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.

Form8・Separate Sheet
(様式8・別紙)

Results of the doctoral thesis screening

博士論文審査結果

Name in Full
氏　名　　NARARATWONG Rungsiman

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

本論文は，タイ語における単語分割に関する新しい手法を提案しこれをソーシャルメディアにおけるブランドクライシスの分析に応用することを目的とする．第１章は目的を示す．タイ語の言語処理では，単語および文の境界に関する曖昧性が課題となっていた．

第２章は問題の所在を示す．タイ語においては連続した語の組み合わせが新しい意味を持つ例として，合成語，複合語，反復語がある．複合語を構成する語は相互に類似した意味を持つ．反復語は同一の語の反復から成る．これに対して合成語を構成する語は互いに異なる意味を有する．このため，言語処理において合成語を分割することによって，意味の変化による曖昧性が生じるという問題があった．

第３章は先行研究の限界と解決方法を示す．第１の論点として，タイ語の単語分割においては，既存の機械学習モデルとしてConditional Random Fields (CRF)を適用することで良好な結果が得られることが知られていた．CRFを基礎とした既存の単語分割モデルは，National Electronics and Computer Technology Center (NECTEC)が開発したBenchmark for Enhancing the Standard of Thai Language Processing (BEST2009)コーパスを学習して完成された．だがBEST2009コーパスは合成語問題の解決を意識して作成されたものではないため，言語分割のための機械学習に用いることには限界があった．

この問題を解決するため，本論文はBEST2009のルールに修正を加え，合成語を構成する語の意味が合成語の意味と関連性を有する場合は，これを分割してはならないとする禁則を置いた．BEST2009のルールではこれを分割して処理するため，提案手法では，既存のCRFモデルを適用して単語分割を行ったのちに，辞書的アルゴリズムに従って連続した語を合成語として再構成する方法をとる．このための準備として，BEST2009コーパスに禁則に該当することを示すラベルを付した．提案手法を評価するため，２つの条件下で実験を行った．条件１では先行研究が適用するBEST2009のルールに準拠し，合成語を構成する語の意味が合成語の意味と関連性を有する場合にはこれを分割し，合成語を構成する語の意味が既存の合成語の意味と関連性を有しない場合および複合語もしくは反復語はこれらを分割しないというルールを適用した．条件２ではBEST2009のルールを修正し，合成語，複合語，反復語はこれを構成する語との関連性に関わらず常に分割しないとするルールを置いた．いずれの実験でも，CRFモデルによって単語分割を行った後にアルゴリズムを適用した．アルゴリズムが候補語を探索する対象としてタイ語の辞書データWiktionary, LibThai, LEXiTRONおよび固有表現データBEST2009, LibThai, GeoNamesを利用した．実験の結果，提案アルゴリズムは既存手法に比べて全体として正確性を12.14%改善することが示され，６種類の辞書データ全てにおいて有意な改善が示された．

第２の論点として，タイ語の言語処理においては，文境界の曖昧性においても課題があった．CRF モデルは文境界の検出においても良好な結果を示すが，part-of-speech (POS) タグを用いていることに起因する限界があった．すなわち，コーパスに依存する機械学習の制約から，POS タグを持たない語が増加する傾向にあり，CRF モデルの正確性が減少していた．そこで本論文は，POS タグを持たない語が分割可能であって，かつ，分割された語の全てにタグを付すことが可能である場合には，これを分割された語として扱うというアルゴリズムを提案した．機械学習の対象として，POS タグおよび文境界情報のアノテーションを持つ ORCHID コーパスを利用し，文境界検出の正確性が達成された時点でこれをベンチマークとして，２つの条件下で実験を行った．条件１では CRF モデルを適用して単語分割を行い，次に CRF モデルを適用して文境界検出を行った．条件２では CRF モデルを適用して単語分割を行い，次に提案アルゴリズムを適用して単語分割を行い，続いて CRF モデルを適用して文境界検出を行った．その結果，条件２では条件１と比べて POS を付す語の数が 1.39%向上し，ロスマージンを示す F1 スコアの平均値が 3.58%改善された．次に，提案アルゴリズムをトピック抽出に適用して検証した．トピック抽出においては単語分割を伴うため，これを人が解釈する際に誤りが生じることから，分割された語を合成語に再構成する提案手法が正確性を改善すると仮定した．仮説を検証するため，タイ語による洪水に関するツイート 1000 件および交通に関するツイート 1000 件から成る，合計 2000 件のコーパスを対象として，Latent Dirichlet Allocation (LDA)および Hierarchical Dirichlet Process (HDP)を適用してトピックを抽出した．２つの条件を置き，条件１では CRF モデルによって単語を分割した．条件２では CRF モデルによって単語を分割した後，提案アルゴリズムに従って単語を再分割し，かつ再合成した．条件１との比較において条件２の結果をみると，LDA を適用した例では洪水の関連語が 7.60%，交通の関連語が 10.00%それぞれ多く抽出された．HDP を適用した例では，洪水の関連語が 23.60%，交通の関連語が 16.00%それぞれ多く抽出された．この結果から，提案アルゴリズムのソーシャルメディアへの応用可能性が示唆された．

第４章では，提案アルゴリズムの一般的な応用可能性を検証するため，ソーシャルメディアにおけるブランドクライシスの分析に適用した．ブランドクライシスに関する先行研究を調査し，Graham etc. (2009)が提案する倫理的評価のための５要素を基礎として，これにエンターテイメント分析のモデルを融合した分析フレームワークを設定した．フレームワークの前段では，Graham etc. (2009)が提案する倫理的評価の 5 要素に準拠して，Harm/care, Fairness/unfairness, Ingroup/loyalty, Authority/respect, and Purity/sanctity の軸に関連する発言を抽出する．後段では，エンターテイメント分析のモデルに準拠して，Hedonic/non-hedonic の軸に関連する発言を抽出する．これらを統合する分析フレームワークを設定した．フレームワークの前段を検証するため，ブランドクライシスに関するソーシャルメディアの発言を対象に分析を行った．Graham etc. (2009)が構築した moral foundation dictionary をソーシャルメディアの英語公開データにおけるブランドクライシスを対象とした発言にあてはめたところ，倫理的評価の５要素の全てに関連する語が抽出されることが確認された．次に，Graham etc. (2009)が構築した moral foundation dictionary に拡張を加え，これを同一のデータにあてはめたところ，抽出される関連語が増加することが観測された．これらをタイ語に応用する可能性について検証す

るため，ソーシャルメディアのタイ語公開データにおけるブランドクライシスを対象とした発言に，拡張された moral foundation dictionary をあてはめた．その結果，倫理的評価の 5 要素の全てに関連する語が抽出されることが確認された．ここで，本論文が提案するアルゴリズムの有用性を検証するため，2 つの条件下で実験を行った．条件 1 では，既存の CRF モデルに従って単語分割を行った．条件 2 では，既存の CRF モデルで単語分割を行ったのち，提案アルゴリズムに従って合成語として再構成した．その結果，提案アルゴリズムによる場合は，抽出できる関連語の数を 12.47%増加させることが確認された．フレームワークの後段を検証するため，これをソーシャルメディアの英語公開データにおけるブランドクライシスを対象とした発言にあてはめたところ，Hedonic/non-hedonic の軸に関連する語として humor, satisfaction, schadenfreude の 3 つが抽出された．これをタイ語環境で適用するためには，タイ語のコーパスにおいて感情分析モデルを適用できることが前提となるため，ここにおいて提案アルゴリズムの有用性を検証する可能性については今後の課題とする．

第 5 章では，タイ語における単語分割および文境界の正確性に関する研究と，その応用としてのソーシャルメディアにおけるブランドクライシスに関する検証を統括し，その有用性と限界について議論する．そして，提案アルゴリズムがタイ語環境のソーシャルメディアの分析において有用性を発揮することが，ブランドクライシスをはじめとする現象の正確な分析に貢献すると結論する．

出願者は，本論文に関して，電子情報通信学会英語論文誌 IEICE TRANSACTIONS on Information and Systems, E101-D (12)に査読付論文が採択されている．また，査読付の国際会議 International Conference on Communication and Computer Engineering において研究を報告している．出願者は，審査会において本論文の内容を章立てに沿って説明し，審査委員からの質問に対して的確に回答した．なお，審査委員から本論文の題目を研究内容の明確化のため変更するよう提案があり，出願者は論文題目を変更することとした．以上の理由により，審査委員会は，本論文が学位の授与に値すると判断した．