Automatic Thai Subjective Examination using Cosine Similarity

Conference Paper · August 2018

DOI: 10.1109/ICAICTA.2018.8541276

2 authors, including:

Pusadee Seresangtakul
Khon Kaen University

Automatic Thai Subjective Examination using Cosine Similarity

Pongsakorn Saipech and Pusadee Seresangtakul
Natural Language and Speech Processing Laboratory
Department of Computer Science, Khon Kaen University
Khon Kaen, Thailand, 40002
s.pongsakorn@kkumail.com and pusadee@kku.ac.th

***Abstract* — This paper presents an automatic subjective examination system for Thai language based on semantic similarity, applying cosine similarity techniques together with the applicable synonym. The examination answers are input into the system in short Thai sentences, without word boundaries. The input answers which are reference answers, and student answers, were segmented into a sequence of Thai words; using the longest matching algorithm and dictionary (Lexitron®). These answers were transformed into vectors using TF-IDF, which is a frequency computation technique. The synonyms were also considered. Cosine similarity was used to measure the similarity of the answers. In order to evaluate the performance of the proposed system, we prepared a dataset of five questions from the final exam of the "Database Management System and Database Design" course, and asked 70 students to provide their answers. The datasets were prepared for both systems, with and without synonyms, in order to score the student answers. We crosschecked the scores of both systems with an expert, their teacher. In comparing the results of both systems, we found that the scores produced from the system using cosine similarity with synonyms were similar to those obtained from the expert.**

Keywords— Thai subjective examination; TF-IDF; cosine similarity; synonym

I. INTRODUCTION

Subjective examination is the usual method for educational evaluation in Thailand [1]. Capable of evaluating student comprehension, the subjective examination makes it difficult for the student to guess the answer, unlike in an objective examination [2]. This type of examination can assess the knowledge, skill, attitudes, and concepts of a student. However, the subjective examination is not without drawbacks. In some cases, a student with poor handwriting may not be clearly understood by the evaluator. And, the quality of the evaluation may change according to the mood of the evaluator, as within a time-consuming evaluation [3]. Because computer-based tests are a popular format, we developed an Automatic Thai Language Subjective Examination System. The system is based on semantic similarities between reference answers and student answers, in which the system verifies the similarity of a student's answer set with a series of reference answers. The Longest-matching approach [4] was utilized for word segmentation. Stop word elimination was used for removing words that do not affect the meaning of the sentence. Synonyms were used for considering words that are synonymous or very similar. Term Frequency-Inverse Document Frequency (TF-IDF) was the technique applied to convert data into vector form. Finally, Cosine Similarity was applied to compare the similarity of sentences. The similarity between two sentences (student answers and reference answers) determined the students' scores.

II. LITERATURE SURVEY

This section presents a review of the research and literature related to our work. The research in the field of automatic examination using cosine similarity technique has been referred to as an online examination, automated essay scoring, and online essay assessment. These research studies were conducted in a variety of languages; such as English, Indonesian, Arabic, Malayalam, and Thai.

Pramukantoro and Fauzi, 2016; [5] proposed an automatic scoring system for essay examination in English. Study data was obtained from a course entitled Operating Systems, which consisted of three questions, and the subsequent answers from the lecturers, as well as 50 selected students. The level of accuracy evaluated through Pearson correlation and MAE (Mean Absolute Error) achieved results of 0.588 and 5.330, respectively.

Lahitani et al., 2016; [6] presented the implementation of the TF-IDF method along with the cosine similarity approach to measure the degree of similarity within their Indonesian essay assessment. Test datasets consisted of ten student documents, derived from an e-learning source, which were then compared for similarity with the documents provided by their five experts. Similarity measurement was expressed in ranked based form, in which the best level of cosine similarity was 0.39.

More recently, Rababah and Al-Taani, 2017; [7] proposed an automated scoring approach for short answers to Arabic essay questions using cosine similarity. Again, the scoring process was based on the similarity between the student answers and the model answers. The dataset consisted of an exam of eleven questions, answered by 50 students, and for which there was a single model answer for each question. The accuracy of the proposed approach was evaluated by comparing the human scores with the cosine scores for all test questions. The experiment results showed scores near their human counterparts, with a percentage value of correlation equal to 95.4%.

978-1-5386-4804-9/18/$31.00 @2018 IEEE

Gokul et al., 2017; [8] presented the observations of sentence similarity between two Malayalam sentences using the cosine similarity method. The test data of 900 and 1400 sentence pairs within the FIRE 2016 Malayalam corpus were utilized as test data in two iterations, which obtained an accuracy of 0.8 and .59, respectively.

Aungkaseraneekul and Jaruskulchai, 2010; [9] presented an Automated Thai-Language Essay Scoring system using K-NN. The cosine similarity technique computed the similarity between the student and teacher answer sets. Five questions posed to 55 students produced accuracies between 76.73% and 92.73%.

III. RELATED WORKS

A. Word Segmentation

Thai language, unlike English, contains an unspecified language boundary, in which text processing requires word segmentation to identify word boundaries. Researchers have proposed numerous techniques for word segmentation, which can be classified into three approaches: dictionary based (DCB), machine learning based (MLB), and hybrid approaches [10-12]. The DCB uses a matching algorithm with the dictionary to segment the input text into word sequences. The main problem within the DCB occurs with unknown and ambiguous words. The MLB built statistical models based on the features of a tagged corpus, in an attempt to solve these problems. The hybrid approach is the combination of the DCB and the MLB approaches, which employs the advantages of each. In this study, the DCB approach was applied to the Thai word segmentation process. We applied the longest matching algorithm with the Lexitron® dictionary [13], consisting of 76,314 Thai words.

B. Stop word elimination

Stop words, by definition, are those words that appear in a text frequently, but do not carry significant information [14]. In Thai, there are some word categories that do not affect the meaning of a text; like prepositions, conjunctions, punctuation, or special symbols. These are removed from the document, while the meaning of content remains unchanged. Stop word elimination should be done before defining the frequency of the sentence, eliminating unwanted features, and reducing the word frequency in the document. Stop word removal reduces memory usage, and decrease time-consuming text processing.

C. Term Frequency – Inverse Document Frequency (TF-IDF)

TF-IDF is a classical feature classification function in the vector space model [15], defined as the product of the Term Frequency (TF) and the Inverse Document Frequency (IDF), where TF denotes term frequency of the word in the document, and IDF denotes the inverse document frequency in all documents. In TF-IDF, it is typically assumed that a good class of distinction occurs if a term has a high frequency in one document, and low frequency in all other documents. Therefore, the weight of the TF-IDF is proportional to the frequency of the word in the current document, and inversely proportional to the number of occurrences in the other documents.

The TF computation in TF-IDF merely considers the frequency of the feature word in a single document and does not take into account the overall distribution in all documents. In this way, feature words, which are uniformly distributed in a class, while important, may be ignored. The IDF takes into account the distribution of feature words in all texts but is more likely to select the more rare vocabulary. As a result, if the original TF-IDF algorithm is applied to a text feature selection, a large number of rare words may be selected. After which, additional dimensions of features will be needed to represent the document text as well. So, a naive TF-IDF feature selection algorithm may result in poor classification performance in the case of a low dimensional feature space. The equation for the TF-IDF is presented in equations (1) and (2), as follows

$$_{ij} = tf_{ij} \times idf_j \tag{1}$$

$$idf_j = 1 + log_e \frac{d}{df_j} \tag{2}$$

Where t is total term in document; tf_{ij} is the frequency of the term t_j in d_i document (term frequency); df_j is the total document that contains the term t_j; idf_j is the inverse from the document frequency; and d is the total document.

The weakness of the TF-IDF technique, in cases where the sentences are of different lengths, resulted in the weight of each sentence having a different value. That is, shorter sentences had a lower weight than those of longer sentences. Therefore, the weights of each sentence needed to be normalized. In order to solve this, the TF-IDF weight is determined by a score and frequency value in a similar or normalized pattern. The equation used to normalize the weight is:

$$tf_{ij} = \frac{n_{ij}}{\sum_k n_{kj}} \tag{3}$$

Where n_{ij} is the number of occurrences of the considered term t_i in document d_j; and n is the sum of the number of occurrences of all terms in the document d_j; that is the size of the document d_j.

D. Cosine Similarity

Cosine similarity [16] is the traditional method used to measure the degree of similarity between two vectors, obtained from the cosine angle multiplication value of two vectors, and is often combined with the TF-IDF. Weighted term results are used for the calculation of similarity between the reference answers and the student answers. Such similarities are often used in the process of data classification having like characteristics. Cosine similarity measures can be exploited to calculate the distance of the similarities being compared, and improve the accuracy of information retrieval. The equation for cosine similarity [17] is as follows (4):

$$SC(Q,D) = \frac{\sum_{j=1}^{t} W_j d_j}{\sqrt{\sum_{j=1}^{t} (d_j)^2 \sum_{j=1}^{t} (w_j)^2}} \tag{4}$$

Where t is total term; W_j is weight of j^{th} term in query; d_j is weight of j^{th} term in document; $\sum_{j=1}^{t} W_j d_j$ is Dot product or Cross product, through equation (5); $\sum_{j=1}^{t}(d_j)^2$ is a summary of document vector, through equation (6) and $\sum_{j=1}^{t}(w_j)^2$ is a summary of query vector, through equation (7).

$$\sum\nolimits_{j=1}^{t} W_j d_j = W_1 d_1 + W_2 d_2 + \cdots + W_n d_n \tag{5}$$

$$\sum\nolimits_{j=1}^{t} (d_j)^2 = \sqrt{{d_1}^2 + {d_2}^2 + \cdots + {d_n}^2} \tag{6}$$

$$\sum\nolimits_{j=1}^{t} (W_j)^2 = \sqrt{{W_1}^2 + {W_2}^2 + \cdots + {W_n}^2} \tag{7}$$

Cosine similarity measures the similarity between two sentences that are always between 0 and 1; thus having the cosine similarity value of 1.

IV. Automatic Thai Subjective Examination Architecture

In this section we present the automatic Thai subjective examination system, which is divided into two main stages; the preparation stage, and the similarity scoring stage, as shown in Fig. 1. First, the preparation stage aims to prepare a reference answer in terms of the TF-IDF vectors, consists of two processes; preprocessing and TF-IDF calculation. Secondly, the similarity scoring stage aims to evaluate the student answers, and determine the student scores. There are three processes in this stage; preprocessing, TF calculation, and similarity and scoring computation.

A. Preparation stage

The aim of this stage is to prepare the reference answer, which is a short Thai sentence, and transform it into vectors; namely, the TF-IDF vectors. The preparation stage consists of the following processes:

1) Preprocessing: Since the reference answer is in Thai, which is un-segmented language, the reference answer (sentence) is divided into a sequence of words, and any stop words are removed. The longest matching algorithm and Lexitron® dictionary are applied to segment the answer into words. These words are operated to eliminate stop words.

2) TF-IDF calculation: After the preprocessing process, word frequency is calculated in terms of TF-IDF. The TF-IDF of the reference answer is stored in the reference answer database.

B. Similarity scoring stage

In the similarity scoring stage, a student answer is also evaluated in a preprocessing process, along with the consideration of the word's synonym, which is a word, or phrase that means exactly or nearly the same as another word or phrase in the same language, and then transformed into TF vectors. The student vectors are then compared with the reference vectors, using the cosine similarity method, stored in the database, to evaluate the score of the student answer.

1) Preprocessing: The student's answers are operated similarly to the preparation stage, utilizing both word segmentation and stop word elimination. Additionally, the sequence of words is considered to determine the synonym.

2) TF calculation: A sequence of words, after undergoing the preprocessing process, is calculated for frequency in terms of TF vectors.

3) Similarity and Scoring computation: In this process, cosine similarity is applied to evaluate the similarity of the reference answer and the student answer. The TF-IDF of the reference answer is queried from the database, and compared with the TF of the student answer. The cosine similarity of the two answers ranged from 0 to 1. A value close to 1 means that the student's answer is very similar to the reference answers, whereas a value close to 0 implies that the student's answer is distant from the reference answer. Then, cosine similarity is used to calculate the actual score of the student's answer with the maximum score possible in each question, which is defined as follows, in equation (8)

$$score_{ij} = exScore_i \times cosine_{ij} \tag{8}$$

Whereas $score_{ij}$ is score of the j^{th} student's answer of i^{th} question; $exScore_i$ is full score of the i^{th} question; $cosine_{ij}$ is cosine similarity of the j^{th} student's answer of i^{th} question.

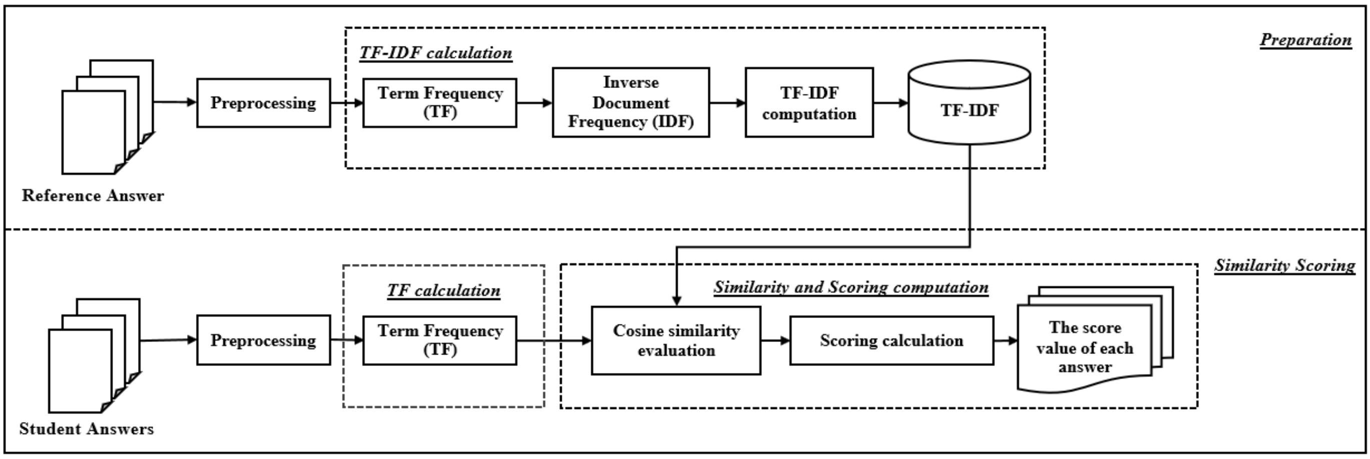

Figure 1. Overview of the Automatic Thai Subjective Examination Architecture

V. Experiments And Results

In order to evaluate the performance of the proposed system, we conducted the following experiment: We asked the professor of the "Database Management System and Database Design" course to prepare a question dataset and provide an 'expert' answer for each question. We then asked 70 students to join the experiment by giving their answers to the questions. The professor then evaluated the answers, and provided a score for their answers, which were then compared with scores obtained from the proposed system in both methods; namely with and without the synonym, in which to compare the similarity with the reference answer.

A. Data Set

The dataset utilized in our experiment was collected from 70 students enrolled in the "Database Management System and Database Design" course and was acquired from the descriptive answers of five subjective questions present in the course exam. The students' answers were compared with the reference answers, resulting in their scores. The answers were scored between '0' to '1', with '1' as the best or closest answer. A sample question from the exam was "อธิบายระบบฐานข้อมูล" (in Thai), which means "Describe a database system?" The answer, written as a short explanation in Thai was: "ระบบฐานข้อมูลคือระบบเก็บบันทึกข้อมูลด้วยคอมพิวเตอร์โดยมีจุดประสงค์เพื่อจัดเก็บข้อมูลและเพื่อให้ผู้ใช้สามารถเรียกดูและปรับปรุงอัพเดตข้อมูลตามต้องการ" meaning: "A database system is a computerized record keeping system, whose overall purpose is to store information, and to allow users to retrieve and update that information on demand."

B. Effectiveness Evaluation

To evaluate the effectiveness of the proposed system, the answers from both the student and the system undergo the preprocessing process, in which cosine similarity computes the similarity value, which is between 0 and 1. A value close to 1 means that the student's answer was very similar to the reference answers, whereas a value close to 0, implies that the student's answer is distant from the reference answer. The output similarity represents the conversion of the actual score with the maximum score possible in each question; which is defined as three. The results of the student's answers for each question are compared with the expert's answers, shown in Fig. 2-6.

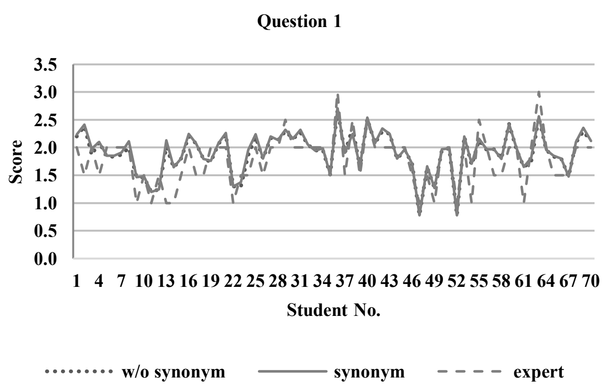

Figure 2. Scores (Question 1) evaluated by the expert vs. the system; with synonym and without synonym.

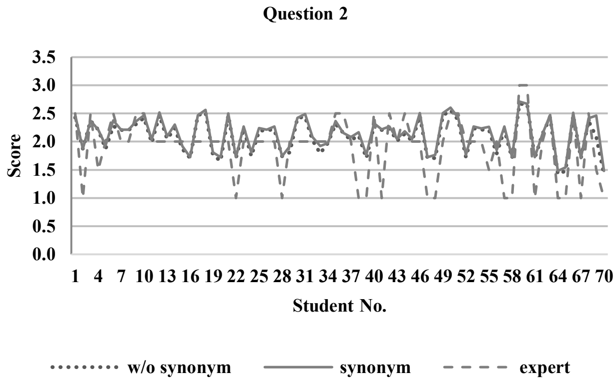

Figure 3. Scores (Question 2) evaluated by the expert vs. the system; with synonym and without synonym.

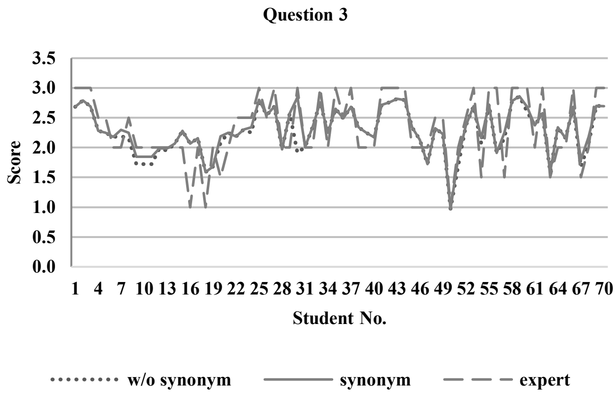

Figure 4. Scores (Question 3) evaluated by the expert vs. the system; with synonym and without synonym.

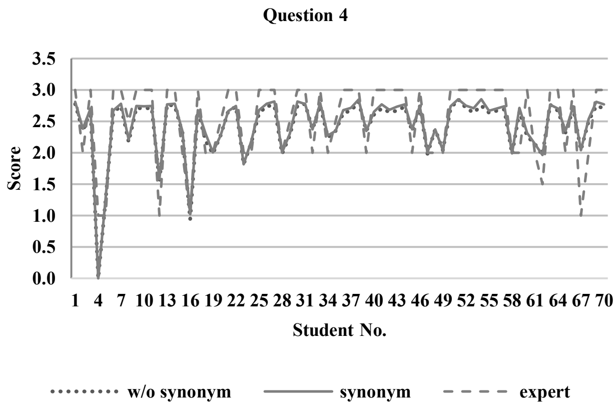

Figure 5. Scores (Question 4) evaluated by the expert vs. the system; with synonym and without synonym.

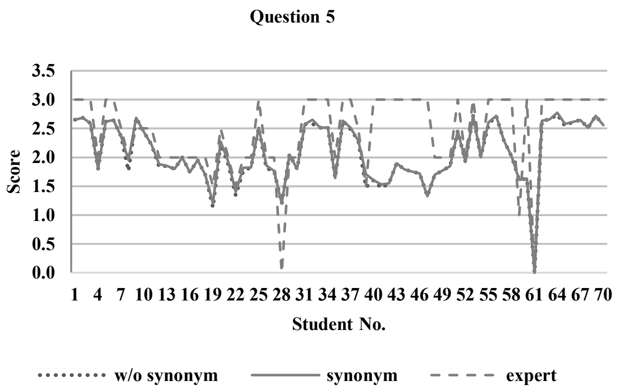

Figure 6. Scores (Question 5) evaluated by the expert vs. the system; with synonym and without synonym.

The results of the experiment revealed that the scores obtained from the system in both methods, compared with the expert evaluation were nearly identical to the scores obtained from the experts. In addition, the system which utilized cosine similarity computation with synonyms was more effective than the computation without use of the synonym. Although most of the scores indicated that the performance of the proposed system proved more accurate, errors did occur; for instance, the answer of the fourth student in Fig.5. The answer was scored as '0' by the proposed system, while the expert scored it as '1'. We determined that this was caused within the preprocessing process, in which there is a lack of words used to compare the similarity with the reference answers. Moreover, the proposed system assesses by comparing the similarity between answers of students and reference answers in term of word frequency, while the expert assesses by meaning and understanding of the students' answers. That is why the score of some case, which gotten by the proposed system is very different from the score gotten from the expert such as the answer of student number 28^{th}, 40^{th} and 46^{th} in Fig. 6. Furthermore, the proposed system evaluated the student's answer and provided the appropriate score. The range of scores obtained from the system is rather detailed in comparison with obtained from the expert for instance the system scores as 0.1, 0.2, 0.3,..., 2.8, 2.9, 3.0 while the expert scores as 1, 1.5, 2.5.

VI. CONCLUSION

This article presents an automatic Thai subjective examination using cosine similarity within short descriptive answers. The students of 70 participated in the examination, which consisted of five subjective questions. To evaluate the efficiency of the system, we compared the scores of the students' answers, provided by our system along, together with synonyms and without synonyms, with those of the expert. The results suggest a similar trend in the scores of the proposed system and those of the expert. However, in some cases, we noticed that scores from our system seemed rather low, as a result of the preprocessing process. System efficiency was reduced, in that we considered only the frequency of the words used and synonym without the benefit of context analysis. To improve the efficiency of the proposed system, we intend to integrate semantic and context analysis into future work.

ACKNOWLEDGMENT

This work was supported by scholarship from the Computer Science Department, Faculty of Science, the Graduate School of Khon Kaen University, Khon Kaen, Thailand.

REFERENCES

[1] C. Suwanroui, "Comparative analysis of subjective and objective questions," in *Payap Technology and Business College*, 2012. (in Thai)

[2] S. Aungkaseraneekul and C. Jaruskulchai, "Automated Thai-Language Essay Scoring using K-NN," in *2010 The 1st International and the 14th National Computer Science and Engineering Conference (ICSEC & NCSEC)*, 2010. (in Thai)

[3] P. Kudi, A. Manekar, K. Daware and T. Dhatrak, "Online Examination with short text matching," in *2014 IEEE Global Conference on Wireless Computing & Networking (GCWCN)*, 2014, pp. 56 - 60.

[4] Yuen Poowarawan, "Dictionary-based Thai Syllable Separation," in *Proceedings of the Ninth Electronics Engineering Conference*, 1986.

[5] E. S. Pramukantoro and M. A. Fauzi, "Comparative analysis of string similarity and corpus-based similarity for automatic essay scoring system on e-learning gamification," in *2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 2016, pp. 149–155.

[6] A. R. Lahitani, A. E. Permanasari, and N. A. Setiawan, "Cosine similarity to determine similarity measure: Study case in online essay assessment," in *2016 4th International Conference on Cyber and IT Service Management*, 2016, pp. 1–6.

[7] H. Rababah and A. T. Al-Taani, "An automated scoring approach for Arabic short answers essay questions," in *2017 8th International Conference on Information Technology (ICIT)*, 2017, pp. 697–702.

[8] P. P. Gokul, B. K. Akhil, and K. K. M. Shiva, "Sentence similarity detection in Malayalam language using cosine similarity," in *2017 2nd IEEE International Conference on Recent Trends in Electronics, Information Communication Technology (RTEICT)*, 2017, pp. 221–225.

[9] S. Aungkaseraneekul and C. Jaruskulchai, "Automated Thai-language essay scoring using K-NN," in *Proceedings of 48th Kasetsart University Annual Conference: Science*, 2010. (in Thai)

[10] A. Srithirath and P. Seresangtakul, "A hybrid approach to Lao word segmentation using longest syllable level matching with named entities recognition," in *International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology*, 2013, pp. 1-5.

[11] C. Haruechaiyasak, S. Kongyoung and Ma. Dailey, "A comparative study on Thai word segmentation approaches," in *International Conference on Electrical Engineer-ing/Electronics, Computer, Telecommunications and Information Technology*, 2008, pp. 125-128.

[12] S. Somsap and P. Seresangtakul, "Isarn Dharma Word Segmentation," in *Proceeding of the In-ternational Conference on Control Automation and Information Sciences (ICCAIS)*, 2013, pp. 53-57.

[13] NECTEC. LexTo., [Online]. Available: http://www.sansarn.com/lexto/lincense-lexto.php

[14] N. Sura, "Automatic Thai Text Categorization Using FPTC Algorithm," *Master Thesis, Computer Science, King Mongkut's Institute of Technology North Bangkok*.

[15] J. Ramos, "Using TF-IDF to Determine Word Relevance in Document Queries," [Online]. Available: https://www.cs.rutgers.edu/~mlittman/courses/ml03/iCML03/papers/ramos.pdf

[16] AR. Lahitani, AE. Permanasari and NA. Setiawan, "Cosine similarity to determine similarity measure: Study case in online essay assessment," in *International Conference on Cyber and IT Service Management*, 2016.

[17] D. A. Grossman and O. Frieder, Information retrieval: Algorithms and heuristics. *Springer Science & Business Media*, 2012, vol. 15