Co-Occurrence-Based Error Correction Approach to Word Segmentation

Ekawat Chaowicharat ekawatchaow@gmail.com
Kanlaya Naruedomkul scknr@mahidol.ac.th
Department of Mathematics, Faculty of Science
Mahidol University, Bangkok, Thailand

Abstract

To overcome the problems in Thai word segmentation, a number of word segmentation has been proposed during the long period of time until today. We propose a novel Thai word segmentation approach so called Co-occurrence-Based Error Correction (CBEC). CBEC generates all possible segmentation candidates using the classical maximal matching algorithm and then selects the most accurate segmentation based on co-occurrence and an error correction algorithm. CBEC was trained and evaluated on BEST 2009 corpus.

1. Introduction

Word segmentation is a fundamental task of natural language processing (NLP) in certain Asian languages including Thai, Chinese, Japanese and Vietnamese (Dang, Tran and Pham 2009; Wang and Huang 2005; Wang, Araki and Tochinai 2008). The lack of reliable word segmentation delays the progress of works in NLP. It was for these reasons that the researches on word segmentation have been carried on still, not excluding The National Electronics and Computer Technology (NECTEC) who has arranged the Thai word segmentation contest Benchmark for Enhancing the Standard of Thai language processing (BEST) during 2009 -2010. (Kosawat et al 2009).

Over decades, a number of word segmentation methods have been proposed. These methods can be classified into two main categories, the dictionary based (DCB) and machine learning based (MLB) (Haruechaiyasak, Kongyoung and Dailey 2008). DCB is more reliable than MLB when dealing with common text. The classic examples of DCB are the longest matching and maximal matching which produce high precision result from testing corpus without named entities (Haruechaiyasak, Kongyoung and Dailey 2008). MLB gains more interest since it is faster and able to recognize a new word it has never seen before. The following are some examples of MLB. In 2005, Wang and Huang used A-proiri based and adjacent characters to help extracting unknown word and then applied the classic maximal matching. In 2006, Kruengkrai and Isahara proposed a novel method using the conditional random field (CRF). Kang and Hwang employed the n-gram model to build a language independent word segmentation process. In 2008, Charoenpornsawat and Schultz applied the left-to-right entropy information in generating the segmentation candidates and selecting the best segmentation that provides the highest average entropy per word.

Although the researches on word segmentation have been carried on since the early period of NLP development, the mission is still incomplete. The researchers have challenged the accuracy of their approaches. Some methodologies were proposed to improve an accuracy of a word segmentation by dealing with an unseen word (i.e., proper noun, named entities) recognition (Huang and Sun 2007; Sutheebanjard and Premchaiswadi 2009; Tirasaroj and Aroonmanakun 2009; Wang 2003).

In this paper, we propose a co-occurrence based approach incorporated with error correction to increase the accuracy of the segmentation. The proposed approach was designed to resolve the segmentation ambiguity problem occurred in a Thai string such as "ตากลม". "ตากลม" can be segmented into two different forms, "|ตา|กลม|" and "|ตาก|ลม|", which conveys different meanings, "round eye" and "expose to the wind", respectively.

2. Relevant Issues

We focused on developing a segmentation approach which is able to resolve the segmentation ambiguity problem and to provide a reliable result for any NLP applications, some issues concerned in designing such an approach are addressed in this section.

2.1 What is "word"?

Before the word segmentation method can be designed, we need to clearly define what the "word" is. According to Longman dictionary, the word is the smallest unit of language that people can understand if it is said or written on its own (Longman 2011). Similarly, the working group who developed the training corpus for BEST 2009 suggested that the small language unit has more benefits

for many kinds of research in NLP. To them, then word is the possible-smallest-meaningful unit (Boriboon et al 2009).

The definition of "word" seems to vary from application to application. In Text To Speech (TTS), word should be defined as a syllable unit for synthesizing the correct utterance whereas word should be defined as a meaningful unit for generating an accurate translation in machine translation. For example, it is possible to segment the string "น้ำปลา" into two words "น้ำ" and "ปลา", for TTS, which refers to "water" and "fish", respectively. However, such segmentation will result in the poor translation since a single word "น้ำปลา" means "fish sauce".

In this research, therefore, a definition of "word" will be defined based on a training corpus provided at that moment to best suited to a specific application. The training corpus used in our experiment, provided by BEST 2009, is a collection of segmented Thai texts including online articles, encyclopedia, news and novels.

2.2 Maximal Matching

Maximal Matching (MM) is a dictionary based word segmentation algorithm. It generates all possible segmentation candidates and then selects the one that contains the fewest number of words. MM is fast, simple and accurate when all the words in the input text found in the available dictionary (Haruechaiyasak, Kongyoung and Dailey 2008; Wang and Huang 2005).

Unfortunately, not unlike the other DCB word segmentation approaches, MM's efficiency is limited to the size of the dictionary. If the input string contains words not listed in the dictionary, the correct segmentation then will not be possible included in the set of candidates. A number of researches were proposed to resolve this limitation (Aroonmanakun 2009; Wang 2003; Huang and Sun, 2007; Sutheebanjard and Premchaiswadi 2009).

The other shortcoming of MM is that the candidate with the fewest numbers of words not always represents an accurate segmentation. For example, the input string "โคลงเรือ" will be segmented into two words "|โคลง|เรือ|" (shaking a ship) by MM but the correct segmentation is three words, "|โค|ลง|เรือ|" (A cow gets on a ship). Our CBEC was, therefore, designed to incorporate a co-occurrence based approach with an error correction technique to ensure the accuracy of segmentation in such a case. Three different error types occurred will be discussed in the next section.

2.3 Word Segmentation Error

Basically, the error occurred in segmentation result can be classified into three different types, the false negative (FN), false positive (FP) and both false positive and negative (FPN).

False negative (FN) is an error considered when a segmentation output exclude some correct delimiter tagging. In this case, the number of tags in the output is less than that of the reference segmentation. Most FN errors occurred if the compound word has higher priority to be included in the segmentation result than its substrings. For example, the single word "แล้วแต่" (up to) has a higher priority than its substrings "แล้ว| แต่" (already| but)

False positive (FP) refers to an error considered when the output gives an undesired tagging. The number of tags in the output is greater than that of the reference segmentation. In this case, the input string contains words (i.e., compound words, proper noun or named entity) which do not exist in the available dictionary.

False positive and negative (FPN) is an error when the number of segmented tags is correct but the tags are not in the right positions. For example, the tag was placed in between "|ตา|อม|" while the correct tag position is "|ตาอ|ม|" according to the reference corpus.

The maximal matching will be used to segment the training corpus, any segmented word that differs from the reference in the training corpus are called error. All error detected will be put into a set so called *error risk bank* and marked by one of the three error types.

2.4 Word Co-occurrence

Word co-occurrence refers to the relationship between a pair of words that appear together in natural language (Atlam et al. 2003; Wartena, Brussee and Slakhorst 2010; Liu et al 2010). In CBEC, the word co-occurrence is used to determine where the segmented tags should be placed. The string "แล้วแต่" in Example 1 and Example 2 are segmented differently according to the co-occurrences words in the context.

Example 1: ฉันกินข้าวเช้า |**แล้ว** | **แต่**|ยังไม่ได้อาบน้ำ
I already had breakfast, but not yet to take a bath.
Example 2: ฉันไปไหนก็ได้ |**แล้วแต่**| เธอ
I can go anywhere, up to you

In CBEC, the word co-occurrence incorporated with error-correction is employed to resolve the segmentation ambiguities.

3. Co-occurrence-Based Error Correction Approach

Co-occurrence-Based Error Correction Approach (CBEC) is our proposed solution to increase accuracy of the word segmentation for any NLP applications. CBEC performs the segmentation in four main phases: candidate generator, best candidate selection, error risk categorization and scoring (Figure 1). In the first phase, all possible segmentation candidates will be generated. In the second phase, the best candidate will be selected using maximal

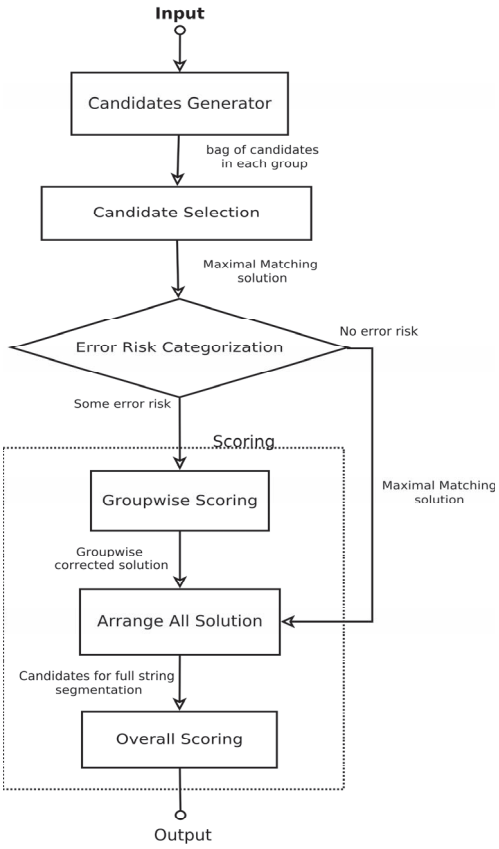

Figure 1: CBEC Architecture

matching algorithm. In the third and fourth phase, an error of the selected candidate will be detected (if any) and will be corrected.

3.1 Candidate Generator

CBEC begins the segmentation process by generating all possible segmentation candidates to ensure that the most accurate result is included. In generating the candidates, first CBEC directly segments the input string based on the words available in the dictionary. Then, it segments the input string based on the *word-like string set*. The word-like string set is a set containing string from named entities and compound words extraction algorithm. All candidates are stored in a *segmentation candidates bag*.

However, to avoid the space and time consuming during the generation process, some not-possible candidates will not be generated. In doing so, CBEC lists all the words in a must-cut group before the candidates can be formed. All possible words from the dictionary of the Example 3 are {ฉัน, ต้อน, โค, ลง, โคลง, เรือ}. Since no word lies across the words "ฉัน" and "ต้อน", this position is therefore, called a must-cut position. "โค" and "ลง" are excluded in the must-cut list because there is a word "โคลง" lies across them.

Example 3: "ฉันต้อนโคลงเรือ"
Must-cut group: {[ฉัน],[ต้อน],[โคลง],[เรือ]}

From the must-cut group, CBEC can generate the candidate as follows:
(1) [ฉัน, ต้อน] forms one candidate - (ฉัน, ต้อน).
(2) [ต้อน, โคลง] forms two candidates- (ต้อน, โค, ลง) and (ต้อน, โคลง).
(3) [โคลง, เรือ] forms two candidates - (โค, ลง, เรือ) and (โคลง, เรือ).

3.2 Candidate Selection

In this phase, the best segmentation in the candidate bag will be selected based on Maximal matching algorithm. The selected candidate will be known as *MM solution.*

The candidate (ต้อน, โคลง) from the last phase is the MM solution since it contains the fewer number of words than the other.

3.3 Error Risk Categorization

The accuracy of the MM solution will be determined in this phase. Each word in MM solution will be checked against the words in error risk bank. If it contains no word in error risk bank, it then will be deemed a segmentation result. But if it contains words in some error category, all the candidates in a bag will be re-examined in the next phase.

3.4 Scoring

If the MM solution contains an error risk, CBEC will then assign the *score vector* to each candidate in the bag and sort them in order to select the better solution than that of MM. In scoring, the four factors are calculated: mean of *inter-relation ratio*, number of non-zero inter-relation pair, *work-like score* per item (if any) and mean of *word occurrence frequency*.

a) Mean of Inter-relation Ratio

Let $W_1, W_2, ..., W_n, W_{n+1}, ...W_N$ be a segmentation candidates. The mean of inter-relation ratio of the candidate is defined by

$$MeanInter = \frac{1}{N-1}\sum_{i=1}^{N-1}\frac{Inter_{n,n+1}}{Inter_{n,n+1} + Intra_{n,n+1}}$$

whereas:
$Inter_{n,n+1}$ = frequency of (W_n, W_{n+1}) appear together as two adjacent words.
$Intra_{n,n+1}$ = frequency of (W_n, W_{n+1}) appear together as one word.

MeanInter varies from 0 to 1. The value 0 means there is no pair that interlexically related and the value 1 means all pairs are fully related and they cannot be concatenated to form a longer word.

b) Nonzero Inter-relation per Number of Pair

We prefer the segmentation with the higher inter-relation score word pairs to that with the lower score word pairs. The higher inter-relation score indicates the stronger the co-occurrence. We want to maximize the number of pair with nonzero inter-relation with respect to the number of pair in a segmentation candidate. So we define

$$NumInterRatio = \frac{\text{Number of word pair with nonzero of inter-ralation score}}{\text{Number of word pair}}$$

c) Work-like Score per Item

A word-like string from the first phase has its own score. This score represents the likelihood that the string can be a word. It is calculated by the sum of Thai character cluster (TCC) co-occurrence score in every pair divided by the number of TCC pairs used in the string.

Word-like score comes into play when the first two score is absent. The candidate containing a word-like string with higher score is more likely to be correct than that with the lower score.

d) Mean of Word Occurrence Frequency.

If the word co-occurrence information is missing, the relation strength between word pair cannot be used to determine the best candidate. Then the word occurrence frequency will be used. This number indicates that how often a word in the candidate occurs in the reference corpus, regardless of the relation between words.

After the score vectors are evaluated, the scoring phase calculates the scalar product between the score vector and the specific weight vector corresponding to the error categories. The candidate with the highest scalar product will be chosen to be the substitute for the MM solution. The score vectors of the group [ต้อน, โคลง] and [โคลง, เรือ] are illustrated in Table 1. Mean of co-oc ratio of (ต้อน, โค, ลง) is higher than that of (ต้อน, โคลง) because the pair (ต้อน, โค) is in the word co-occurrence database, whereas (ต้อน, โคลง) is not.

Table 1 : the score vector of the group [ต้อน, โคลง] and [โคลง, เรือ]

		Mean co-oc	Nonzero co-oc	Word-like	Word oc Freq.	Product score
[ต้อน,โคลง]	(ต้อน, โค, ลง)	1	1	0	1	2.5
	(ต้อน, โคลง)	0	0	0	0.0208	0.0104
[โคลง, เรือ]	(โค, ลง, เรือ)	0.5588	1	0	1	2.0588
	(โคลง, เรือ)	1	1	0	0.1458	2.0729

4. Experimental Results and System Evaluation

We used the BEST 2009 segmented corpus in training and testing our CBEC approach. The Thai dictionary used in CBEC is the Royal Institute 2525. Approximately 4000 random Thai people names were used in training TCC co-occurrence.

The accuracy of segmentation results generated by CBEC were evaluated via the percentage of correctness measurement: the precision (P) and the recall (R).

$$P = \frac{\text{Number of correct words in the solution}}{\text{Number of words in the solution}}$$

$$R = \frac{\text{Number of correct words in the solution}}{\text{Number of words in the reference string}}$$

The harmonic mean of P and R is called the f-measure

$$F = \frac{2PR}{P+R}$$

BEST 2009 training set contains four different text categories: article, encyclopedia, news and novel. We used the first 30 files of each category in training and developing knowledge-bases required in CBEC including word co-occurrence, TCC co-occurrence, error risks bank and weight vectors. The test set contains approximately 600,000 words. The evaluation results, the number of correct word delimeter tagging, exceeding tagging and missing tagging of both MM and CBEC, are shown in Table 2. By comparing F_{CBEC} and F_{MM}, CBEC improved the accuracy of the MM solution by around 8%. The error correction rate in the last row is evaluated by

$$\text{Error correction rate} = \frac{F_{CBEC} - F_{MM}}{1 - F_{MM}}$$

Error correction rate shows that CBEC can correct about 60% error from MM.

Table 2 : comparative evaluation result for each category in the BEST corpus

		Article	Ency	News	Novel	Total
MM	Correct	135559	134437	131270	131731	532997
	Exceeding	14717	14209	14830	18822	62578
	Missing	25612	23644	30463	26783	106502
	F_{MM}	0.870511	0.876591	0.852865	0.852443	0.863101
CBEC	Correct	154083	147856	149623	149880	601442
	Exceeding	7333	7955	7541	11125	33954
	Missing	7544	6431	12053	8634	34662
	F_{CBEC}	0.953947	0.953608	0.938546	0.938160	0.946035
Error Correction Rate		0.644351	0.624082	0.582329	0.580909	0.605806

5. Concluding Remarks

CBEC was designed to be able to resolve the segmentation ambiguity problem and to provide accurate results for

different NLP applications. CBEC generates all possible candidates based on available dictionary and the word-like string set to ensure that the correct solution is not excluded. The maximal matching is employed in selecting the best candidate at first since it is fast and simple. An error of the selected candidate will be detected (if any) and will be corrected using word co-occurrence and error-based scoring.

The four different Thai text categories were used in both training and testing. CBEC has proved that it can increase the segmentation accuracy generated by classical maximal matching by 8%.

References

Aroonmanakun, W. 2009. Extracting Thai Compounds Using Collocations and POS Bigram Probabilities without a POS Tagger. *International Conference on Asian Language Processing*, 118-122.

Atlam, E.-S.; Ghada, E.; Fuketa, M.; and Aoe, J. 2003. A new algorithm for construction specific field terms using co-occurrence words information. *IEEE International Symposium on Micro-NanoMechatronics and Human Science,* 990- 993.

Boriboon, M.; Kriengket, K.; and Chootrakool, P.; Phaholphinyo, S.; Purodakananda, S.; Thanakulwarapas, T.; Kosawat, K. 2009. BEST Corpus Development and Analysis. *International Conference on Asian Language Processing,* 322-327.

Charoenpornsawat, P.; and Schultz, T. 2008. Improving word segmentation for Thai speech translation. *Spoken Language Technology Workshop,* 241-244.

Dang Duc Pham; Giang Binh Tran; and Son Bao Pham. 2009. A Hybrid Approach to Vietnamese Word Segmentation Using Part of Speech Tags. *Knowledge and Systems Engineering,* 154-161.

Haruechaiyasak, C.; Kongyoung, S.; and Dailey, M. 2008. A comparative study on Thai word segmentation approaches. *The fifth International Conference on Electrical Engineering/ Electronics, Computer, Telecommunications and Information Technology,* 125-128.

Huang, Degen; and Sun, Xiao. 2007. An Integrative Approach to Chinese Named Entity Recognition. *The Sixth International Conference on Advanced Language Processing and Web Information Technology,* 171-176.

Kang, S. S.; and Hwang, K. B. 2006. A Language Independent n-Gram Model for Word Segmentation. *AI 2006: Advances in Artificial Intelligence*, 557-565. Springer Berlin / Heidelberg.

Kosawat, K.; Boriboon, M.; Chootrakool, P.; Chotimongkol, A.; Klaithin, S.; Kongyoung, S.; Kriengket, K.; Phaholphinyo, S.; Purodakananda, S.; Thanakulwarapas, T.; and Wutiwiwatchai, C. 2009. BEST 2009 : Thai word segmentation software contest. *Eighth International Symposium on Natural Language Processing,* 83-88.

Kruengkrai, C.; and Isahara, H. 2006. A conditional random field framework for thai morphological analysis. *Proceeding of the Fifth International Conference on Language Resources and Evaluatio.*

Longman. 2011. Longman dictionary of contemporary English Online. available at *http://www.ldoceonline.com/* as of February.

Sutheebanjard, P.; and Premchaiswadi, W. 2009. Thai personal named entity extraction without using word segmentation or POS tagging. *Eighth International Symposium on Natural Language Processing,* 221-226.

Tirasaroj, N.; and Aroonmanakun, W. 2009. Thai named entity recognition based on conditional random fields. *Eighth International Symposium on Natural Language Processing, 2009.* 216-220.

Wartena, Christian; Brussee, Rogier; and Slakhorst, Wout. 2010. Keyword Extraction Using Word Co-occurrence. *Workshop on Database and Expert Systems Applications (DEXA)*, 54-58.

Ye Wang; and Shang-Teng Huang. 2005. Chinese word segmentation based on A-priori and adjacent characters. *Proceedings of 2005 International Conference on Machine Learning and Cybernetics,* 3808-3813.

Zhong-Hua Wang. 2003. Name entity recognition using language models. *IEEE Workshop on Automatic Speech Recognition and Understanding,* 554- 559.

Zhongjian Wang; and Araki, K.; Tochinai, K. 2008. Word Segmentation Method Based on Inductive Learning and Segmentation Rule. *International Symposium on Computational Intelligence and Design,* 95-98.

Zitao Liu; Wenchao Yu; Yalan Deng; Yongtao Wang; and Zhiqi Bian. 2010. A feature selection method for document clustering based on part-of-speech and word co-occurrence. *Seventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD),* 2331-2334.