An Overview on the Development of Thai Natural Language Processing

Chalermpol Tapsai* Phayung Meesad** and Herwig Unger***

Received : November 30, 2019
Revised : December 24, 2019
Accepted : December 26, 2019

Abstract

At present, Natural Language is a leading technology that plays an important role and has a wide range of applications in human daily life. For the Thai language, studies and applications of Natural Language have been developed since 1980, starting with the research related to syllable segmentation and word segmentation, which are the basic elements of the language. After that, some studies in a higher level were conducted to apply the Natural Language in various fields. Unfortunately, according to the complexity of the Thai language, the improvement of Thai Natural Language Processing has been relatively slow over time. However, at present, the AI FOR THAI platform has been developed by NECTEC to be a digital ecosystem, used as a center for disseminating the results of research studies on Natural Language Processing and artificial intelligence. This platform offers many API services related to Natural Language Processing (NLP) to help developers to create applications for business, which may help driving Thai NLP to be success more effectively.

Keywords: Natural Language Processing, Development, Soundex, Trie.

1. Introduction

Natural Language Processing is a processing technique to make computers learn and understand human language, which allows users to interface and command computers easier by their language. Early studies on Natural Language Processing became to be known and interested since the 1960s. With many techniques, Natural Language Processing algorithms were developed and applied for many purposes, for example, language translation [1], text summarization for information [2], [3], [4], humans-computers interaction [5], and data retrieval from database [6], [7], etc. Though many NLP studies have been conducted, many problems that obstruct the progression of the NLP studies still occurred. Androutsopoulos [8] mentioned that essential factors in the development of NLP are the expertise in linguistics and specialization of research works. The lack of expertise would hinder the progression of NLP researches and developments. Rodolfo [9] mentioned four major problems of Natural Language to Interface Database, including various grammatical forms of Natural Language, missing some essential words, querying for information relates to many tables and using an aggregate function, and problems caused by human errors.

2. Processes in Natural Language Processing

Natural Language Processing can be divided into two types, including grammar-based methods and non-grammar-based methods. With non-grammar-based methods, instead of using linguistic principles, other techniques, such as statistical analysis or Data mining, will be applied to analyze words or texts to obtain the expected results. For grammar-based methods, grammar rules and linguistic principles are used to analyze the meaning of Natural Language sentences and produce outputs related to the meaning. The process of grammar-based methods consists of 4 work steps, which are Lexical Analysis, Syntactic Analysis, Semantic Analysis, and Output Transformation [10].

1) Lexical Analysis: This step segment Natural Language sentences into small items called Token.

2) Syntactic Analysis: This step parsed all tokens with predefined sentence syntax to verify the sentence pattern and

* *Department of Information Technology Management, Faculty of Information Technology, King Mongkut's University of Technology North Bangkok..*
** *Department of Information Technology, Faculty of Information Technology, King Mongkut's University of Technology North Bangkok.*
*** *Chair of Communication Networks, Faculty of Mathematics and Computer Science, FernUniversitat in Hagen.*

provided some information used in the next step.

3) Semantic Analysis: This step analyzed all tokens from step 2 by semantic structures, such as an ontology or a semantic web structure, to infer the meaning of a sentence.

4) Output Transformation Process: This step transforms outputs derived from Semantic Analysis into the results that meet the user requirements, such as SQL commands for information retrieval from databases.

3. Thai Word Segmentation

There are many Natural Languages that are Non-segmentation language. These languages, for example, Thai, Laos, Japanese, Chinese, etc., written all words in consecutive without any spaces or delimiters, which make Lexical Analysis to be a non-trivial task, and need effective algorithm to define the precise word boundaries. For the Thai language, the early studies on NLP related to syllable segmentation started in 1981. The first research [11] is a rule-based algorithm, which divided Thai characters into five groups, including consonant, vowel, tonal mark, numeral, and special character. This algorithm parsed inputted text from the rightmost character to the left with the rules to define the boundaries of each syllable. In the case that the syllable cannot be defined according to the rules, the undefined characters will be stored in the data file as an exception. This algorithm works well with high accuracy of 85%. However, all rules were written within the source code of the program, which causes difficulty in editing and adding rules for better performance.

The other research [12] was conducted to improve the precision of segmented outputs with new rules designed according to Thai spelling principles, divided into two groups: the rules used to define the front boundary and the rules used to define the end boundary. Besides, some rules for other syllables forms segmentation, such as acronyms, foreign language syllables, special characters, and numbers also applied in this algorithm. The last syllable segmentation algorithm [13] is a dictionary-based, which parsed the inputted text from the leftmost character to the right with the syllables stored in the dictionary to define each syllable boundaries. By using the "longest matching technique," this algorithm prefers the output with the longest length. In the case of cannot detect the beginning of the next syllable, "Backtracking" technique is used to step back for a shorter syllable and start parsing at a new position for a better result. In the case that no syllable in the dictionary matched to letters of the inputted text, these letters will be processed by eight grammar rules to find suitable boundaries.

The first algorithm for Thai word segmentation research [14] is performed by a mixed-method algorithm of Rules-based and Dictionary-based method. In the first step, the inputted text is parsed with 43 grammar rules to define boundaries of the longest token without considering the final consonant (except in some particular cases). Then, all tokens are grouped and parsed with words in the dictionary, which implemented by a relational database to output the correct words.

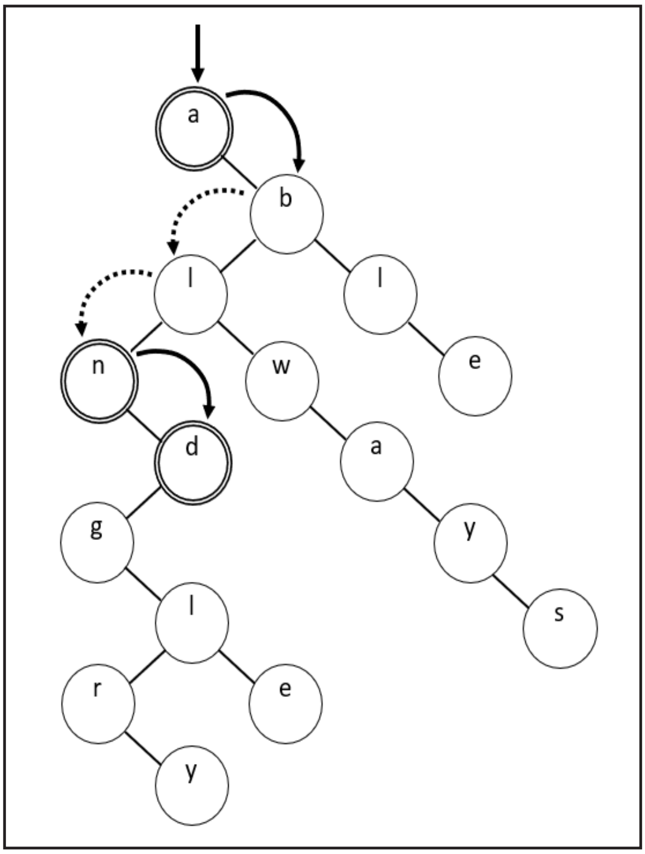

Figure 1. Trie parsing for the word "and".

The second research [15] focuses on increasing the efficiency of word segmentation by using "Trie" [16], [17] that is a tree-like data structure in which each node stores just a letter rather than a word, to reduce the size of the dictionary as well as the number of comparisons. The example of a Trie structure, which stores five words: able, always, and angle, angry, shown in Figure 1.

Parsing a text with Trie, each letter of the text must be compared with Trie structure starting from the root node to find the matched characters and skip to next node (if matched: skip to the right node, mismatched: skip to the left node) until the ending of the expected word founded or reach the end of the text. The average time for parsing text with Trie structure is related to the number of letters in the longest word stored in the dictionary. While storing the dictionary with other tree structure such as Binary Tree or B-Tree cost the average parsing time related to a total number of words in the dictionary, which take a longer time. For this reason, the Trie structure is popular and wildly used in most of word segmentation algorithm.

Another technique for Thai word segmentation named "Maximal Matching" [18] was proposed to correct the error of the Longest Matching technique, which sometimes chooses a too long word that includes the initial consonant of the next word. The process of this algorithm firstly segments the inputted text into all possible words, then analyzed all outputs for the best result that provided the least number of words.

In 1997, a new word segmentation algorithm [19] using the part of speech (POS) to analyze the boundaries of words was conducted to solve the problem of ambiguous words and misspelling words. The work process divided into two steps: In the first step, the inputted text is segmented and tagged each word with all possible POS types, then calculate the probability of each output with the Markov model and choose the output with the highest probability as a result. The second step is a process for the misspelling correction based on the statistic values obtained from the corpus analysis that were segmented and tagged by humans (hand-tagged).

Another Thai word segmentation with a learning algorithm was presented to solve the problem of ambiguous words [20]. By using two features, including characteristics of context words and the collocation of words, derived from a text corpus, the learning process with two methods, Ripper and Winnow to identify the unknown words. In the case of unknown words are found, the algorithm will create new words by some rules that are set by the researcher to find the correct word in the dictionary. These new words may be made up of only the unknown word or combine with the surrounding words.

Currently, the most popular Thai word segmentation program named "LexTo" developed and distributed by The National Electronics and Computer Technology Center (NECTEC) [21]. LexTo is a dictionary-based method using a Trie structure with Longest Matching and backtracking technique. The dictionary used by LexTo is Lexitron Dictionary [22] that contains 42,222 words.

4. Comparison of Thai Segmentation Algorithms

The comparison study for the effectiveness of Thai word segmentation algorithms, which used different methods, was conducted in 2008 [23]. This research compared a Dictionary-based method (DCB) that used the Trie structure to four Learning-based methods (LB) including, Naive Bayes (NB), Decision Tree (DT), Support Vector Machine (SVM) and Conditional Random Field (CRF). The results showed that the Dictionary-based method and Conditional Random Field method provide accurate results at the level of high with the best Precision and Recall.

5. A High-performance Thai Word Segmentation

In 2019, A high-performance Thai word segmentation algorithm named "TLS-ART-MC" [24] was presented to cover four crucial problems of Thai segmentation, including the segmentation efficiency, misspelling words, multiple spelling patterns of names and foreign language vocabulary, and Compound words. For the segmentation efficiency improvement, the Automatic Ranking Trie that used the actual Word Usage Frequency (WUF) to reorganized the Trie structure to located the words with higher frequency at the beginning of Trie structure, which help reduces the number of comparison tasks significantly. Moreover, WUF is also used to reduce the number of words in the dictionary used to create Trie as well. For misspelling correction and multiple spelling patterns problems, a new Soundex technique named "Completed Soundex," which solved the problem in Traditional Soundex [25-30] and improved for better similarity analysis implemented in this algorithm. In the case of Thai Compound words, which are the words created from more than one base-word in many characteristics and purposes. According to the fact that several compound words in Thai are too many to be stored in the dictionary. Therefore, many errors of inconsistency result always occurred in all previous Thai word segmentation algorithm.

In order to solve this problem, the segmentation algorithm divided into two steps. In the first step, the inputted text will be segmented into base-word with word type. Then, in step 2, all base-words are analyzed and composed to be correct compound words based on the rules of Thai grammar.

6. Other Thai NLP Studies

Some studies on Thai NLP other than word segmentation presented in various purpose, for example, A Semantic Model for analyzing Natural Language Arithmetic Sentences to generate results instead of using mathematical formulas or functions [31], speech translation [32], Information retrieval using Natural Language [33], searching data using Soundex [30], Information Processing and Retrieval from CSV File by Natural Language [34], etc.

7. AI FOR THAI

In 2019, The National Electronics and Computer Technology Center (NECTEC) developed the AI FOR THAI platform to become a Digital Infrastructure for Thai people. This platform, include with NECTEC Artificial Intelligence (AI) studies, is provided in the form of API services for users and developers to be able to create and develop applications to benefit both business and society. As well as being an important mechanism for driving the AI Ecosystem in Thailand by allowing researchers, teachers, students, and developers to work together to create more new APIs in the future. NECTEC launches "AI FOR THAI" at the NECTEC-ACE 2019 Annual Conference on 9 September 2019. The AI FOR THAI consists of 11 services, including Basic NLP, Tag Suggestion, Machine Translation, Sentiment Analysis, Character Recognition, Object Recognition, Face Analytics, Persons & Activity Analytics, Speech to Text, Text to speech, and Chatbot. General users can use these services without registration at https://aiforthai.in.th/index.php#home.

7.1 Basic NLP

This service is a Thai text processing service. which consists of eight services as follows:

7.1.1 Word Segmentation: There are two-word segmentation services, including Tlex Plus and LexTo Plus. Tlex Plus offers word processing using machine learning techniques using the Conditional Random Fields algorithm. LexTo Plus offers a Dictionary-based word processing using the Longest Matching algorithm.

7.1.2 POS tagging: This service offers a word processing with Part of speech (POS) tagging

7.1.3 Named Entity Recognition: This service offers the Named Entity Recognition with POS using the Conditional Random Fields algorithm.

7.1.4 Grapheme to Phoneme: This service offers Thai text to Phoneme conversion.

7.1.5 Soundex: This service is searching for similar names or words from two data sources, which are a list of popular Thai names and the royal dictionary, using Soundex techniques

7.1.6 Word Approximation: This service is searching for the nearby spelling names or words from two data sources, which are a list of popular Thai names and the royal dictionary.

7.1.7 Word Similarity: This service is searching for similar contexts from two data sources, which are Thai Wikipedia and Twitter using Word2Vec algorithm

7.1.8 Text Cleansing: This service is editing for Thai text that is a misspelling in character misplaced typing.

7.2 Tag Suggestion

This service offers analyzation of the inputted text for some keywords related to the inputted text meaning.

7.3 Machine Translation

This service offers the translation from Thai to Chinese and vice versa.

7.4 Sentiment Analysis

This service offers the sentiment analysis of the inputted sentences to evaluate the Sentiment and Intend.

Emoji Prediction offers the sentiment analysis of Thai messages for felling evaluation and output as the Emoji icons.

7.5 Character Recognition

This service offers consists of two sub-services.

7.5.1 OCR: The Optical Character Reader for converting an image of a text document into text. This service supports image files of .jpg and .png formats, which must be greater than 200 dpi resolution, Or A4 size files with a file size less than 1 MB.

7.5.2 LPR: The Label Plate Recognition for converting an image of Label plate into characters and numbers, which appeared in the image. This service supports image files of .jpg formats, which file size must be less than 1 MB.

7.6 Object Recognition

This service offers the analysis and recognition for objects from images using Deep Learning techniques. For example, identify Thai food name from pictures.

7.7 Face Analytics

The service for Face Analytics is still under development and will be ready for service in 2020.

7.8 Persons and Activity Analytics

This service analyzed an image to identify and located each person in that image.

Motion Heatmap analyzed and evaluated the density of moving peoples in a specific area.

7.9 Speech to Text

This service, named "Partii," converts an inputted audio file into text. The audio file must be.wav extension recorded in mono format, 16bit resolution, at 16kHz frequency, a maximum of 30 seconds in length, and file size of no more than 1MB.

7.10 Text to speech

This service offers the conversion from text to speech.

7.11 Chatbot

This service, named "ABDUL (Artificial chatBot, which Does Understand Language)," is a platform that offers a service to create chatbots for an automated chat.

8. AI FOR THAI: Corpus

This section of AI FOR THAI service offers Thai corpus for developers who register tobe the AI FOR THAI members. There are eight corpuses available for download with the following details:

8.1 BEST

This corpus, named "Benchmark for Enhancing the Standard of Thai Language Processing (BEST)," is a collection of Thai text with word boundary tagging used in important software competitions related to Thai language processing. More details available at http://thailang.nectec.or.th/best.

8.2 Lexitron 2.0

This corpus is a collection of Thai/English vocabulary, with part of speech (word type), translation meaning, synonyms, and usage examples. With a 6 MB text file, this corpus consists of 53,000 Thai -> English vocabularies, and 83,000 English -> Thai vocabularies. More details available at http://lexitron.nectec.or.th

8.3 Lotus

This corpus is an extensive Thai speech database developed for use in research and development of a continuous speech recognition system (Large Vocabulary Continuous Speech Recognition: LVCSR). The database consists of a set of balanced phonemes. (Phonetically Balanced Set) from a database of news articles or general articles. Speech in the LOTUS database recorded via two types of microphones: high-quality Close-talk microphones and Unidirectional microphones, with medium quality of the recording in 2 environments: a quiet room environment and an office environment. The LOTUS database contains speech data and transcripts from 24 speakers.

8.4 Thai OCR

This corpus is an image database for developing Thai Optical Character Recognition programs, divided into two sets:

1) Training set consists of BMP type images of 162 Thai characters in which each character has 5,000 styles.

2) Test set is JPG image files of Thai documents, which divided into two categories, including books and academic journals, 100 pages per category.

8.5 Thai Plagiarism

This corpus is a data set of creating text that has been copied by humans used as a suspicious document for Plagiarism evaluation. With a total of 1,050 files, the suspicious documents divided into four types of copied styles, which are copy-based change, lexicon-based change, structure-based change, and semantic-based change. The Source documents in this corpus are Thai Wikipedia articles and webpages. More details available at http://copycatch.in.th/thai-plagiarism-task.html

8.6 Thai QA

This corpus consists of a group of question-and-answer pairs created by general users with various contents from Thai Wikipedia, such as science, tourism, sports, and more. With a total of 4000 pair of Q/A mixed with both simple and difficult questions, these questions covered most used question words, such as ใคร (who), อะไร (what), ที่ไหน (where), เมื่อไร (when), ใด (which), เท่าใด (how many), เท่าไร (how much), etc. More details available at http://copycatch.in.th/thai-qa-task.html

8.7 TSynC2_Nun

This corpus, named "Thai Speech Database for Speech Synthesis Version 2", is created for use as an automated database for creating acoustic models, and also used as a basic sound unit for Thai speech synthesis programs. Database properties have details as follows:

1) The text data section is 2710 Thai sentences from news messages in which each sentence has already segmented into words with phonetic transcription.

2) The speech data section is female sounds recorded in 44.1 kHz at 16 bits/sample, approximate SNR 32.68 dB with 5 hours 25 minutes duration.

8.8 THAI-NEST

This corpus is a collection of news with tagging on name entities, including person's name, location, organization name, group, date, time, and quantity specification words.

9. Developer

This section provides many API services, including Soundex, and...., for developers who register as a member. Each API service has a usage description with example source codes, which can be used as a guideline for programming in many languages, such as PHP, JAVA, Python.

10. Conclusion

From the content presented in the previous section, even though the Thai Natural Language Processing has been researched and developed continuously for a long time, many problems still occurred due to the fact that the Thai language is very complex in the usage of words and sentences. These problems obstruct the development of Thai Natural Language Processing from successful in higher-level research. However, at present, NECTEC has developed a platform, named AI FOR THAI, which aims to be a center for the dissemination of NLP and AI research for users and developers to create applications of NLP for real usage in business. This platform is a new hope for big success in Thai NLP.

11. References

[1] O. Kazar, S. Hamza, B. Hind, and L. S. Bourekkache, "Semantic natural language translation based on ontologies combination." *Proceedings of The 8th International Conference on Information Technology ICIT (The Amman*, Jordan, 17-18 May, 2017.

[2] I. G. Harris. "Extracting design information from natural language specifications." *Proceedings of the Design Automation Conference (DAC), 2012 49th ACM/EDAC/ IEEE*, pp. 1252-1253, June, 2012.

[3] O. M. Foong, S. P. Yong, and F. A. Jaid. "Text Summarization Using Latent Semantic Analysis Model in Mobile Android Platform." *Proceedings of The 2015 9th Asia Modelling Symposium (AMS),* pp. 35-39, September, 2015.

[4] A. Fatwanto. "Software requirements specification analysis using natural language processing technique." *International Conference on QiR*, pp. 105-110, 2013.

[5] H. J. Hyeon, J. Taylor, and E. T. Matson. "Natural Multi-Language Interaction between Firefighters and Fire Fighting Robots." *In Web Intelligence and Intelligent Agent Technologies, IEEE/WIC/ACM International Joint Conferences* DOI=10.1109/ WI-IAT.2014.166., pp. 183-89, 11-14 August, 2014.

[6] R. Agrawal and et al. "DBIQS — An intelligent system for querying and mining databases using NLP." *International Conference on Information Systems and Computer Networks (ISCON)*, 2014.

[7] R. Priyadarshini, L. Tamilselvan, T. Khuthbudin, S. Saravanan, and S. Satish. "Semantic Retrieval of Relevant Sources for Large Scale Virtual Documents." *Procedia Computer Science*, Vol. 54, pp. 371-379, 2015.

[8] G. Androutsopoulos, D. Ritchie, and P. Thanisch. Natural Language Interfaces to Databases-An Introduction." Natural Language Engineering, Vol. 1, No. 1, pp. 29–81, 1995.

[9] A. Rodolfo, R Pazos, J. Juan, B. Gonzalez, A. Marco, and L. Aguirre. "Semantic Model for Improving the Performance of Natural Language Interfaces to Databases." *In: Advances in Artificial Intelligence: 10th Mexican International Conference on Artificial Intelligence, MICAI 2011*, Vol. 7094, pp.277-290, 2011.

[10] A.Shah, J. Pareek, H. Patel, and N.Panchal. "NLKBIDB-Natural Language and keyword based interface to database." *In: IEEE International Conference on Advances in Computing, Communications and Informatics (ICACCI 2013)*, pp. 1569–1576, 2013.

[11] Y. Thairatananond. *Towards the Design of a Thai Text Syllable Analyzer*. Master Thesis of Science, Asian Institute of Technology, Pathumthani, 1981.

[12] S. Chanyapornpong. *A Thai Syllable Separation Algorithm.* Master Thesis of Engineering, Asian Institute of Technology, Pathumthani, 1983.

[13] Y. Poovarawan and W. Imarrom. "Thai Syllable Separator by Dictionary." *In: 9th Annual Meeting on Electrical Engineering of the Thai Universities*. Khonkaen, 1986.

[14] D.Sawamipuk. *Development of Thai grammar analysis software under UNIX system*. Bangkok, Thammasat University Press., 1990.

[15] S. Raruenrom. *Dictionary-based Thai Word Separation. Senior Project Report*, Department of Computer Engineering, Chulalongkorn University, Bangkok, 1991.

[16] P. Smith. *Applied Data Structures with C++*. Jones and Bartlett publisher. Massachusetts, USA, pp. 253-273, 2004.

[17] H. T. Cormen and et al. *Introduction to Algorithms. 2nd Edition*. USA : The MIT Press, pp. 434-493, 2002.

[18] V. Sornlertlamvanich. "Word Segmentation for Thai in Machine Translation System." *In: Machine Translation, National Electronics and Computer Technology Center*. Bangkok, pp. 50–56, 1993.

[19] A. Kawtrakul, C. Thumkanon, and S. Seriburi. "A Statistical Approach to Thai Word Filtering." *Proceedings of the 2nd Symposium on Natural Language Processing*, Bangkok, pp. 398-406, 1997.

[20] P. Chaloenpomsawat. *Feature-Based Thai Word Segmentation,Chulalongkorn University*, Master Thesis of Chulalongkorn University, 1998.

[21] National Electronics and Computer Technology Center. *Thai Lexeme Tokenizer : LexTo*. Available Online at http://www.sansarn.com/lexto/ accessed on 1 December 2019

[22] National Electronics and Computer Technology Center. *Thai Lexeme Tokenizer : Lexitron dictionary*. Available Online at http://www.sansarn.com/lexto/ accessed on 1 December 2019.

[23] C. Haruechaiyasak, S. Kongyoung, and M. Dailey. "A Comparative Study on Thai Word Segmentation Approaches." *IEEE Proceedings of ECTI-CON 2008. 5th International Conference,* pp. 125–128, 2008.

[24] C. Tapsai, P. Meesad, and C. Haruechaiyasak."Thai Language Segmentation by Automatic Ranking Trie with Misspelling Correction." *Proceedings of The Autonomous Systems 2019*, Cala Millor, Spain, pp. 121-134, October, 2019.

[25] National Archives and Records Administration. *The Soundex Indexing System.* Available Online at https://www.archives.gov/ research/census/soundex. html. accessed on 1 December 2019.

[26] T. Karoonboonyanan and et al. "A Thai Soundex System for Spelling Correction." *Proceedings of NLPRS*, pp. 633-644, 1997.

[27] Datamat Co., Ltd. "Local tax collection monitoring system." *Report to the Office of Policy and Planning*, Bangkok, 1978.

[28] N. Angkawattanawit, C. Haruechaiyasak, and S. Marukatat. "Thai Q-Cor: Integrating Word Approximation and Soundex for Thai Query Correction." *Proceedings of International Conference on Electrical Engineering/Electronics*, *Computer, Telecommunications and Information Technology*, Krabi, Thailand, pp.121-124, 2008.

[29] V. Lorchirachoonkul "A Thai soundex system." *Information Processing and Management*, Vol. 18, No. 5, pp. 243-255, 1982.

[30] W. Udompanich. *String searching for Thai alphabet using Soundex compression technique*. Master Thesis of Department of Computer Engineering Graduate School, Chulalongkorn University, 1983.

[31] C. Tapsai, P. Meesad, and C. Haruechaiyasak. "Natural Language Semantic Model for Arithmetic Sentences." *The 3rd International Conference on Communication and Information Processing (ICCIP 2017)*, Tokyo, Japan, November, 2017.

[32] P. Charoenpornsawat, and T. Schultz. "Improving word segmentation for Thai speech translation." *IEEE Spoken Language Technology Workshop*, 2008.

[33] C. Tapsai. "Searching Model on Learning Document by Soundex and Semantic Analysis." *International Academic Conference on Economy and Management Innovations*, *AC-EMI-2018*, Budapest, Hungary, March, 2018.

[34] C. Tapsai. "Information Processing and Retrieval from CSV File by Natural Language." *IEEE 3rd International Conference on Communication and Information Systems (ICCIS),* 2018.

[35] National Electronics and Computer Technology Center. Available Online at https://aiforthai.in.th/about.php. accessed on 1 December 2019.