氏　　名	NARARATWONG Rungsiman
学位(専攻分野)	博士 (情報学)
学 位 記 番 号	総研大甲第 2076 号
学位授与の日付	平成 31年 3 月22日
学位授与の要件	複合科学研究科　情報学専攻 学位規則第6条第1項該当
学 位 論 文 題 目	A Study on Thai Word Segmentation and an Analysis of Brand Crisis as Its Application
論 文 審 査 委 員	主　査　准教授　岡田　仁志 教授　越前　功 准教授　水野　貴之 准教授　南　和宏 教授　木下　宏揚　神奈川大学　工学部

(Form 3)

Summary of Doctoral Thesis

Name in full NARARATWONG Rungsiman

Title A Study on Thai Word Segmentation and an Analysis of Brand Crisis as Its Application

Chapter 1 introduces a Thai word segmentation problem and its role in an analysis of a brand crisis. Word boundary ambiguity has been a challenge in Thai language processing. Incorrect word segmentation may result in misleading interpretations. Chapter 2 explains Thai language fundamentals that are related to word segmentation. The chapter describes word formation, which is a sequential combination of words that form a new word. The roots of a compound word may have different meanings or can be interpreted differently from the word. Because of this difference, word segmentation may not produce a meaning that is similar to the meaning of the whole word, making the outcome ambiguous.

Chapter 3 proposes word segmentation rule and two post-processing algorithms to the existing machine-learning model, a Conditional Random Fields (CRF). The two proposed algorithms are word-merging and word-splitting algorithms. CRF is one of the most accurate word segmentation models among Thai word segmentation methods. The existing CRF-based word segmentation model was trained on Benchmark for Enhancing the Standard of Thai Language Processing (BEST2009) corpus developed by National Electronics and Computer Technology Center. The first problem is that the corpus does not address the compound word issue. In solving this problem, this study proposes changes to the original BEST2009 rule to prevent compound words with semantically relevant roots from being segmented and their meanings being altered. The rule of BEST2009 corpus stated that compound words with semantically relevant components should be segmented, but compound words with irrelevant components should not be segmented. The proposed rule stated that compound words, regardless of their relations to their components, should not be segmented. Based on this changed rule, this study proposed a dictionary-based algorithm that merges compound words after the CRF-based word segmentation. The algorithm merges any sequential combination of segmented words if the combined words are in a dictionary.

In the evaluation of the word-merging algorithm, one native Thai speaker relabeled part of BEST2009 for testing. The relabeling was done according to the proposed rule. The algorithm looks up its candidate words in three dictionaries – Wiktionary, LibThai, and LEXiTRON – and three named-entity dictionaries – BEST2009, LibThai, and GeoNames. The experiment consists of two conditions: condition (1) segments words using the CRF model alone, which is the method used in the previous study. The CRF

model was trained using BEST2009 corpus, which was created based on the original BEST2009 rule. Condition (2) performs the word-merging algorithm after the CRF model segmented the words. The CRF model in condition (2) was also trained on the same BEST2009 corpus as in condition (1). However, the segmented words were later merged by the word-merging algorithm, which followed the proposed rule. Finally, the result of each condition was compared to the relabeled corpus to measure the accuracy. The evaluation result indicates that applying the algorithm to condition (2) improves the accuracy by 12.14 percent on the test using the relabeled corpus. The evaluation of all combinations of the six dictionaries indicates a moderately positive correlation between the number of dictionaries and accuracy.

The second problem this study address is a sentence boundary ambiguity. A CRF model is among the most accurate sentence segmentation methods. The CRF model uses part-of-speech (POS) tags to increase its accuracy of sentence segmentation. The limitation is that POS-tagging algorithms cannot recognize some of the words due to limited training corpus. As a result, these words do not have POS tags, thus decreasing the accuracy of the CRF model. The proposed POS-based word-splitting algorithm in this study addresses this problem by splitting words that do not have POS tags if all of the segmented words can be tagged.

Since BEST2009 does not include POS tags, the word-splitting algorithm was instead tested against ORCHID corpus. ORCHID contains the POS tags, as well as word boundary and sentence boundary annotations necessary for the evaluation. Before the experiment, a benchmark had been established by training the CRF-based sentence segmentation model using ORCHID corpus with word and sentence annotations and POS tags. The CRF model was then tested using ORCHID corpus with only word annotations and POS tags. The experiment consists of two conditions: condition (1) segments words with the CRF model alone, which is the existing method, while condition (2) performs the proposed word-splitting algorithm after the CRF-based word segmentation. The result shows that the word-splitting algorithm in condition (2) tagged 1.39 percent more POS and was able to recover the average F1-score of sentence segmentation by 3.58 percent in relation to the loss margin. The recovery percentage was computed from the improvement of the F1-score from condition (1) to condition (2) divided by the loss of F1-score from the benchmark to condition (1).

The applications of the proposed algorithms were evaluated in three language processing tasks: Thai-to-English translation, summarization, and topic extraction. For the Thai-to-English translation, the proposed method looks for words that are not in dictionaries. These unrecognizable words are split if any parts of them can be found in the dictionaries. Finally, the method applies the word-merging algorithm to the text. This study hypothesized that the proposed method would repair incorrectly segmented words. The test corpus includes 50 Thai and English abstracts from journal articles. In condition (1) of the experiment, the words in the Thai abstracts were segmented by the

CRF model. In condition (2), the segmented texts were split and merged. All Thai abstracts were fed into a machine translation model created in this study and Google Translate. The English translations were compared to their human-translated references using Recall-Oriented Understudy for Gisting Evaluation (ROUGE) metrics. The test using Google Translate indicates an improvement in condition (2) over condition (1): ROUGE-1 = 1.12 percent, ROUGE-2 = 1.34 percent, and ROUGE-L = 1.24 percent.

For summarization, a TextRank summarization algorithm decides which sentences are the most important and should be in a summary. With inaccurate sentence segmentation, parts of important sentences may be omitted, while a segment of their less-important neighbors may be included. This study hypothesized that utilizing the word-splitting algorithm would improve sentence segmentation, which would eventually improve summarization. In the summarization experiment, the test corpus was 50 online articles across different topics, summarized by one native Thai speaker. In condition (1), the articles were segmented using the CRF model before being summarized. In condition (2), the segmented words were split before the summarization. The result indicates improvement in in condition (2) over condition (1): ROUGE-1 = 2.41 percent, ROUGE-2 = 2.08 percent, and ROUGE-L = 1.70 percent.

The problem with a topic extraction in Thai is that the segmented topic keywords with altered meaning can mislead human interpretation. This study hypothesized that by merging compound words, preserving their original meaning, would make the interpretation more accurate. The topic extraction model was evaluated using 2,000 tweets, half of which were related to flooding and the rest were related to traffic. Both corpora were fed into the Latent Dirichlet Allocation (LDA) and Hierarchical Dirichlet Process (HDP) topic extraction models. The words in the corpora were segmented by the CRF model, then merged. In the case of LDA, the result shows that 7.60 percent of topic keywords of the flood corpus and 10.00 percent of the keywords of the traffic corpus were merged. For HDP, the percentages were 23.60 and 16.00, respectively. The results show that the proposed methods can be used effectively in analyzing data obtained from social media. Hence, the following chapter explores the possibility of enhancing the proposed algorithms to be applied for social media analysis.

The results of the topic extraction showed that the proposed methods could be applied to social media analysis. Hence, Chapter 4 utilizes the proposed word-merging algorithm and the summarization method in order to examine whether this study can enhance analysis of a brand crisis in Thai social media. The analysis investigates the entertainment aspect of the crisis. The chapter proposes a conceptual framework that underlines a psychological process of the entertainment experience. The process begins with social media users, who are the audience, making a moral judgment of the brand and other involved parties based on five moral foundations. The foundations include care/harm, fairness/unfairness, loyalty/disloyalty, authority/subversion, and sanctity/degradation. This judgment then triggers their hedonic and non-hedonic

entertainment experience. For the hedonic dimension, the audience develops an affective disposition, leading to anticipation and enjoyment, while the non-hedonic dimension involves reflective thoughts, reinforcement of moral self and appreciation.

The framework was validated using content analyses in three studies. The first study used an English moral foundations dictionary created by Graham, Haidt, and Nosek (2009) to quantify moral foundations in Facebook comments related to brand crises. The study found evidence of moral judgment in all five moral domains. The second study extended the moral dictionary and found more topics of discussion related to the moral foundations. The third study summarized comments from Thai social media, then extracted moral words and validated their consistency with the English moral dictionaries. The study found that the public's moral judgment can be classified into five moral foundations. However, some of the dictionary's compound keywords were not found in the data which compound words were segmented. To solve this problem, the analysis was conducted in two conditions: condition (1) uses only the CRF model for word segmentation, and condition (2) merges compound words after the CRF-based word segmentation. The result shows that in condition (2), 12.47 percent more moral words were found in the data.

The chapter also demonstrates the possible application of the proposed word segmentation methods in analyzing the hedonic dimension. Its fourth study analyzed the dimension from the English Facebook comments. The study found three types of enjoyment in the comments: humor, satisfaction, and schadenfreude. This analysis can be conducted in Thai once a corpus is available to train a Thai sentiment analysis model, and the proposed word segmentation methods can be used for preparing data for the analysis in the future.

Lastly, Chapter 5 discusses the results of the word segmentation study, as well as the brand crisis study, including limitations of both studies. The chapter concludes that the proposed algorithms improve Thai language processing and facilitate human interpretation in the study of a brand crisis in social media.