. Therefore, the Word Filtering, also, includes the scanning process that detect and correct these errors. Both filtering and scanning process use a statistical data information collected from the hand-tagged corpus.

The experiment has shown that word filtering can eliminate most of the alternative word sequences. Moreover, this technique is fairly good at the implicit error correction.

1. Introduction

One of the major problems in many languages, such as Japanese, Chinese, Korean and Thai, is word boundary ambiguity because these languages do not have any delimiters between words. The second problem is tagging ambiguity which occurs when there is more than one tag for one word. Another problem is implicit spelling error that occurs because some incorrect words can be found in a dictionary .This problem is very hard to solve with a simple approach, such as dictionary approach. Thai morphological analysis must face these three problems which cause many possible alternative or/and the erroneous chains of words. These problems generate a lot of unnecessary work for the parser. In order to simplify the parser and speed it up, three important points to bear in mind when considering the morphological processing are neat segmentation of characters into words, part of speech tagging selection, and implicit spelling error detection. This work attempts to provide a computational solution, called Word Filtering, to handle those three points prior to parsing.

The proposed model of Thai morphological analysis consists of three steps: sentence segmenting, spell checking and word filtering. Using word formation rules and a dictionary look up algorithm in the first step, all possible word groups with all possible tags will be given. If there is any explicit error, the second step, that is spell checking, will give a suggestion about a set of most likely words. However, the implicit spelling error may still exist and will affect the parser. That is, the parser must search a large set of tagged word combinations in order to choose the right one. Thus, the main goal of word filtering is to reduce the combination of unuseful tagged words and to identify implicit spelling error.

*A Statistical Approach to Thai Morphological Analyzer is a part of WPA (Writing Production Assistant) Project supported by the National Research Council of Thailand.

The proposed Word Filtering method consists of two steps: a filtering process and a scanning process. The first process will try to filter out any incorrect word boundary and any unsuitable tag. The second process detects and corrects the implicit spelling error by generating the new words for the detected error.

The basic idea of the filtering process is to calculate the probabilities of all possible chains of tagged words by using a trigram of the Markov Model. The most likely sequences of tagged words are the ones that maximize chain probabilities. Nevertheless, they may be an erroneous chain which have implicit spelling errors. Thus, the Word Filtering, also includes the scanning process to detect and correct the error. At this step, a set of words will be generated by a generating function and be replaced to the detected word. The most likely sequences of correct words are the ones that maximize chain probabilities. Both filtering and scanning processes use the statistical information collected from the hand-tagged corpus. From results of the experiments on small corpus (about 10,000 sentences), word filter can eliminate alternative word sequences and can correct the implicit error quite well.

In the following section. key problems in Thai morphological analysis are described. Then, we present the overview of a computational morphological processing in section 3. In the section 4, the concept of how to use the statistical information to handle word boundary ambiguity, tagging ambiguity and implicit spelling error will be explained. Finally, we present the conclusion result of the experiment .

2. Key Problems in Thai Morphological Analysis

There are three nontrivial problems of Thai morphological processing: word boundary ambiguity, tagging ambiguity and implicit spelling errors.

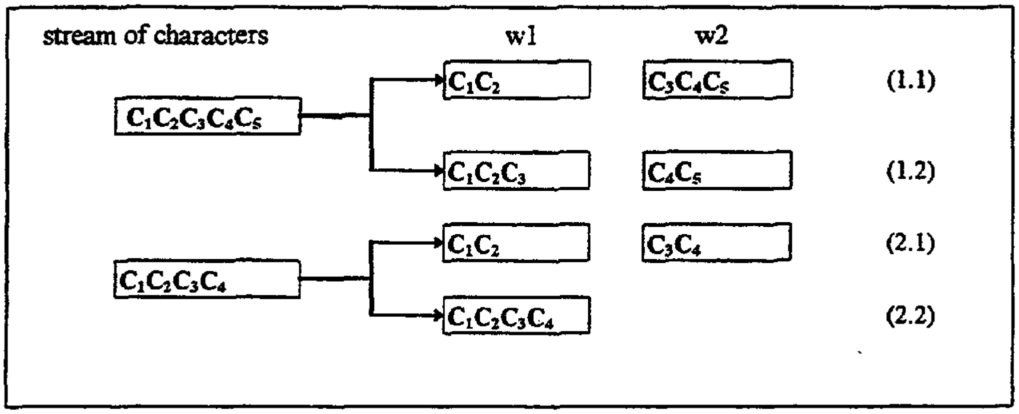
2.1 Word Boundary Ambiguity

Thai sentences are similar to the Japanese's and Chinese's in terms of having no blank space to mark each words within the same sentence. Additionally most of Thai words are multisyllabic words. Some of them contains more than monosyllabic words as parts of its component. This causes word boundary ambiguity.

Let C be a sequence of characters: $C = c_1 c_2 c_3 \ldots c_n$

Let W be a sequence of words: $W = w_1 w_2 \ldots w_n$ where $w_i = c_{i1}..c_{ir}$

Giving a stream of characters, the possible word segmentation is as following :



As shown above, the word in "$C_1C_2C_3C_4C_5$" pattern has two ambiguous forms. One is "C_1C_2" and "$C_3C_4C_5$". The other one is "$C_1C_2C_3$" and "C_4C_5". In our corpus, more than 50% of sentences include word boundary ambiguity.

The assignment of part of speech to the segmented word is also effected by the word boundary ambiguity. This causes the ambiguous pattern in a sentence The example is as shown in the following:

Example 1

เรือ <u>โคลง</u> เพราะ <u>โคลง</u> เรือ

ABCD ABCD

? ?

The ambiguous patterns of the above sentence are :

	เรือ	โค	ลง	เพราะ	โค	ลง	เรือ
a)	N	N	V	conj	N	V	N
	(The boat)	(ox)	(go down)	(because)	(ox)	(go down)	(the boat)
b)	N	N	V	conj	V		N
	(The boat)	(ox)	(go down)	(because)	(shake)		(the boat)
c)	N	V		conj	N	V	N
	(The boat)	(shake)		(because)	(ox)	(go down)	(the boat)
d)	N	V		conj	V		N
	(The boat)	(shake)		(because)	(shake)		(the boat)

In the above example, only c) and d) are the meaningful sentences.

2.2 Tag Ambiguity

A Thai word can have more than one part of speech. This tag ambiguity can cause a large set of tagged word combinations. Consider the following example :

Example 2

อาหารเช้า	ที่	เขา	ทำ	จัด	ไว้	ที่	ที่	ของ	ฉัน	2	ที่
cn: (breakfast)	1) relpron: (which),	pron: (he)	v: (cook)	v: (set)	postv	1) relpron: (which),	1) relpron: (which),	1) prep: (of)	1) pron: (I)	num (2)	1) relpron: (which),
	2) prep: (at),					2) prep: (at),	2) prep: (at),	2) cn: (thing)	2) v: (eat)		2) prep: (at),
	3) cn: (place),					3) cn: (place),	3) cn: (place),				3) cn: (place),
	4) cl: (dish)					4) cl: (dish)	5) cl: (dish)				4) cl: (dish)
	(4 tags)					(4 tags)	(4 tags)	(2 tags)	(2 tags)		(4 tags)

The above multiple-tagged words give 1024 combinations of word chain. However, only one word chain is correct. Figure 1. shows tag ambiguity in our corpus. As we can see, there are about 95% of the words are ambiguous with regards to the tags they take.

Number of tags	Number of words	Percentage
1 tags	130	4.24
2 tags	1750	57.0
3 tags	998	32.5
4 tags	192	6.26

Figure 1. Tag ambiguity found in 3070 words corpus.

Both word boundary and tag ambiguity increase the complexity in syntax analysis. It also increases the amount of time used for parsing the sentences. Besides these two ambiguities, spelling errors in Thai, called implicit spelling errors, also cause a lot more work for the parser.

2.3 Implicit Spelling Error

Implicit spelling errors, one of ill-formedness usually encountered in documents, are caused by either carelessness or lack of knowledge. This type of error can not be detected by simply using a dictionary approach. There are three kinds of typing errors caused by the carelessness: Missing, Keyboard Mistyping, and Swapping as the examples in the following :

Cause Type	carelessness	lack of knowledge
Missing	(t)his → his	free → fee
Mistyping	fa(t) → far	both → boat
Swapping	(n)o → on	form → from

In case of lacking of knowledge, the errors occured from the unclear speech confuse to the typist. Additionally, they also occurs from the confusion in writing since there are many forms in one sound. See the following example.

Example 3

ฉัน เข็น แพ ลง น้ำ จน ขา แพลง

C_1C_2 C_3C_4 $C_1C_2C_3C_4$

(N) (prep) (V)

I push raft down water until leg twist
(I push the raft to the river and twisted my leg.)

1.The ambiguous patterns caused by word boundary ambiguity are :

	ฉัน	เข็น	แพ	ลง	น้ำ	จน	ขา	แพ	ลง
a)	pron	V	N	prep	N	conj.	N	N	prep
b)	pron	V	N	prep	N	conj.	N	V	
a)	pron	V	V		N	conj.	N	N	prep
a)	pron	V	V		N	conj.	N	V	

Example 3 (continue)

2.The erroneous chain of words caused by implicit spelling error :

ฉัน	เข็น	แพ	ลง	น้ำ	จน	ขา	แพง *
							$C_1C_2C_4$ (C_3 is missed)
pron	V	N	prep	N	conj.	N	mod
I	push	raft	down	water	until	leg	expensive

The implicit spelling errors can occur much easier in Thai than in English and Japanese (in Hiragana) because the errors always involve using a word that has a similar pronunciation. There are about 20-30% of Thai words that can cause this kind of the confusion to typist. Additionally, there are 2 characters in one key pad (see figure 2). Thus, keyboard mistyping increases the way of implicit misspelling which can not be detected easily using the dictionary - based approach.

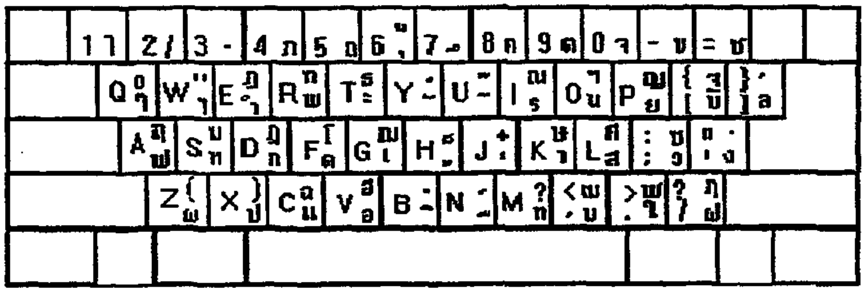


Figure 2. Two level key pads for Thai character.

In this work, we attempt to provide a computational solution to handle these three nontrivial problems for making the job of a parser much easier. The next section will present the overview of the system.

3. An Overview of a Computational Morphological Processing for Thai

A computational model consists of word segmenting, spelling checking and word filtering processes is proposed to handle the morphological problems mentioned earlier. (see figure 3)

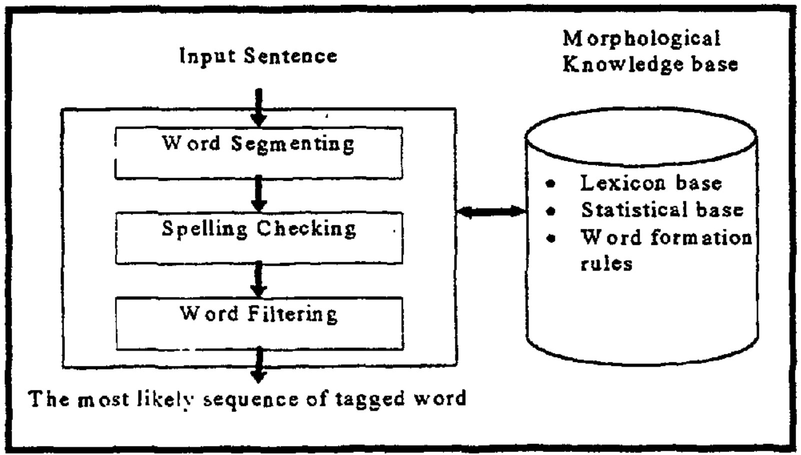


Figure 3. An overview of Thai Morphological Analysis.

Input sentence is a stream of characters without explicit delimiters. Using word formation rules and Lexicon base look up algorithm [KAW95(a)], the word segmenting process, will provide all possible word grouping with all possible tags. If there is any explicit error then the second step, the spell checking process, will be called to give a suggestion with a set of most likely word [KAW95(a)]. However, an implicit spelling error may still exist. In order to choose the right tagged word combination, word filtering process will use the statistical association among words, collected as a statistical base, to eliminate the alternative and/or erroneous chain of words which is caused by word boundary and tagging ambiguities and implicit spelling error. This paper concentrates only on the word filtering process. The detail of the process will be discussed next.

4. Word Filtering

All of word boundary, part of speech tag and implicit spelling error can be disambiguated by using a trigram model [CHAR 93] to calculate the probabilities of word cluster. The sentences shown in example 1, 1c) and 1d) are meaningful sentences. In other words, they have the strength of association of word in a chain more than 1a) and 1d) have. The association between words in ``the boat shake" is stronger than in ``the boat ox". In example 2, we can also can find the most likely sequence of parts of speech by considering the previous part of speech. Since an implicit spelling error affects both meaning and tag, (such as : บิน (fly) : v → บน (on): preposition) the special process is needed . Consequently, word filtering will consists of two processes : a filtering process used to eliminate unuseful tagged word combinations and a scanning process used to detect and correct an implicit spelling error by generating a new set of words according to the cause of errors and selecting the one that maximizes the probabilities of word cluster. Both processes need to look up a statistical information collected from the hand-tagged corpus.

4.1 The Training Corpus

The training corpus is a set of sentences, divided into two groups. Each sentence in the first group is prepared to give a context for a word which has a possibility to become an implicit spelling error, and a context for a sequence of words that have word boundary ambiguity. The second group are sentences prepared to give a context for a multiple-tagged word. All of these sentences have already segmented and tagged. A statistical information will be collected as a statistical base to support both filtering process and scanning process. Thus, collected statistics not only emphasize on the frequency of using individual words but also on the cluster of words.

4.2 Markov Model as a Statistical Model of Filtering Process

4.2.1 Trigram Model

A Trigram Model [CHAR 93] is utilized to calculate the probabilities of word cluster, i.e. how the previous two words affects the probabilities of next word. This can be explain in equation (1).

$$P(w_{1,n}) = \prod_{i=1}^{n} P(w_i | w_{i-2,i-1}) \qquad (1)$$

In order to estimate the probability of $P(w_i|w_{i-2,i-1})$ in (1), the following equation is used :

$$P_e(w_i|w_{i-2,i-1}) = \frac{C(w_{i-2,i})}{C(w_{i-2,i-1})} \qquad (2)$$

where $P_e(X)$ is the estimated probability for X based on some count C.

So to estimate the probability of w_i appear after "w_{i-2},w_{i-1}", we count how many times the pair "w_{i-2},w_{i-1}" appears in our corpus and how many times "w_{i-2},w_{i-1},w_i" appears and divide.

Because of the sparse-data problem in trigram model, rather than equation (1), we instead use :

$$\prod_{i=1}^{n} P(w_i|w_{i-2,i-1}) = \prod_{i=1}^{n}\left(\lambda_1 P_e(w_n) + \lambda_2 P_e(w_n|w_{n-1}) + \lambda_3 P_e(w_n|w_{n-2,n-1})\right) \qquad (3)$$

Thus, we can compute the better probabilities although the relevant trigram or bigram data are missing. The result from the experiment shows that the assigned values .1, .3, .6 to λ_1, λ_2, λ_3, [CHAR 93] respectively, will give the satisfied solution for Thai word sequence probability. Using equation (3), the strength of association of words in a chain can be calculated.

In order to handle the tagging ambiguity problem. A trigram part of speech model is also used [DeRose 88]

$$P(t_{1,n}) = P(t_i|w_i)P(t_i|t_{i-2}t_{i-1}) \qquad (4)$$

Since the proposed model is provided for disambiguating both word boundary and tag, we use the average of probabilities calculated by the equation (3) and (4) as the strength of a chain of tagged words and select the higher one as the most likely sequence of corrected word with their tags. For example, the strength of word chain (see the example 1) in 1c) higher than 1a) while the probabilities of the sequence of parts of speech of 1a) and 1c) are equal. Based on the average of the strength of word chain and the most likely sequence of parts of speech, 1c) will be selected as the solution of word segmentation and tagging.

4.2.2 Two parts of Word Filtering

There are two parts in word filtering (see the figure below)

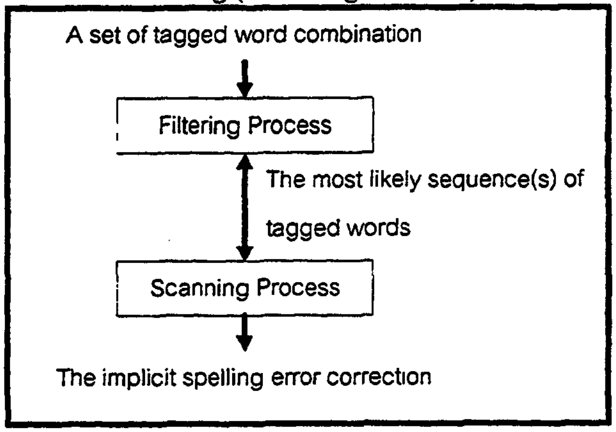

A set of tagged word combination
Filtering Process
The most likely sequence(s) of tagged words
Scanning Process
The implicit spelling error correction

Figure 4. Two Parts of Word Filtering

The first part of word filtering, i.e., the filtering process, calculates the strength of each tagged word combination. The combination(s) that gives the highest value will be the most likely sequence(s) of tagged words. In the second part, scanning process, an implicit spelling error will be detected and corrected [KAW95(b)]. That is, the weakest strength of word cluster will be assumed to have an implicit spelling error. Then a new set of words which are generated according to the causes of error will be replaced to that detected word one by one. A replaced word which gives the highest value of the strength of word chain will be a solution of an implicit spelling error.

5. Conclusion

From the results of the experiment shown below, word filter can eliminate many of alternative word sequences and correct the implicit error. This result makes the job of the parser much easier and speeds it up.

Pruning Approach	Word boundary ambiguity	tagging ambiguity	implicit spelling error	speed (for one sentence)
Corpus based (word filtering)	85.2%	76.6%	61.9%	4msecs-minutes

Acknowledgment

The work reported in this paper was supported by the National Research Council of Thailand. Thanks are also due to Patcharee Varasai, Supapas Kumtanode, Thitiporn Tharapoome and Mukda Suktarajarn for their helpful on the preparation of the training corpus, Puchong Uthayopat and Amarin Deemagarn for their helpful to complete this paper..

References

[ARAKI94] Araki Tetsuo. Ikehara Satoru, Tsukahara Nobuyuki, Komatsu Yasunori, ``An Evaluation to Detect and Correct Erroneous Characters Wrongly Substitued, Deleted and Inserted in Japanese and Englist Sentences Using Markov Models.", COLING94, 1994, pp. 187-193.

[CHAR93] Charniak Eugene, ``Statistical Language Learning", MIT Press, 1993.

[DeRose88] DeRose, S.J., ``Grammatical Category Disambiguation by statistical optimization", Computational Linguistics 14, 1988, pp. 31-39.

[KAW95(a)] Kawtrakul Asanee, Muangyunnan Parinee, Maneekanjanajing Nopparat, ``A Morphological Analyzer for Writing Production Assistant System", A Progress Report to the National Research Council of Thailand, 1995.

[KAW95(b)] Kawtrakul Asanee, ``A statistical Approach to Ambiguity Filtering in WPA System", A Progress Report to the National Research Council of Thailand, 1995.

[SHIHO94] Shiho Nobesawa, ``Segmenting a Sentence into Morphemes Using Statistic Information between Words", COLING94, 1994, pp.227-233.