A Gradual Refinement Model for A Robust Thai Morphological Analyzer

KAWTRAKUL Asanee, THUMKANON Chalatip,
JAMJANYA Thitima, MUANGYUNNAN Parinee, POOLWAN Kritsada
Natural Language Processing Research Laboratory, Department of Computer Engineering Kasetsart University,
Paholyothin Rd.,Bangkok 10900, Thailand.
email: ak@nontri.ku.ac.th

INAGAKI Yasuyoshi
Department of Information Engineering Nagoya University, Chikusa-ku, Nagoya 464, Japan.
email : inagaki@inagaki.nuie.nagoya-u.ac.jp

Abstract

This work attempts to provide a robust Thai morphological analyzer which can automatically assign the correct part-of-speech tag to the correct word with time and space efficiency. Instead of using a corpus based approach which requires a large amount of training data and validation data, a new simple hybrid technique which incorporates heuristic, syntactic and semantic knowledge is proposed. To implement this technique, a three-stage approach is adopted to the gradual refinement module. It consists of preference based pruning, syntactic based pruning and semantic based pruning. Each stage will gradually weeds out word boundary ambiguities, tag ambiguities and implicit spelling errors. From the result of the experiment, the proposed model can work with time-efficiency and increase the accuracy of word boundary segmentations, POS tagging as well as implicit spelling error correction.

1. Introduction

One of the important requirements for developing practical natural language processing system is a morphological analyzer that can automatically assign the correct POS (part-of-speech) tagging to the correct word with time and space efficiency. For non-separated languages such as Japanese, Korea, Chinese and Thai, the more task in morphological analyzer is needed, i.e, segmenting an input sentence into the right words (Nobesawa et.al, 1994; Seung-Shik Kang et.al, 1994). However, there is another problematic aspect, called implicit spelling error, that should be solved in morphological processing level. The implicit spelling errors are spelling errors which make the other right meaningful words. This work attempts to provide a robust morphological analyzer by using a gradual refinement module for weeding out the many possible alternatives and/or the erroneous chains of words caused by those three non-trivial problems: word boundary ambiguity, POS tagging ambiguity and implicit spelling error.

Many researchers have used a corpus based approach to POS tagging such as trigram model (Charniak, 1993); feature structure tagger (Kemp,1994), to word segmentation, such as D-bigram (Nobesawa et.al, 1994), to both POS tagging and word segmentation (Nagata, 1994) and to spelling error detection as well as correction (Araki et.al, 1994; Kawtrakul, et.al, 1995(b)). Eventhough a corpus based approach exhibits seemingly high average accuracy, it requires a large amount of training data and validation data (Franz, 1995). Instead of using a corpus based approach, a new simple hybrid technique which incorporates heuristic, syntactic and semantic knowledge is proposed to Thai morphological analyzer. It consists of word-boundary preference, syntactic coarse rules and semantic strength measurement . To implement this technique, a three-stage approach is adopted to the gradual refinement module : preference based pruning, syntactic based pruning and semantic based pruning. Each stage will gradually weed out word boundary ambiguities, tag ambiguities and implicit spelling errors.

Our preliminary experiment shows that the proposed model can work with a time-efficiency and increase the accuracy of word boundary and tagging disambiguation as well as the implicit spelling error correction.

In the following sections, we will begin by reviewing three non-trivial problems of Thai morphological analyzer. An overview of the gradual refinement module will be given. We will then show the algorithm with examples for pruning the erroneous word chains prior to parsing. Finally, the results of applying this algorithm will be presented.

2. Three Nontrivial Problems of Thai Morphological Processing.

2.1 Word Boundary Ambiguity

Like many other languages such as Japanese, Chinese and Korean, Thai sentences are formed with a sequence of words mostly without explicit delimiters. Especially, for Thai and Japanese written in Hirakana (Nobesawa,1994), a word is a stream of characters. This causes the problem of word boundary ambiguity (see Fig.1).

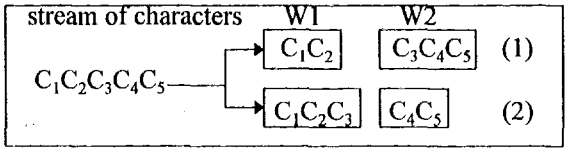


Figure1. Two possible grouping characters into words: longest possible segment or shortest possible segment

There are two possible grouping characters into words, i.e, shortest possible segment such as (1) and longest possible segment such as (2) in Fig.1. Each word given by either way of grouping has a meaning. In our corpus, more than 50% of sentences include word boundary ambiguity. This causes a lot of alternative chains of words where some are meaningless.

2.2 Tagging Ambiguity

Thai word can have more than one part of speech. In our corpus, only 2% of sentences are written by using one-tagged words. Accordingly, tag ambiguity in Thai causes a large set of tagged word combinations. We found that a sentence with 12 words can generate 3027 syntactic patterns of word chain. Both word boundary and tag ambiguity also create complexity in syntactic analysis.

2.3 Implicit Spelling Error

Spelling errors in Thai are classified into two types (Kawtrakul, 1995 (b)): explicit spelling error and implicit spelling error. The former can be detected easily by using a dictionary-based approach. The latter can not be detected by simply using dictionary since the error can lead to words that are unintended, but spelled correctly. Table 1 shows three kinds of spelling errors caused by carelessness and lack of knowledge.

Table 1. Three types of implicit spelling error.

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

In Thai, implicit spelling errors can occur more easily than in English because there are 2 distinctive characters on each-keypad. From the result of our experiment, 2,286 words can generate 6,609 implicit spelling error words where 75.68 % of those errors have new syntactic categories. This will cause an erroneous pattern of word chain which increases a lot of unnecessary job to the parser.

Accordingly, Thai morphological analysis is not only expected to assign the right tag to the right word but should correct the implicit spelling error prior to parsing.

3. An Overview of Thai Morphological Analyzer with a Gradual Refinement Module

Instead of using a corpus based approach which requires a large amount of training data and validation data, a new simple hybrid technique which incorporates heuristic, syntactic and semantic knowledge is proposed to a gradual refinement module which gradually weeds out the alternative and/or the erroneous chains of words caused by those three nontrivial problems. The techniqueis implemented by using word boundary preference, syntactic coarse rules and semantic dependency strength measurement. Fig.2 shows an overview of the system.

The system consists of four steps:

Step 0: This step provides all possible word groupings with all possible tags by using word formation rules and Lexicon base (Kawtrakul et.al, 1995 (a)). If there is any explicit spelling error, it will be detected and suggested for correction. At this stage a temporary dictionary is created for the remaining steps.

Step 1-3 : These steps are preference based pruning, syntactic based pruning and semantic based pruning. Each step will gradually weeds out word boundary ambiguities, tag ambiguities and implicit spelling errors.

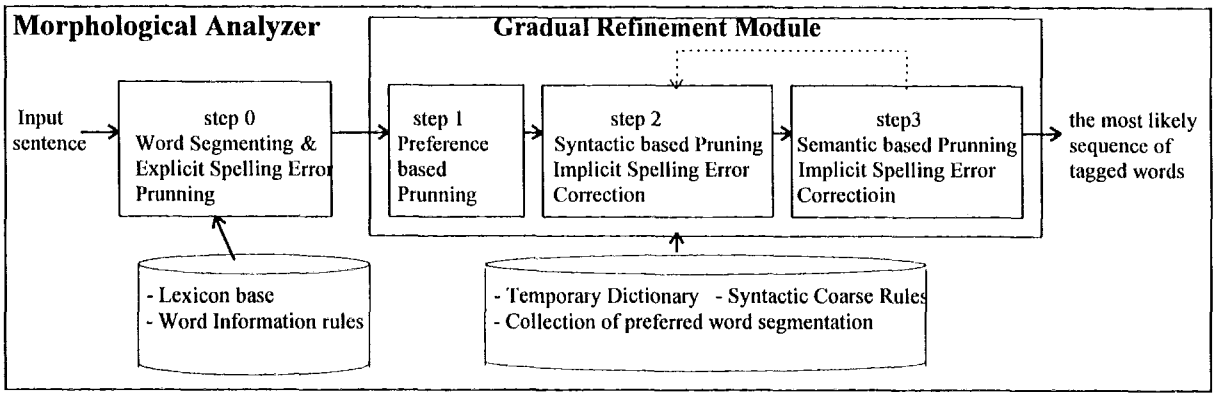


Figure 2. An overview of Thai morphological Analysis with a gradual refinement module

4. A Gradual Refinement Module

4.1 Preference based Pruning

From the Fig.1, Thai Word Segmentation can be implemented as follows:

case i - only longest segmentation or shortest segmentation is possible,

case ii - both longest segmentation and shortest segmentation are possible.

The former will be processed at this stage by looking up the preferred words (see Table 2). Some of them are determined by the cooccurrence word in the left or right. For the latter, it will be processed by the next steps.

Table 2. The Collection of Preferred Words.

stream of char.	segmentation			L_1	R
	longest	shortest	preferred		
มากว่า	มาก-ว่า (much-that)	มา-กว่า (more-than	มา-กว่า (more-than)	*	*
ไฟฟ้า	ไฟฟ้า (electricity)	ไฟ-ฟ้า (fire-sky)	ไฟฟ้า (electricity)	*	*
แพลง	แพลง (twist)	แพ-ลง (raft-down)	แพ-ลง (raft-down)	{เข็น} (push)	*
แพลง	แพลง (twist)	แพ-ลง (raft-down)	แพลง (twist)	{ขา,มือ} (leg,hand)	*

Note : * means any word, L1 means a word in the left., R1 means a word in the right.

In summary, word boundary preference is used to prune the word chains which consist of impossibly occurred or rarely occured word segmentation.

4.2 Syntactic based Pruning and Implicit Spelling Correction

At this stage, the syntactic coarse rules are used for pruning the remaining erroneous word chains caused by the word boundary ambiguities, tagging ambiguities and/or implicit spelling errors.

Syntactic Coarse Rules :An example of the syntactic coarse rules for a set of two consecutive words (W_i, W_{i+1}) in Thai grammar is given as follows :

if W_i is noun then W_{i+1} might be : noun, verb, modifier .....
if W_i is verb then W_{i+1} might be : noun, postverb, mod .....

The POS matrix (PM) given below is used to implement the finite state automaton model of the syntactic coarse rules: where syntactic category of W_i is cat_i and W_{i+1}, is cat_{i+1}.

Table 3. The 46 X 46 POS matrix obtained from 20,000 sentences corpus.

	cat_{i+1}						
cat_i	stop	noun	verb	mod.	postv.	cl.	.....
start	-	1	1	0	0	0	.....
noun	0	1	1	1	0	1	
verb	1	1	1	1	1	0	
mod.	0	0	0	1	0	0	
postv.	0						
cl.	0	0					
.	.	.	.	.	.	.	.

Note : start means the beginning of a sentence, stop means the end of a sentence.

Together with the POS matrix, some constraints, called flag, are used to change the PM_{ij} from 0 to 1. For example :

if there exist ``verb'' before ``modifier'' then flag = 1
else flag = 0

According to the above constraint, PM_{ij}, where i = modifier and j = postverb, can be changed from 0 to 1 if flag equals 1. Based on POS matrix and constraints, now, we can use the following definition to detect the position of error in the word chains.

$$cat_i, cat_{i+1} = \begin{cases} \text{True} & \text{if } PMi_j = 1 \\ \text{True} & \text{if } (PM_{ij} = 0) \wedge (\text{flag} = 1) \\ \text{False} & \text{if } (PM_{ij} = 0) \wedge (\text{flag} = 0) \end{cases}$$

Consider the following example :

W1	W2	W3	W4
กระบอก	อยู่	บน	โต๊ะ
{tube-shape container : n, cl}	{is : v}	{on : prep}	{table : n}

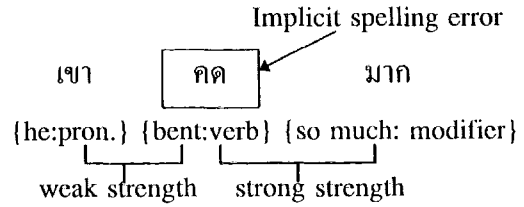
As shown above, W1 has 2 tags : noun and classifier. However, "`classifier" will be pruned since it violates the syntactic coarse rule that "classifier" could not be an initial word. The POS matrix is used to disambiguate word boundary as well.

Finally, if there is no word chain which has all right POS sequences, the erroneous word chain, which has the error marker at the most remote position, will be selected and be expected that there is an implicit spelling error. Then the word generating function will be called for generating a set of candidate words to that position and the process will start pruning at this stage again.

4.3 Semantic based Pruning and Implicit Spelling Correction

Since the syntactic coarse rules only weed out the erroneous POS word chains, some errors still

remain. At this stage, the semantic information from Lexicon Matrix (Kawtrakul et.al, 1995 (a)) is accessed and used to calculate the semantic dependency strength between certain pairs of two words. Consider the following example :



As shown in the above example, there is no POS chain error, but there exists an implicit spelling error which can be detected by using the semantic dependency strength. The word generating function will be called for generating a set of candidate words for the two consecutive words that have weak strength, and the process will return to step 2 for pruning the erroneous POS chains and then goto step 3 for calculating the semantic strength again. The strongest strength chain will be selected as the most likely sequence of the right words in the right tags. The final solution for the above example is

เขา คิด มาก

{he:pron.}{think:verb}{so much: modifier}

5. Experimentation Results

We tested the system on PC-486 DX2 by using two hundred sentences corpus. The percentages of word correctly segmented, tagged and spelled, based on the gradual refinement module and time efficiency are compared with the results based on a statistical approach to word filtering on small training corpus (Kawtrakul, 1995 (b)) as shown in Table 4.

Table 4. Percentage of Accuracy

Approach	Word Segmentation	POS tagging	implicit spelling correction	speed (for one sentence)
Corpus based (word filtering)	85.2%	76.6%	61.9%	msec. - min.
Linguistic based (the gradual refinement model)	92.5%	88.7%	76.6%	msec.

6. Conclusion and Future Work

This paper has described a new simple technique that performs the disambiguation of word boundary, POS tagging and implicit spelling correction by using local information such as lexicon preference, a consecutive POS preference and semantic dependency strength measurement of the associative words in a sentence. From the experimentation results, while a corpus based approach has proven to be efficient, the method seems to be computationally costly and requires a large amount of training data and validation data. For the proposed model, it can work in time efficient and increase the accuracy of word boundary and tagging disambiguation as well as implicit spelling error.

The further directions for this research will concern with unknown word processing and increase the accuracy of the gradual refinement method.

Acknowledgements

The work reported in this paper was supported by the National Research Council of Thailand. Thanks are also due to Patcharee Varasai, Supapas Kumtanode, Mukda Suktarajarn for their linguistic helps..

References

Araki, T., Ikehara, S., Tsukahara, N. and Komatsu, Y., "An Evaluation to Detect and Correct Erroneous Characters Wrongly Substituted, Deleted and Inserted in Japanese and English Sentences Using Markov Model", Coling 94, pp. 187-193, 1994.

Charniak, E., "StatisticalLanguage Learning", MIT Press, 1993.

Franz Alexander, "An Exploration of Stochastic Part of Speech Tagging", Proceeding NLPRLS '95, pp. 217-222, 1995.

Kang, S-S. and Kim, Y.T., "Syllable-Based Model For the Korean Morphology", Coling 94, pp. 221-226, 1994.

Kawtrakul, A., Kumtanode, S., "A Lexicon Model for Writing Production Assistant System", The Proceeding of the second Symposium on Natural Language Processing, pp. 226-236, 1995 (a).

Kawtrakul, A., "A computational Model for Writing Production Assistant System", The Proceeding NLPRL '95, pp. 119-124, 1995 (b).

Kemp, A., "Probabilistic Tagging with Feature Structures", Coling 94, pp. 161-165, 1994.

Nagata, M., "A Stochastic Japanese Morphological Analyzer Using a Forward-DP Backward-A* N-Best Search Algorithm", Coling 94, pp. 201-207, 1994.

Nobesawa, S., Tsutsumi, J., Nitta, T., Ono, K., Jiang, S. and Nakamishi, M., "Segmenting a Sentence into Morphemes using Statistic Information Between Words", Coling 94, pp. 227-232, 1994.