Natural Language Processing for Disaster Management Using Conditional Random Fields

Hathairat Ketmaneechairat and Maleerat Maliyaem
College of Industrial Technology Faculty and Information Technology Faculty, King Mongkut's University of Technology North Bangkok, Thailand
Email: hathairat.k@cit.kmutnb.ac.th

***Abstract*—This research aims to extract name entity mentioned in unstructured text into a predefined category using Conditional Random Field (CRF) and bidirectional Long Short-Term Memory (LSTM). The experiments were conducted using one thousand words which extracted from the collection of twitter massage that collected in the topic related to natural disaster and classify into six classes of the output. There are three scenarios for testing and evaluate: CRF, CRF-optimize and a combination of LSTM and CRF. The results show that CRF-optimize parameter performance is given better than other model with 98.94%, 98.95% and 98.93% for precision, recall and F-measure respectively.**

***Index Terms*—natural disaster, natural language processing, information extraction, conditional random field, named entity recognition**

I. Introduction

Nowadays, the data communication via the Internet has higher communication speeds more than the past. As a resulting, the development of technology is occurring in various fields. The Internet access can be done easily. The most people can access the Internet for all ages. The service provider has developed a system to support a lot of people. One of the most popular services is online social media services. For example, the popular online social media is Facebook, Twitter and Instagram. In this research, we study information extraction method from the messages that users have published on Twitter and Instagram services, which are social media that looks like a microblog. The users will publish information in a short message and concise manner. The Twitter is a social media service that allows users to publish messages up to 280 characters at a time. In the main part of Instagram, the users can send the image and also add subtitles related to that image.

It is widely known that social media users are widely used. Because of the most users can get the information quickly and switch to using information via social media instead of traditional media such as newspapers, television, radio and magazines. With the ease of access and dissemination of information, many users have also changed their role as news publishers. The traditional media producers have also changed the way to broadcast the news and add more channels to broadcast the news on social media. Nowadays, the online social media has a lot of information that is published in each day. The information is diverse in many areas, whether it is the type of information or pattern of information. The users often publish about daily life events, events have been found and disseminate information from other media sources such as text, images and audio. This information will be useful only if the information can select and search methods of extraction information for the purpose of using as much as possible.

Natural disasters are one topic that has a lot of interest topic because of the damage and the consequences of natural disasters are often severe. The occurrence of natural disasters is difficult to wo preventing. If it happens then there will be a method for managing natural disaster events to receive the least damage and impact, to help sufferers quickly and efficiently as possible. If referring to information technology, finding appropriate methods for applying information in various fields is one of the most effective ways to support disaster management and assistance. Which the researcher is interested in bringing information from online social media sources, which is an enormous amount of information and current information from the research found that during the natural disaster event. The users have used social media to disseminate information about the disaster that has occurred [1] using information from Twitter and Instagram. In text format, the information is in the pattern of social media online, the information is quite complicated. Because of this type of information is classified as unstructured data. The use of Natural Language Processing (NLP) technology is one of artificial intelligence. In this research, the researchers aim to find the ways to extract information into classes that represent that massage posted on social media. Conditional Random Field (CRF) and Long Short-Term Memory (LSTM) are selected to use as a classification model for extract name entity from unstructured massage. It is a part of Natural language processing that will be discussed in the next section.

II. Related Works

A. Natural Langauge Processing (NLP)

Natural Language Processing is the technology used to aid computers to understand the human's natural language. NLP is one of the most important technologies of the information age a crucial part of artificial intelligence (AI).

Manuscript received November 6, 2019; revised April 6, 2020.

doi: 10.12720/jait.11.2.97-102

Fully understanding and representing the meaning of language is an extremely difficult goal. The challenges in natural language processing frequently involve speech recognition, natural language understanding, and natural language generation. There is a fast-growing collection of useful applications derived from NLP such as spell checking, keyword search, finding synonyms, machine translation, spoken dialog systems and complex question answering. A natural language understanding system have knowledge about what the words mean, how words combine to form sentences, how word meanings combine to from sentence meanings and so on. Parts of Speech (POS) are an approach to perform semantic analysis and include the process of assigning one of the parts of speech to the given word. POS include nouns, verbs, adverbs, adjectives, pronouns, conjunction and their sub-categories. NLP is considered a difficult problem in computer science. It is the nature of the human language that makes NLP difficult. The rules that dictate the passing of information using natural languages are not easy for the computers to understand. It is very important to know how the state-of-the-art techniques work in many different languages. There are hundreds of natural languages, each of which has different syntax rules. Words can be ambiguous where their meaning is dependent on their context. The example of the word in Thai language is different semantic such as “ตากลม” can be divided into two patterns, “ตา-กลม” and “ตาก-ลม”. Therefore, NLP is challenging and interesting research.

NLP is using a large amount of data source or big data. Most NLP data is come from online social media. There is a lot of researchers do the research about NLP and online social media. M.B. Habib and M. van Keulen [2] presented NEED4Tweet, a Twitterbot for Name Entity Extraction (NEE) and Name Entity Disambiguation (NED) in tweets. F. Liu *et al*. [3] developed automatic identification locative expressions from a variety of Social Media Text. J. Lingad *et al*. [4] presented an experimental study to quantify the potential of Named Entity Recognizers (NER) in location extraction in tweets. NLP and online social media can be used to support disaster management and classification disaster data [5], [6]. D. Küçük and R. Steinberger [7] reported the NER experiments on Turkish tweets in order to determine facilitating and impeding factors during the development of a NER system for Turkish tweets which can be used in social media analysis applications.

B. Named Entity Recognition (NER)

There are a variety of methods for Natural Language Processing (NLP) that aims to extract information from the unstructured text.One of the well-known methods is Named Entity Recognition (NER) which is a process of information extraction that seeks to locate and classify named entities in text into pre-defined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc. Many techniques used to create various types of models, such as Rule Base, Machine Learning and Hybrid Method [8]. The language of each country is different for example in the English language, the sentence started with uppercase letters and the end of the sentence with the full-stop. Each word is divided with clear space, compared to the Thai language, which is not have a clear sentence breaks, no word breaks that caused difficulties in finding the boundaries of words. Some problems of the word ambiguity etc., However NER can help many areas such as extract a person mentioned in the twitter, specified products mentioned in complaints or reviews etc.

Even NER models were created using the same technique but the results can be different. In other words, the source of the message is affected to the model creation. Compared to either the information from Wikipedia or the official news website, it is found that the language usage is quite accurately than the news from social media. There are several techniques can be used to create the model for NER. However, this research tries to find an appropriate technique and compare those techniques in terms of precision recall and F-measure.

C. Conditional Random Field (CRF)

The CRF was presented by Lafferty McCallum and Pereira in 2003, in which the construction of a NER technique created by the CRF technique will use the method of determining the probability that there is a probability of the word being which type of specific name is the most? Which all types of unique names are associated with every word and also have a relationship with the previous specific name. Creating a NER model with CRF technique is one way to get an effective model. It is also used to create other models, such as POS [9]

A CRF is a Discriminative Probabilistic Classifiers. The difference between discriminative and generative models is that while discriminative models try to model conditional probability distribution, i.e., P(y|x), generative models try to model a joint probability distribution, i.e., P(x,y). Logistic Regression, SVM, CRF are Discriminative Classifiers. Naive Bayes, HMMs are Generative Classifiers. CRF’s can also be used for sequence labelling tasks like NER and POS Taggers. In CRFs, the input is a set of features (real numbers) derived from the input sequence using feature functions, the weights associated with the features (that are learned) and the previous label and the task is to predict the current label. The weights of different feature functions will be determined such that the likelihood of the labels in the training data will be maximized.

In CRF, a set of feature functions is defined to extract features for each word in a sentence. Some examples of feature functions are the first letter of the word capitalized, what the suffix and prefix of the word, what is the previous word, is it the first or the last word of the sentence, is it a number etc. These set of features are called state features. In CRF, we also pass the label of the previous word and the label of the current word to learn the weights. CRF will try to determine the weights of different feature functions that will maximize the likelihood of the labels in the training data. The feature function dependent on the label of the previous word is transition feature.

A conditional random field may be viewed as an undirected graphical model, or Markov random field, globally conditioned on X, the random variable representing observation sequences. Formally, define G = (V, E) to be an undirected graph such that there is a node v ∊ V corresponding to each of the random variables representing an element Y_v of Y. If each random variable Y_v obeys the Markov property with respect to G, then (Y, X) is a conditional random field. In theory the structure of graph G may be arbitrary, provided it represents the conditional independencies in the label sequences being modeled. However, when modeling sequences, the simplest and most common graph structure encountered is that in which the nodes corresponding to elements Y_v of Y form a simple first-order chain, as illustrated in Fig. 1.

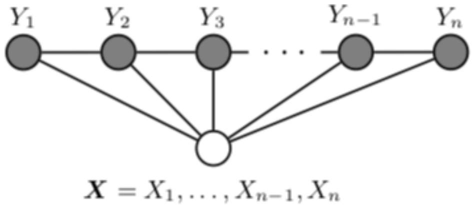

Figure 1. A chain-structured CRFs for sequences.

Fig. 1 is Graphical structure of a chain structured CRFs for sequences. The variables corresponding to unshaded nodes are not generated by the model.

D. Bidirectional -Long Short-Term Memory (Bi-LSTM)

Artificial neural networks have a variety of architectures. One architecture called a Recurrent Neural Networks (RNNs) is a class of artificial neural networks where connections between nodes form a directed graph along a temporal sequence.This allows it to exhibit temporal dynamic behavior derived from feed-forward neural networks.

The main problem of RNNs is the back-propagation process, which will calculate the gradient value to adjust the weight appropriately. When the gradient value used with RNNs with the length of the data sequence, it will have to start calculating since the first node. When reaching at the end of node, the gradient values are reduced to a minimum cause the efficiency are decreased. Since 1997, Hochreiter [10] presented LSTM to solve the problem as mentioned by allowing the status of each node. It will be calculated more efficiently. The LSTM is become a well-known and widely used in order to create a NER model. It is used many researches in year 2003. Huang, Xu and Yu [11] proposed a combination of a Bidirectional LSTM network and a CRF network to form a BI-LSTM-CRF network (Fig. 2) to train a model and test performance compared to LSTM model, Bidirectional LSTM model, CRF model. The experiments show that Bidirectional LSTM-CRF is given the best performance in terms of accuracy. However, Hochreiter and Schmidhuberalso used LSTM to create a NER model by testing with both English and German language and applied with other techniques. Hammerton [12] proposed Named entity recognition with long short-term memory while Chiu and Nichols used Bidirectional LSTM together with Convolutional Neural Network (CNN) for Named entity recognition [13].

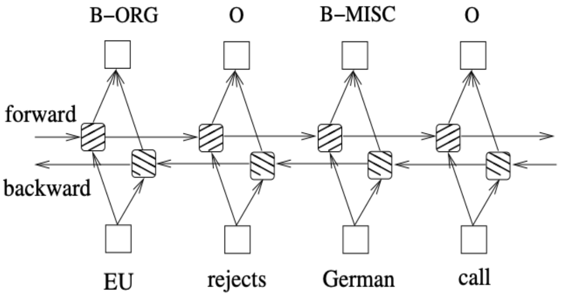

Figure 2. A Bi-LSTM-CRF model.

E. Social Media and Named Entity Recognition (NER)

Social media has become an integral part of people’s lives and daily routines. Using user-generated contents from social media is also an interesting topic for knowledge discovery in order to extract a useful information from online social media. Name Entity Recognition (NER) is one of the techniques that used to extract information from online social media [14]. Limsopatham and Collier are created a NER model by using Bidirectional LSTM technique [15] in many languages. Peng and Dredze are investigate better ways to corporate word boundary information into the NER model for Chinese social media [16] Sodanil and Lungkatoongare presented a NER model from social media in Thai language in order to support natural disaster management [17] event both Chinese and Thai language are difficult to implement the NLP. In [11], Bidirectional LSTM-CRF models for sequence tagging is applied which given such a good performance in English language.

This research focused on increasing the efficiency of the NER model using information from social media content that contained natural disasters. Both CRF and Bi-LSTM are used to test the performance of NER model based on the same data set that collected from online social media.

III. Research Methodology

This section described data collection, data preparation, classification model creation and model evaluation as shown in proposed framework (Fig. 3).

A. Data Collection

For data gathering. Data was collected from a public APIs of Twitter and Instagram that related to earthquake on April, 2015 and related to flood in between January 28, 2015 and September 14, 2015. The keywords used are in the group of disasters, including earthquake, fire, forest fire, flood, flash flood, summer storm, windstorm or hurricane and tsunami. The massages were separated from HTML using regular expression and stored in a pattern of sentences. The total words after extracted with uniques is equal 1,000 words.

B. Data Preparation

There are two main processes, tokenizer and insert label in order to prepare for the model creation. Data will transform into two patterns that suitable for each method:

CRF and Bi-LSTM which required data in different formats. For the first step, each word has to segment in order to prepare for using in the next part. PythaiNLP[1], a Python package for text processing and linguistic analysis is used to process text in the first step. It's similar to Natural Language Toolkit (NLTK) but PythaiNLP is focus on Thai language. The Newmm is selected to segment Thai words by using the maximum matching algorithm.

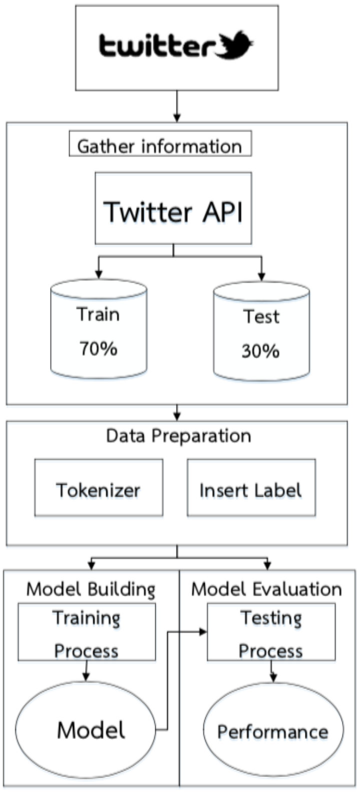

Figure 3. Proposed Framework for NER.

In the next step, the Part-of-Speech (POS) tagging is used for marking up a word in the corpus to a corresponding part-of-speech tagging. The POS Tagger technique is used together with a Thai POS database called "ORCHID" database [18]. After POS tagging has been completed, recognition model is started to create model. The output of NER is divided into six groups such as Type, Location, Organization, Person, Level and Other.

The data is divided into two sets, one CRF and one for the Bi-LSTM. In the Bi-LSTM dataset has to change the words into word embedding, which uses Word2vec[2] as a tool to convert data. However, both of datasets are divided into two parts for creating model and for testing model (training 85% and data testing 15%) and measure the model performance in terms of precision, recall and F-measure.

C. Classification Modeling

The CRF model process is created by using additional specifications. The sentence will be checked the previous word, and the back word, the word is in the beginning position of the sentence and word is in the end position of the sentence? The initial coefficient is c1 = 0.1 and c2 = 0.1. When the initial model has been completed, the coefficient will be randomly adjusted to obtain the most effective model. The Bi-LSTM-CRF is used to create the model by determine the number of cycles in the testing (Epochs) equal to 20 cycles. Finally, the data set that used for testing is separated from training data set.

D. Model Evaluation

Each method has to evaluate the performance in order to find the best solution for NER. The performance are measured using precision, recall and F-measurethat can be denotedas:

$$\mathrm{Pr}\,ecision(P) = \frac{\#\mathrm{Re}\,trieved\ \&\ \mathrm{Re}\,levant}{\#\mathrm{Re}\,trieved} \tag{1}$$

$$\mathrm{Re}\,call(R) = \frac{\#\mathrm{Re}\,trieved\ \&\ \mathrm{Re}\,levant}{\#\mathrm{Re}\,levent} \tag{2}$$

$$F - measure = \frac{2PR}{P+R} \tag{3}$$

IV. Experimental and Results

This section is illustrated the experimental and result of the proposed method. Also compare with other techniques using the same data set.

A. Experiments

Three main experiments were testedbased on the same data set which is devided into training and testingset for 85% and 15% respectively. There are six classes of output that represented input text: Type, Location, Organization, Person, Level and Other.

CRF and CRF with optimization are tested based on cross-validation in order to avoid overfitting.

B. Results

Table I and Fig. 4 shows the performance comparison for all testing model. It shown that the CRF+Optimize model is given the highest performance than other methods with score 98.94%, 98.95% and 98.93% of precision recall and F-measure respectively. Compared to previous the paper (authors presented in the 3rd GI Workshop on Complex Structures 2018, March 25th – 28th, 2018 in Boppard on the Rhine/Germany), all methods are given higher performance than the results from previous paper. It means that these methods are better to use for NER based on Thai twitter messages. However, the results shown that the performance between CRF and bi-LSTM is not significant difference.

TABLE I. Performance Evalutation for Each Method

	Precision	Recall	F-measure
CRF	0.9865	0.9866	0.9863
CRF+Optimize	**0.9894**	**0.9895**	**0.9893**
bi-LSTM+CRF	0.9820	0.9830	0.9820
Previous Paper	0.9058	0.8465	0.8798

[1] https://github.com/PyThaiNLP/pythainlp
[2] https://code.google.com/archive/p/word2vec/

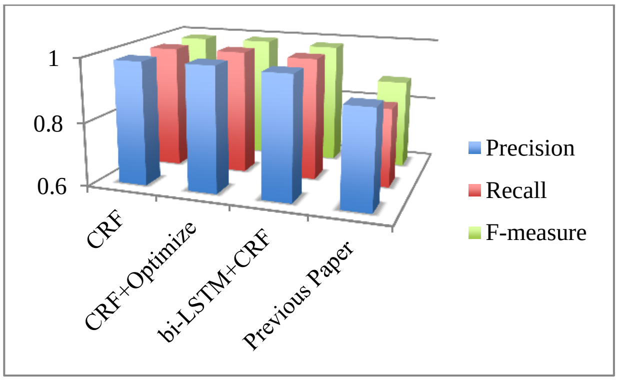

Figure 4. Performance evaluation graph.

Table II and Fig. 5 show the performance for each group of entity type. The total average score for all types is equal with 93.89%, 82.39% and 87.60% of precision recall and F-measure respectively. According to the results, CRF-optimize method is applied to use with text cloud that can be displayed in the form of text visualization as shown in Fig. 6 and Fig. 7 for a sample text cloud of disaster type of earthquake and flood respectively.

TABLE II. Performance for Each Group

Group	Precision	Recall	F-measure
Type	0.9941	0.9849	0.9895
Location	0.9384	0.8294	0.8805
Organization	0.8947	0.6623	0.7612
Person	0.9439	0.8211	0.8783
Level	1.0000	0.8462	0.9167
Other	0.8625	0.7995	0.8298
Total	**0.9389**	**0.8239**	**0.8760**

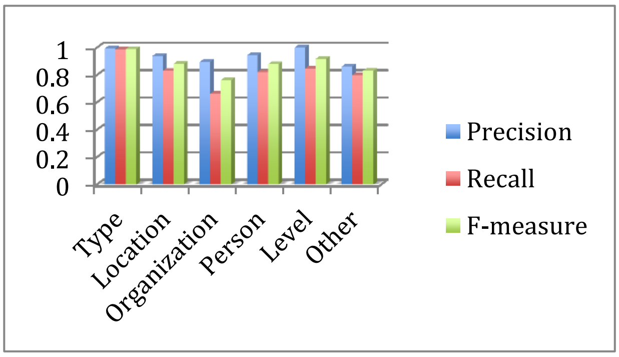

Figure 5. Performance evaluation graph for each group.

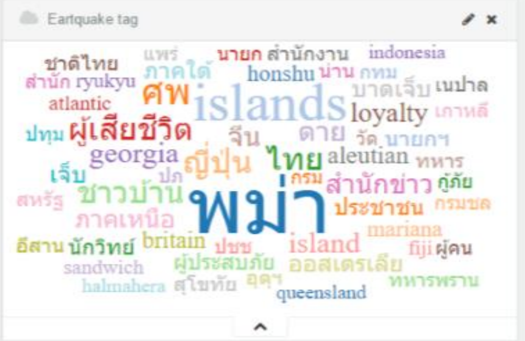

Figure 6. A sample of text cloud for earthquake.

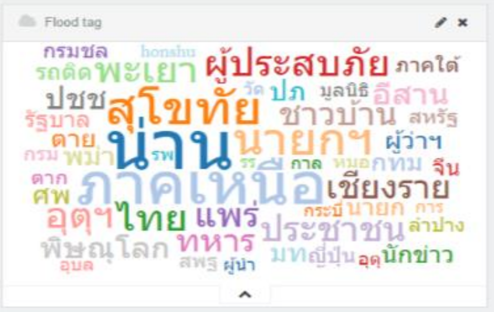

Figure 7. A sample of text cloud for flood.

V. Conclusions

In this paper, Named Entity Recognition (NER) is proposed using the resource from online social media content such as Twitter and Instagram, which is about natural disasters information. Two techniques are applied and combine for three scenarios: CRF, CRF+Optimize and Bi-LSTM CRF and testing with Thai language dataset. The results shown that CRF adjusted optimization is achieved a highest performance score 98.94%, 98.95% and 98.93%of precision recall and F-measure respectively. However, when compared to previous work, all methods are given better performance.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Hathairat Ketmaneechairat conducted the research; Maleerat Maliyaem analyzed the data; All authors wrote the paper and had approved the final version.

Acknowledgment

This research is supported by the College of Industrial Technology Faculty (CIT) of the King Mongkut's University of Technology North Bangkok (KMUTNB).

References

[1] A. Kongthon, C. Haruechaiyasak, J. Pailai, and S. Kongyoung, “The role of Twitter during a natural disaster: Case study of 2011 Thai Flood,” in *Proc. PICMET'12*: *Technology Management for Emerging Technologies*, Vancouver, BC, 2012, pp. 2227-2232.

[2] M. B. Habib and M. V. Keulen, “NEED4Tweet: A Twitterbot for tweets named entity extraction and disambiguation,” in *Proc. 53rd Annual Meeting of the Association for Computational Linguistics*, 2015, pp. 31-36.

[3] F. Liu, M. Vasardani, and T. Baldwin, “Automatic identification of locative expressions from social media text: A comparative analysis,” in *Proc. the 4th International Workshop on Location and the Web*, Shanghai, China, 2014, pp. 9-16.

[4] J. Lingad, S. Karimi, and J. Yin, “Location extraction from disaster-related microblogs,” in *Proc. the 22nd International Conference on World Wide Web*, Rio de Janeiro, Brazil, 2013, pp. 1017-1020.

[5] M. S. Y. Lungkatoong, “Classification of open source earthquake disaster information in Thai language,” in *Proc. 11th National Conference on Computing and Information Technology*, 2015.

[6] P. K. Prasetyo, D. Lo, P. Achananuparp, T. Yuan, and L. Ee-Peng, “Automatic classification of software related micro-blogs,” in *Proc. IEEE International Conference on Software Maintenance*, 2012, pp. 596-599.

[7] D. Küçük and R. Steinberger, "Experiments to improve named entity recognition on Turkish tweets," arXiv preprint arXiv:1410.8668, 2014.

[8] X. Liu, M. Zhou, F. Wei, Z. Fu, and X. Zhou, "Joint inference of named entity recognition and normalization for tweets," in *Proc. Meeting of the Association for Computational Linguistics: Long Papers*, 2012, pp. 526-535.

[9] H. M. Wallach, "Conditional random fields: An introduction," Technical Report (CIS), February 2004.

[10] S. Hochreiter and J. Schmidhuber, "LSTM can solve hard long time lag problems," in *Proc. the 9th International Conference on Neural Information Processing Systems*, 1996, pp. 473-479.

[11] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," arXiv preprint arXiv:1508.01991, 2015.

[12] J. Hammerton, "Named entity recognition with long short-term memory," in *Proc. Seventh Conference on Natural Language Learning at Hlt-naacl*, 2003, pp. 172-175.

[13] J. P. Chiu and E. Nichols, "Named entity recognition with bidirectional LSTM-CNNs," arXiv preprint arXiv:1511.08308, pp. 357-370, 2015.

[14] N. Peng and M. Dredze. (2017). Supplementary results for named entity recognition on Chinese social media with an updated dataset. Tech. Rep. [Online]. Available: http://www.cs.jhu.edu/~npeng/papers/golden horse supplement.pdf

[15] N. Limsopatham and N. H. Collier, "Bidirectional LSTM for named entity recognition in Twitter messages," in *Proc. the 2nd Workshop on Noisy User-generated Text*, 2016, pp. 145-152.

[16] N. Peng and M. Dredze, "Improving named entity recognition for chinese social media with word segmentation representation learning," arXiv preprint arXiv:1603.00786, pp. 149-155, 2016.

[17] M. S. Y. Lungkatoong, "NER detection from Thai tweets related to natural disaster," in *Proc. National Conference on Computing and Information Technology*, 2017.

[18] V. Sornlertlamvanich, N. Takahashi, and H. Isahara, "Thai part-of-speech tagged corpus: ORCHID," in *Proc. the Oriental COCOSDA Workshop*, 1998, pp. 131-138.

Hathairat Ketmaneechairat received PhD in Electrical Engineering with the thesis title "Smart Buffer Management for Different Start Video Broadcasting" from the King Mongkut's University of Technology North Bangkok, Thailand. Currently, she is a lecturer at the College of Industrial Technology, King Mongkut's University of Technology North Bangkok. Her Re$$search areas are Natural Language Processing and Data Mining, Machine Learning and Artificial Intelligence.

Maleerat Maliyaem received PhD in Information Technology (International Program) with the thesis title "Thai Speech-to-Text Translation of Spontaneous Speech" from the King Mongkut's University of Technology North Bangkok, Thailand. Currently, she is a lecturer at the Faculty of Information Technology, King Mongkut's University of Technology North Bangkok. Her research areas are Natural Language Processing and Information Retrieval, Machine Learning and Artificial Intelligence.