เครื่องมือตรวจจับข้อความที่เป็นการระรานทางไซเบอร์สำหรับภาษาไทยบนสื่อสังคมออนไลน์

TGF-GRU: A Cyber-bullying Autonomous Detector of Lexical Thai across Social Media

[1]Pakpoom Mookdarsanit , [2]Lawankorn Mookdarsanit

[1]Department of Computer Science, Faculty of Science, Chandrakasem Rajabhat University

[2]Department of Business Computer, Faculty of Management Science, Chandrakasem Rajabhat University

Email: pakpoom.m@chandra.ac.th[1*]

Received: June 28, 2019

Revised: August 13, 2019

Accepted: September 5, 2019

บทคัดย่อ

ความขัดแย้งของคนบนสังคมออนไลน์ในประเทศไทยที่ดำเนินมาอย่างต่อเนื่องนั้น หนึ่งในสาเหตุหลักมีผลมาจากการระรานทางไซเบอร์บนเครือข่ายทางสังคมออนไลน์แทบทั้งสิ้น โดยมีวัตถุประสงค์เพื่อให้ร้ายต่อบุคคลอื่นให้เกิดความเสียหาย โดยเฉพาะอย่างยิ่งในเรื่องของการเมือง บทความวิจัยนี้ได้พัฒนาแบบจำลองทางภาษาเพื่อใช้ในการตรวจจับข้อความที่เป็นการระรานทางไซเบอร์บนเครือข่ายสังคมออนไลน์ แบบจำลองที่พัฒนาขึ้นใช้พื้นฐานของ “จีอาร์ยู” ซึ่งก่อนถึงขั้นตอนการดำเนินงานของจีอาร์ยูนั้น จะมีอัลกอริทึมสำหรับลดมิติของข้อมูล โดยตั้งชื่อว่า “ทีจีเอฟ” ซึ่งรวมกันเรียกว่า “ทีจีเอฟ-จีอาร์ยู” โดยที่แบบจำลองดังกล่าวนี้ได้ถูกสร้างและคำนวณมาจากโพสต์หรือความคิดเห็นข้อความภาษาไทยจากเฟซบุ๊กจำนวน 10,900 ข้อความ ผลการทดลองแสดงให้เห็นว่า ทีจีเอฟสามารถช่วยเพิ่มความแม่นยำให้กับจีอาร์ยูได้ถึงร้อยละ 8.41 กับเวลาในการประมวลผลที่เพิ่มมาอีกเล็กน้อย อย่างไรก็ตาม การใช้ภาษาไทยเพื่อการสื่อสารนั้นเป็นที่รู้กันในท้องถิ่นว่ามีคำพ้องความหมาย คำพ้องรูป และการประชด ซึ่งจะมีผลให้การตรวจจับข้อความภาษาไทยโดยใช้คอมพิวเตอร์มีโอกาสผิดพลาด โดยสรุปแบบจำลองทีจีเอฟ-จีอาร์ยู สามารถใช้เป็นเครื่องมือเสริมทางปัญญาประดิษฐ์ของเฟซบุ๊กสำหรับตรวจจับข้อความภาษาไทยที่ไม่เหมาะสม ก่อนที่ผู้ใช้จะทำการโพสต์หรือแสดงความคิดเห็นได้ ในอนาคตอันใกล้นี้ข้อความที่เป็นการระรานทางไซเบอร์ (เช่น การดูหมิ่น ความคิดเห็นที่เกี่ยวกับเพศ คำพูดที่สร้างความเกลียดชัง ข่าวลือ การใส่ร้ายป้ายสี และการระรานรูปแบบอื่น ๆ) จะถูกตรวจจับและกรองออกอัตโนมัติทันที และเมื่อนั้นชนวนที่จะทำให้เกิดความแตกแยกในสังคมก็จะค่อย ๆ ลดลง โดยให้เครื่องมือตรวจจับข้อความที่เป็นการระรานทางไซเบอร์ทำงานด้วยตัวของมันเอง

คำสำคัญ: การตรวจจับข้อความที่เป็นการระรานทางไซเบอร์, การวิเคราะห์อารมณ์จากความคิดเห็น, การทำความเข้าใจภาษาธรรมชาติ, จีอาร์ยู, การปรับและคัดกรองไวยากรณ์ภาษาไทย, การแบ่งคำภาษาไทย

Abstract

Continually, one of the most the fragile states in Thailand are originated from cyber-bullying across social media networks (OSNs). Cyber-bullying intentionally is plotted to offend other people, particularly in politics. This paper develops a novel linguistic model to detect the Thai-bullying label on OSNs. Our model is

based on “Gated Recurrent Unit (GRU)” that has a pre-process for dimensional reduction algorithm called “Lexical Thai Grammatical Filtering (TGF)”. Our developed TGF-GRU is formulated by the 10,900 Thai texts from posts/comments on Facebook. From the results, TGF can improve the accuracy of normal GRU as 8.41% with a little time consumption. Notwithstanding, some synonyms, homographs or insinuations of Thai jargons can easily confuse the detection. In a nutshell, TGF-GRU model will be able to be used as an additional AI feature to autonomously detect the inappropriate Thai text before a user posts or comments on Facebook. For years, some cyber-bullying labels (e.g. pejorative, sexual comment, hate speech, rumor, slandering, etc.) will have been autonomously detected and filtered out; the causes of social fragile state will be gradually mitigated by Thai-bullying detector.

Keywords: Bullying Detection, Sentiment Analysis, Natural Language Understanding, Gated Recurrent Unit (GRU), Thai Grammatical Filtering, Thai Word Segmentation

1. Introduction

Thai is a spoken-language in a family of Kra-dai that was derived from Pali, Sanskrit, Khmer and Mon [1]. From the historical evidence, Thai ancient alphabets were originally inscribed on many memorial stones by King Ramkhamhaeng of Sukhothai [2]. Heretofore, Thai is one of an official language in ASEAN that has almost 70 million native speakers [3]. Grammatically, Thai is a tonal language that the “same pronunciation but different tones” may easily communicate in the different meaning. Such a same word as “ma”, one tone of ma is a verb, “come (Thai: มา)”, a higher tone means “grand-mom (Thai: ม่า)” and “horse (Thai: ม้า)” and the highest tone as “dog (Thai: หมา)”. In statistical machine translation, there is no space segmentation between Thai words [4][5] (unlike English) that is one of a well-known challenge in researches about Thai natural language processing (Thai-NLP), particularly in linguistic textual corpus.

Until the digital era, all amounts of public textual contents on the internet are written in Thai [6] around 0.3% of other languages on the whole internet. Despite this, some of them can be seen as “cyber bullying (aka online bullying)” [7][8] that plot to offend other people across Online social media networks (OSNs) e.g. pejorative, sexual comment, hate speech, rumor, slandering or any other negative posting against others, particularly in politics. From the statistics, more than 80% of overall bullying-contents in OSNs are produced on Facebook that has the 2.38 billion active users [9].

Intrinsically, Thai natural language processing [10][11] that can be categorized into Text-to-speech conversion [12][13][14]. Thai language understanding [15-18]. Thai word segmentation [19-23] and statistical machine translation [24-26].

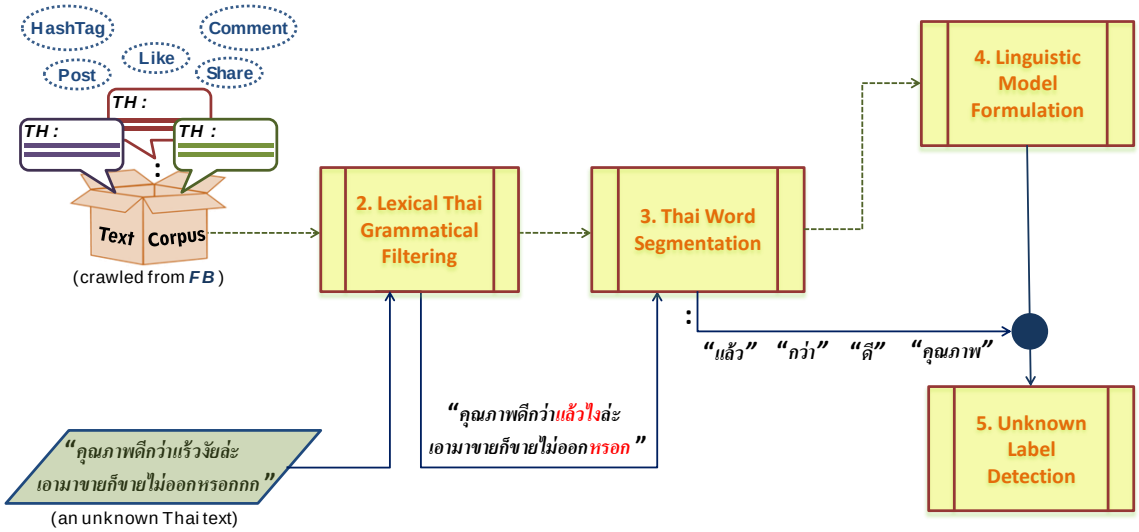

Figure 1 : The framework of Thai-textual cyber-bullying detection using TGF-GRU model

This paper develops an intelligent linguistic model based on Gated Recurrent Unit (GRU) [27] to detect the bullying labels from lexical Thai contents (comments or posts) on OSNs. For the pre-processing, the lexical Thai grammatical Filtering is used to segment the sentence/phrase into many words that our proposed model is denominated "Lexical Thai Grammatical Filtering Lexical Thai Grammatical Filtering and Gated Recurrent Unit (TGF-GRU)". The overall TGF-GRU model is shown in Figure 1.

For the model formulation, the corpus (or textual dataset) keeps the 10,900 Thai texts that are crawled from posts and comments from Facebook. Prior to the linguistic model formulation (in section 4), these texts are filtered and corrected some grammatical errors and noises/outliers (as "lexical Thai grammatical filtering" in section 2) and sentence-by-sentence segmented into many Thai words (as "Thai word segmentation" in section 3).

For unknown label detection (in section 5), the unknown Thai text are also checked by section 2 and 3. Finally, the TGF-GRU analyzes whether the bullying label or not. From the experimental results, our TGF-GRU improves the detection accuracy as 8.41% compared to the traditional GRU.

This paper is divided into 6 sections. The section 2 describes "Lexical Thai Grammatical Filtering". The topic "Thai Word Segmentation" is written in section 3. Section 4 and Section 5 talk about "Linguistic Model Formulation" and "Unknown Label Detection". The conclusion is in section 6.

2. Lexical Thai Grammatical Filtering

To formulate the bullying label detection, the computer model must understand the sentimental contents within the textual information across online social media networks (OSNs).

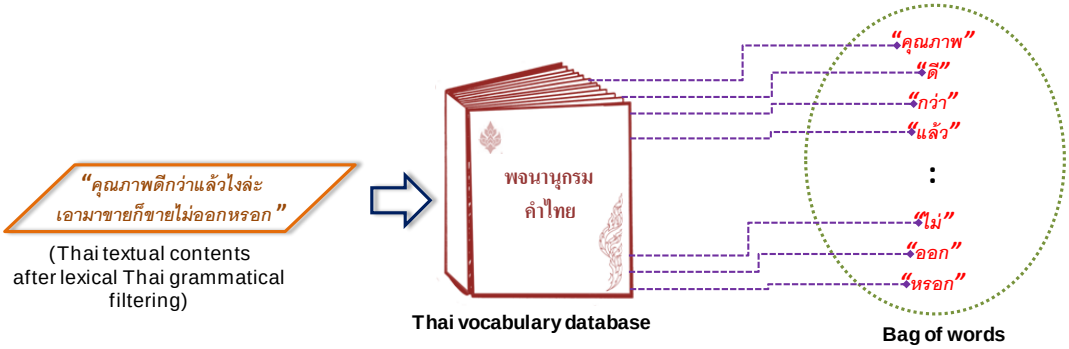

Figure 2 : The workflow of Thai word segmentation

Notwithstanding, some many unofficial words are widely used over the internet that easily make the model wrongly understands the meaning of text. All most often used unofficial words are listed in our "volatile linked-list" to check.

Prior to the word segmentation, all ungrammatical words are replaced by the correctly grammatical Thai words as shown in Table 1.

Table 1 Some Thai Word Replacement

Unofficial Words	Replaced by
แร่ว, แร้น, แร้ว, แล้น, ล้าว	แล้ว
งัย, งาย	ไง
เตง, ตะเอง	ตัวเอง
เมพ	เทพ
จิม, จริม	จริง

For the dimensionality reduction, some repeated alphabets (such as a number of "ว (s)" in the text "รู้ แ ร้ ว ว ว ว ว ว ว ว ว ว ว ว ว ว ว ว ") are seen as "noise/outlier" that is eventually filtered out.

3. Thai Word Segmentation

The filtered Thai text is segmented into Thai words by "vocabulary entity matching". For the vocabulary dataset, all vocabularies are automatically crawled from the source of online Thai dictionary (Thai: พจนานุกรมออนไลน์) and covers almost all Thai vocabularies. All crawled vocabularies are stored in "Thai vocabulary database" term of "vocabulary entity". As shown in Figure 2, Thai text is segmented into many words using the matching between those entities in the dictionary and words within the text. All words are stored in a storage called "bag of words".

4. Linguistic Model Formulation

Since the internet users read a post/comment across online social media networks (OSNs) as Thai textual information, the one is trying to understand each word within the text that needs to understand all previous words. For this reason, all segmented words (from the previous section) are formulated by Recurrent Neural Network (RNN).

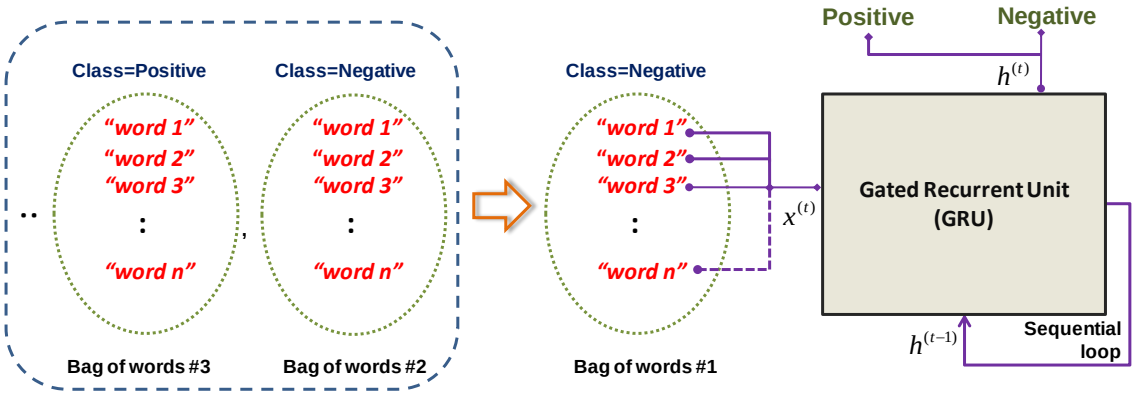

Figure 3 : Linguistic model formulation using gated recurrent unit (GRU)

RNN is well-known and appropriate for a sequence of events with the connection between previous tasks, particularly in language understanding. According to the speed and performance, Gated Recurrent Unit (GRU) – a simple version of RNN, is used for the model formulation. Since GRU can be seen a light-weight version of Long-Short Term Memory (LSTM). GRU is faster and less processing power than LSTM. In our model, the classification of the answer is categorized into "Non-bullying" and "Bullying" label. Technically, all words of a text (aka a bag of words) with its label tagging are input to the GRUs. Each word is analyzed by one GRU circuit. And the previous state affects the next GRU, repeatedly as shown in Figure 3.

All 10,900 textual information (aka a number of 10,900 bags) from any posts/comments are collected from Facebook.

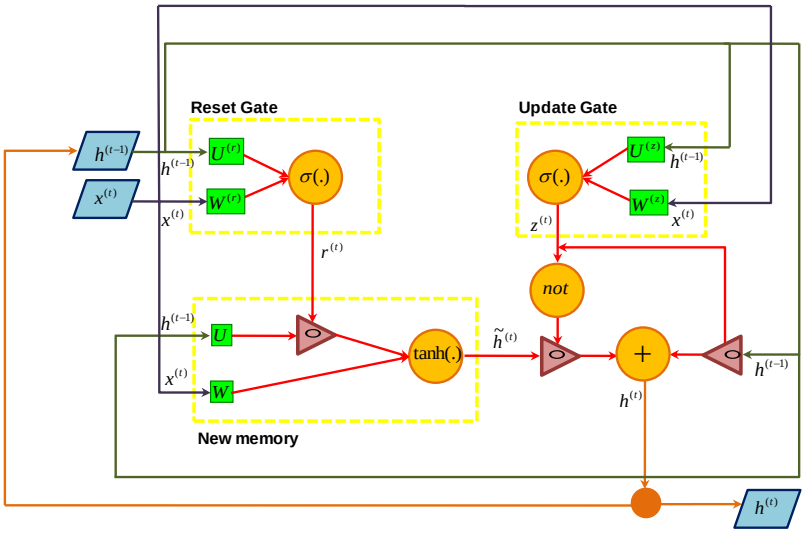

Figure 4 : Architecture of gated recurrent unit (GRU)

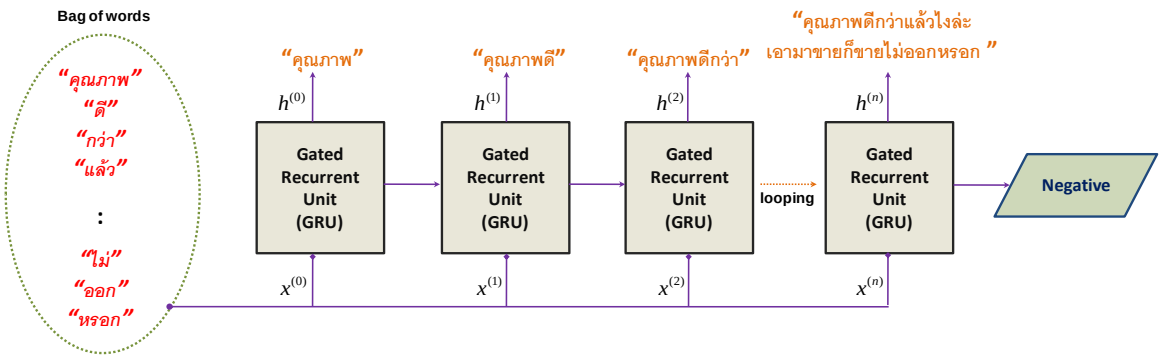

Figure 5 : The **w**orkflow of unknown label detection

Basically, the architecture of GRU circuit (as shown in Figure 4) consists of update gate ($z^{(t)}$), reset gate ($r^{(t)}$), new memory ($\tilde{h}^{(t)}$) and hidden state that can be computed by the equation (1)-(4) [27], respectively.

$$z^{(t)} = \sigma\left(W^{(z)}x^{(t)} + U^{(z)}h^{(t-1)}\right) \quad (1)$$

$$r^{(t)} = \sigma\left(W^{(r)}x^{(t)} + U^{(r)}h^{(t-1)}\right) \quad (2)$$

$$\tilde{h}^{(t)} = \tanh\left(r^{(t)} \bullet Uh^{(t-1)} + Wx^{(t)}\right) \quad (3)$$

$$h^{(t)} = \left(1 - z^{(t)}\right) \bullet \tilde{h}^{(t)} + z^{(t)} \bullet h^{(t-1)} \quad (4)$$

5. Unknown Label Detection

Since an unknown Thai label is input to detect whether bullying label or not. The unknown label can either pass or bypass the "Lexical Thai Grammatical Filtering" in section 2.

Table 2 Comparison between TGF-GRU and GRU

Method	Accuracy		Time Consumption
	AVG	SD	
TGF-GRU	84.21	0.0046	160.1 s
GRU	77.68	0.0052	151.9 s

Table 3 Some Example of Wrong Detection in Thai

Thai Textual Information	Wrong Detection
"ท่านจะยุในจัยผมไปตลอดกาล ผมจะเก็บทุกเรื่องราวสุดยอดเดรัจฉานวิชาของท่านไว้เป็นอย่างดี"	Positive
"ยี่ห้อนี้เมพชิง ๆ ราคาสูงอย่างกับต้นตาล แต่พอเวลาผ่านไป ราคาตกลงเร็วอย่างกับน้ำเชื่อมบูด ขวัญใจรถยกตัวจริงเลยล่ะครับ"	Positive
"ไม่เห็นแปลกใจเลย... ถึงแม้ว่าจะเป็นโฉมใหม่ รถซื้อแกง ก็คือ รถซื้อแกง ออพชั่นน้อย เครื่องเก่า แถมราคาแพง แต่ก็ยังขายดีอยู่ คงเป็นเพราะแบรนด์นี้ขึ้นชื่อว่าซ่อมง่าย และที่สำคัญคือ ทนอย่างกะแรด"	Negative

Later, the section 3 is to segment those Thai label into many words. Finally, the bullying label from the summarization of these words is recursively detected by the linguistic model formulation. The overall workflow of unknown label detection is shown in Figure 5.

We compared the results using Accuracy that can be computed by TP, TN, FP and FN, respectively.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

Our designed TGF-GRU (aka Thai Word Segmentation and Gated Recurrent Unit) and general GRU are compared in Table 2 (based on 2,200 unknown Thai labels).

Even if TGF-GRU spends more time than GRU, the TGF-GRU improves the accuracy of bullying detection. Due to the various synonyms (Thai: คำพ้องความหมาย), homographs (Thai: คำพ้องรูป) and also insinuations (Thai: การประชด) of Thai used in the internet that produce some understanding errors. (Note that: we cannot show such dirty textual contents in this paper)

In heritage view, these vague, tonal and sarcastic meanings, unique to Thai jargon with its alphabets, can be seen as "Thai-ness" that distinctively and attractively differs from any other languages.

6. Conclusion

Since Thai is an official language in Thailand, ASEAN that is spoken by almost 70 million people. The specialty of Thai is the different tones of the same word and no space between words that can be seen as the research challenge in Thai natural language processing (Thai-NLP). From the statistics, the 0.3% of internet contents is written in Thai. Some of them is cyber-bullying that intends to attack other people, particularly in online social media network (OSNs) like Facebook. This paper develops a linguistic model called "Lexical Thai Grammatical Filtering and Gated Recurrent Unit (TGF-GRU)" to detect the bullying contents. Our developed TGF-GRU is formulated by 10,900 Thai textual contents that are crawled from any posts/comments on Facebook. The average accuracy of TGF-GRU is 84.21%, except for some synonyms, homographs and insinuations of Thai jargons. For future work, many new Thai vocabularies will be used on the internet according to the language revolution. The Thai linguistic corpus always needs to be updated that should be used the domain adaptation algorithms to grow the knowledge. Without regarding to a large number of new Thai vocabularies, the human's understanding is still based on the understanding of previous words until the end of Thai textual sequence.

7. Reference

[1] Satienkoses, Y. (1981). **Essays on Thai Folklore.** Bangkok: Duang Kamol.

[2] Inthajakra, L., Prachyapruit, A. & Chantavanich, S. (2016). The Emergence of Communication Intellectual History in Sukhothai and Ayutthaya Kingdom of Thailand. *Social Science Asia*, 2(4): 32-41.

[3] World Bank. (2018). Population, Total. Retrieved on May 15, 2019, from https://data.worldbank.org/indicator/SP.POP.TOTL?locations=TH

[4] Haruechaiyasak, C., Kongyoung, S. & Dailey, M. (2008). A comparative study on Thai word segmentation approaches. In proceedings of the 5th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (125-128). Krabi, Thailand: IEEE.

[5] Klahan, A., et al. (2018). Thai Word Safe Segmentation with Bounding Extension for Data Indexing in Search Engine. In Proceedings of the 14th International Conference on Computing and Information Technology (83-92). Chiang Mai, Thailand: Springer.

[6] W3 Techs. (2018). Historical trends in the usage of content languages for websites. Retrieved on May 15, 2019, from https://w3techs.com/technologies/history_overview/content_language

[7] Smith, P. K., et al. (2008). Cyberbullying: its nature and impact in secondary school pupils. *The Journal of Child Psychology and Psychiatry*, 49(4): 376-385.

[8] Sittichai, R. & Smith, P. K. (2018). Bullying and Cyberbullying in Thailand: Coping Strategies and Relation to Age, Gender, Religion and Victim Status. *Journal of New Approaches in Educational Research*, 7(1): 24-30.

[9] Social Media Today. (2018). Facebook Reaches 2.38 Billion Users, Beats Revenue Estimates in Latest Update. Retrieved on May 15, 2019, from https://www.socialmediatoday.com/news/facebook-reaches-238-billion-users-beats-revenue-estimates-in-latest-upda/553403/

[10] Koanantakool, T., Karoonboonyanan, T. & Wutiwiwatchai, C. (2009). Computers and the Thai Language. *IEEE Annals of the History of Computing*, 31(1): 46-61.

[11] Sornlertlamvanich, V., et al. (2000). The state of the art in Thai language processing. In proceedings of the 38th Annual Meeting on Association for Computational Linguistics (1-2). Stroudsburg, PA, USA: ACM.

[12] Mittrapiyanuruk, P., et al. (2000). Issues in Thai Text-to-Speech Synthesis: The NECTEC Approach. In proceeding of NECTEC Annual Conference (483-495). Bangkok, Thailand: NECTEC.

[13] Chinathinmatmongkhon N., Suchato, A. & Punyabukkana, P. (2008). Implementing Thai Text-to-Speech Synthesis for Hand-held Devices. In proceeding of the 5th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (125-128). Krabi, Thailand: IEEE.

[14] Wutiwiwatchai, C. & Furui S. (2007). Thai speech processing technology: *A review. Journal Speech Communication*. 49(1): 8-27.

[15] Assawinjaipetch, P., et al. (2016). Recurrent Neural Network with Word Embedding for Complaint Classification. In proceedings of the Third International Workshop on Worldwide Language Service Infrastructure and Second Workshop on Open Infrastructures and Analysis Frameworks for Human Language Technologies (36-43). Osaka, Japan: ACL Technology.

[16] Nomponkrang, T. & Sanrach C. (2016). The Comparison of Algorithms for Thai-Sentence Classification. *International Journal of Information and Education Technology*. 6(10): 801-808.

[17] Sukakanya, U. & Porkaew , K. (2009). Comparison of Classification Techniques for Thai Web Document Classification. World Academy of Science, Engineering and Technology. 2009: 895-901.

[18] Haruechaiyasak, C., et al. (2013). S-Sense: A Sentiment Analysis Framework for Social Media Sensing. In proceedings of the 6th International Joint Conference on Natural Language Processing (6-13). Nagoya, Japan: The Association for Computational Linguistics.

[19] Nararatwong, R., et al. (2018). Improving Thai Word and Sentence Segmentation Using Linguistic Knowledge. IEICE Transactions on Information and Systems. E101.D(12): 3218-3225.

[20] Lapjaturapit, T., Viriyayudhakom, K. & Theeramunkong, T. (2018). Multi-Candidate Word Segmentation using Bi-directional LSTM Neural Networks. In proceedings of the 2018 International Conference on Embedded Systems and Intelligent Technology & International Conference on Information and Communication Technology for Embedded Systems (1-6). Khon Kaen, Thailand: IEEE.

[21] Theeramunkong, T. & Usanavasin S. (2003). Non-dictionary-based Thai word segmentation using decision trees. In proceedings of the first international conference on Human language technology research (1-5). San Diego: ACM.

[22] Theeramunkong, T. & Tanhermhong T. (2004). Pattern-Based Features vs. Statistical-Based Features in Decision Trees for Word Segmentation. IEICE TRANSACTIONS on Information and Systems. E87-D(5): 1254-1260.

[23] Boonkwan, P. & Supnithi, T. (2017). Bidirectional Deep Learning of Context Representation for Joint Word Segmentation and POS Tagging. In proceedings of the 5th International Conference on Computer Science, Applied Mathematics and Applications. Berlin, Germany: Springer.

[24] Lyons, S. (2016). Quality of Thai to English Machine Translation. In proceedings of Pacific Rim Knowledge Acquisition: Workshop on Knowledge Management and Acquisition for Intelligent Systems (261-270). Phuket, Thailand: Springer.

[25] Luekhong, P., et al. (2016). A Study of a Thai-English Translation Comparing on Applying Phrase-Based and Hierarchical Phrase-Based Translation. In proceeding of international Symposium on Natural Language Processing (38-48). Ayutthaya, Thailand: Springer.

[26] Sutantayawalee, V., et al. (2017). Improvement of Statistical Machine Translation using Charater Based Segmentation with Monolingual and Bilingual Information. In proceedings of the 28th Pacific Asia Conference on Language, Information and Computation (145-151). Phuket, Thailand: ACL Technology.

[27] Chung, J., et al. (2014). Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. arXiv. 1412.3555: 1-9.