SOLVING UNKNOWN WORD PROBLEMS IN NATURAL LANGUAGE PROCESSING

Chalermpol Tapsai[*]& Wilailuk Rakbumrung[**]

[, **] College of Innovation and Management, Suan Sunandha Rajabhat University, Thailand*
Email: [*] chalermpol.ta@ssru.ac.th, [**]wilailuk.ra@ssru.ac.th

ABSTRACT

Unknown words are a major problem that makes the Natural Language Processing (NLP) impossible to correctly analyze the meaning of the sentence. This research aim is to provide a model that will allow the NLP to correctly diagnose unknown words and replaced by the correct words. To complete this, the researcher firstly analyzes the characteristics of the unknown words that are not recognized by the NLP model. By collecting 12,800 text files of messages from various sources including both online and offline, cover all levels of language, formal, semi-formal, and non-formal. These text files were analyzed for the characteristics of the unknown words and classified into 7 types: Excess of alphabets, Missing of alphabets, Repetition of alphabets, Typo error, Misplacement of alphabets, Slang words and Mixed type error. To overcome the unknown words' problem. Complete Soundex is used to correct the misspelled words, the diversity of spelling, slangs and modification of traditional words by analyzing the unknown words to provide the correct words with the highest similarity. Evaluation of the model is performed by inputting the test dataset, which is natural language sentences collected from a sample group of 125 people with a total of 3,750 sentences, into the model to detect the unknown words and analyze to provide the correct words that can be used to replace the unknown word. Then collect all outputs and calculate the precision, recall and F1-score value. The result showed that the performance of the model was very good. The precision, recall and F1-score value are all greater than 90% and the unknown words that cannot be corrected by the model are 6.88% of the overall unknown words. There are 2 reasons that makes this model unable to solve these unknown words: 1) too much misspelling position in the unknown word and 2) the Word segmentation module cannot specify the boundaries of the unknown words correctly. In order to solve this problem, there should be more research to improve the process of analysis of unknown words' boundaries and using of co-occurrence word analysis which will help improve the model's efficiency.

Keyword: unknown word, natural language processing, complete Soundex, Thai language.

INTRODUCTION

Unknown word is a common thing that we often find in various types of documents that are used in everyday life, especially in informal documents and semi-formal documents. In general, although most of these words can be analyzed and understood by humans, contrary, for Natural Language Processing, unknown words are a major problem that causes serious errors in all processes and produce the wrong outputs for users. In order to make understand in the characteristics of these unknown words which can be used as the guidelines for solving this problem, we collected messages from various sources including both online and offline, cover all levels of language, formal, semi-formal, and non-formal. Then save these messages as text files with a total of 12,800. These text files were used as input to analyze for the characteristics of the unknown words which is finally classified into 7 types as follows: Excess of alphabets, missing of alphabets, Repetition of alphabets, Typo error, Misplacement of alphabets, Slang words and mixed type error [1]. Moreover, there are also some studies that discuss about causes of misspelling problem [2].

In the past period, there have been a lot of research studies that present various methods of unknown words handling and correcting typos' error. For example [3], [4], [5]. These solutions can work well but not cover for Thai language. For this reason, we are interested in conducting the research on word correction of Thai language. In this research, we propose a new model for unknown word analysis and correction. By using Soundex Similarity analysis, this model will analyze the input sentences for the unknown word, then encode this unknown word into a soundex code which will be used to compare to the Soundex code of the words in the dictionary to find the most similar word that will be used to replace the unknown word. The Soundex code used in this research is the Completed Soundex [6], a new encoding technique that was developed using all of the phonetic components of the word for encoding in order to get a Soundex code that is accurate to the actual pronunciation. This technique can be used to define the value of similarity between 0 - 1 while the traditional Soundex will result in just only two results, similar (1) and not similar (0). The remaining content in this article is divided into 6 topics: 1) Completed Soundex and similarity parsing, 2) The Unknown word Correction by Completed Soundex (UCCS) model, 3) Research steps, 4) Evaluation of the model, 5) Results and 6) Conclusion with the following details.

COMPLETE SOUNDEX

As mentioned above, the Complete Soundex (CS) is a code created from every phonetic component of a word. This code is divided into subtypes called "syllable codes". Each syllable codes is a 7-letter code which is encoding from initial consonant, vowels, tones, final consonant, and a clustering alphabet. As shown in Figure 1.

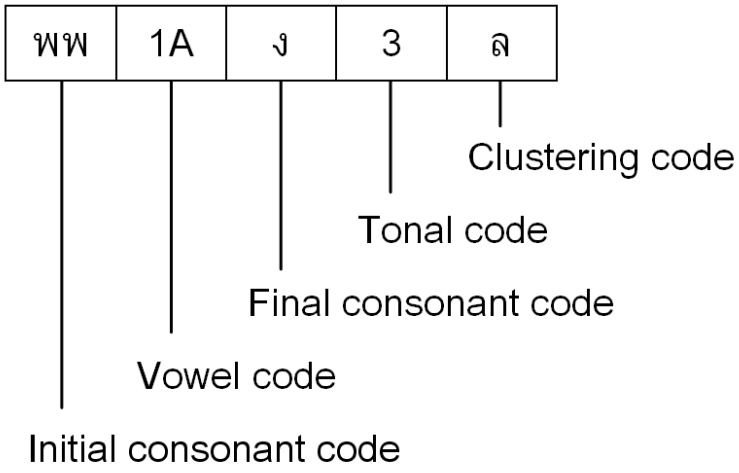

Figure 1. The example of a Syllable code for example Thai language word "พลั้ง".

The comparison of two CS to find the similarity value is a syllable-to-syllable comparison as shown in Table 1, by comparing each letter at the same position, the result will be 1 if both letters are the same, and result 0 in another case. Then, all results are summed and divided by the total number of 1 and 0 digit to obtain the similarity value.

Table 1. The example of a Syllable code for Thai language

Words	Initial consonant code	Vowel code	Final consonant code	Tonal code	Clustering code	Similarity value
พลั้ง	พพ	1A	ง	3	ล	-
พลาง	พพ	1B	ง	0	ล	-
Similarity	11	10	1	0	1	5/7 = 0.71

UNKNOWN WORD CORRECTION BY COMPLETED SOUNDEX MODEL

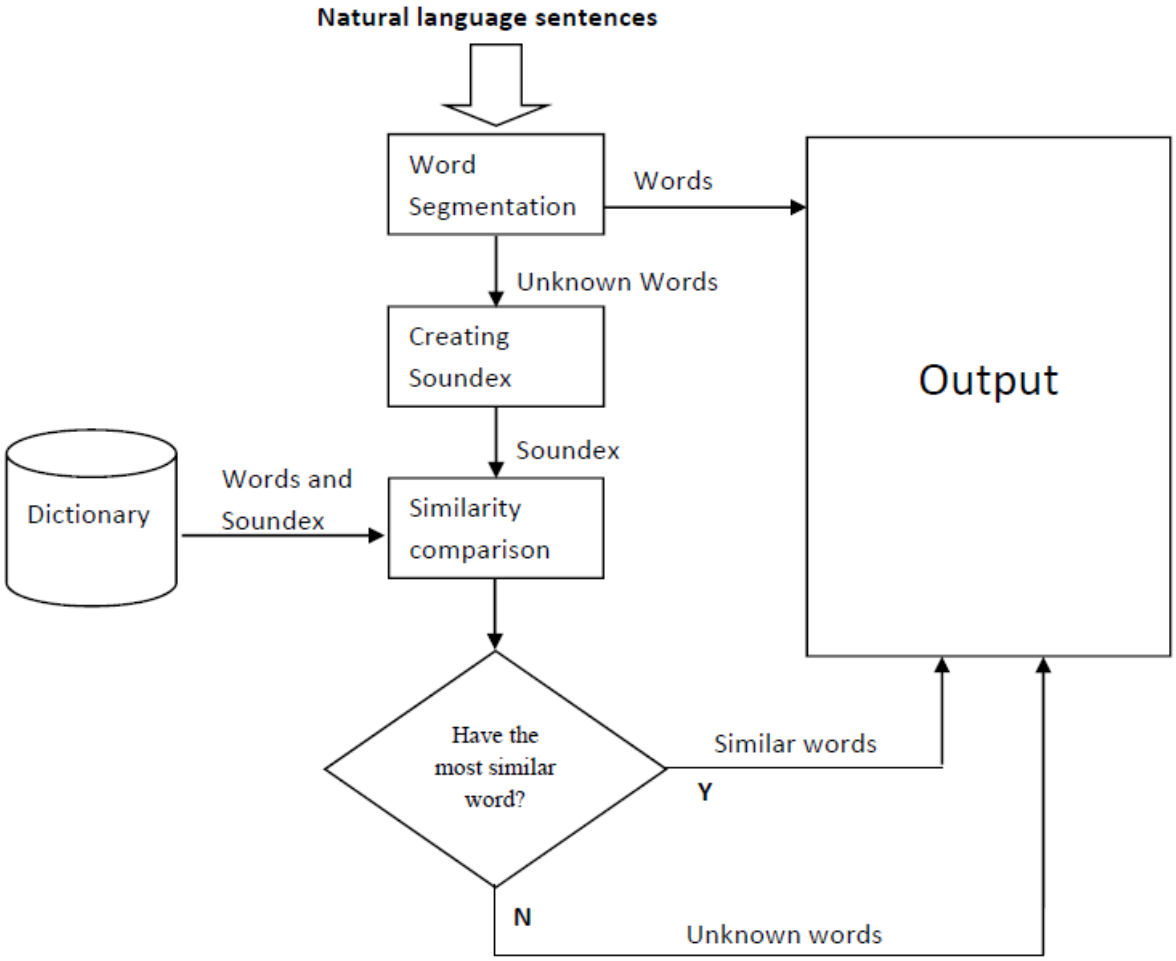

Figure 2. The Unknown word Correction by Completed Soundex (UCCS) model diagram

As shown in Figure 2, the UCCS model can be divided into 2 parts. The first part, Word segmentation, is the module that analyzed natural language sentences and separates into words. By using TLS-ART [7], a dictionary-based word segmentation program with Ranking Trie structure, each character in sentences is parsed to Ranking Trie to define word's boundaries by the longest matching with the backtracking technique. In the case of an unknown word was detected, TLS-ART will send that unknown word into the second part, correcting the unknown word, which consists of two important steps: Creating Soundex, and Similarity comparison, to determine the word in the dictionary with the most similarity value to the unknown word. The third part is the Similarity Comparison. This process begins by parsing the Soundex code of the unknown words to the Soundex code of words in the dictionary to find the most similar word which will be used to replace the unknown word. In the case that there is more than one word with the most similarity, the backtracking technique will be used by ignoring one-by-one phonetic component starting from the clustering character and then parse both Soundex codes again to define the most similarity word. If there are still words that have more than 1 word with the most similarity, the next phonetic components: tone and vowel will be ignored respectively until only one word can be found with the most similarity. In the case of there still have more than one word with the most similarity, random methods are used to select a word to be used as a replacement for the unknown word. On the other hand, if there are no words at all that have the similarity value more than 0.75 (the minimum threshold for this research), the model will report this unknown word as the output.

EVALUATION OF THE MODEL

In order to evaluate the performance of the model. We collected natural language sentences in which each sentence must include at least 1 misspelled word from a sample group of 125 people, 30 sentences for each people, with a total of 3,750 sentences. These sentences are used as input to test the performance of the model. Then, all results from this experiment were processed to calculate the Precision, Recall, and F1-Score.

RESULTS

From the experiment, The number of unknown words that can be detected, the number of unknown words that are replaced by correct words (True Positive), the number of unknown words that are replaced by wrong words (False Positive), and the number of correct words that are not selected (False Negative) are shown as Table 2.

Table 2. The experimental result of model evaluation

Number of unknown words detection	True Positive	False Positive	False Negative	Number of unknown words
4,272	3,783	195	356	294

The Precision, Recall, and F1-Score values are 95.1, 91.4 and 93.2 respectively.

CONCLUSION AND DISCUSSION

As shown in Table 2, although the performance of the UCCS model was very good. The Precision, Recall, and F1-Score values are all greater than 90%. However, the unknown words that cannot be corrected by the model are 294 or 6.88% of all unknown words that are detected. There are two reasons that make this model unable to solve these unknown words: 1) too much misspelling position in the unknown word and 2) the Word segmentation module cannot specify the boundaries of the unknown words correctly. In order to solve this problem, there should be more research to improve the process of analysis of unknown words' boundaries and using co-occurrence word analysis which will help improve the UCCS model's efficiency.

ACKNOWLEDGEMENTS

The author would like to thank the Research and Development Institute, Suan Sunandha Rajabhat University, Bangkok, Thailand for financial support.

REFERENCES

[1] Tapsai, C. (2018). Analysis of Patterns and Causes of Misspelling and Slang Words for Natural Language Processing, *Proceedings of the 2018 International Conference on Science, Technology and Management*, Moscow, Russian Federation, pp. 25-30

[2] Bunnatham, P. (2555). Chinese student and error in Thai writing. Research report, Suansunandha Rajabhat University. pp. 53-62

[3] Kukich, K. (1992). Techniques for Automatically Correcting Words in Text. *ACM Computing Surveys (CSUR) 24(4)*: pp. 377-439.

[4] Sharma, S. and Gupta, S. (2015). A Correction Model for Real-word Errors. *Procedia Computer Science 70*: pp. 99-106.

[5] Golding, A. R. and Roth, D. (1999). A Winnow-Based Approach to Context-Sensitive Spelling Correction. *Machine Learning 34(1)*: pp. 107-130.

[6] Tapsai, C., Meesad, P., Haruechaiyasak, C. (2019). Complete Soundex for Similarity Analysis of Thai Words. *Information Technology Journal*, Year 16, Volumn 1, January- June 2019.

[7] Tapsai, C., Meesad, P., Haruechaiyasak, C. (2016). TLS-ART: Thai Language Segmentation by Automatic Ranking Trie. *Paper presented at The 9th International Conference Autonomous Systems*, October, 2016.

[8] Tapsai C. and Rojanabenjakun P. (2019). Value Added of Agricultural Products by Information Technology Through Eletronic Marketing. Proceedings of the ICBTS & ICTBH 2019 Conference in London. 7-8 March 2019.pp.72-77