ThaiQCor 2.0: Thai Query Correction via Soundex and Word Approximation

Santipong Thaiprayoon, Alisa Kongthon, and Choochart Haruechaiyasak
National Electronics and Computer Technology Center (NECTEC), National Science and Technology Development Agency (NSTDA), Pathumthani, Thailand

Abstract—**Nowadays, search engine is an important tool for enabling users to search for information on the Internet. One of the most important problems of searching is inaccurate typing due to typographical and cognitive errors. Typographical errors are normally resulting from typing mistakes from adjacent letters on a keyboard layout. Cognitive errors are due to the lack of user knowledge in query term spelling. To solve the problems, we designed and developed a new version of Thai query correction program called *ThaiQCor 2.0* that can handle both typographical and cognitive errors. Our program consists of two main approaches, word approximation and soundex. Word approximation employs the approximate string retrieval technique including character edit distance calculation. This approach aims to solve the typographical errors. Soundex applies the grapheme-to-phoneme conversion and then performs string matching approximation by calculating the edit distance of weighted phonemes from phoneme sequences. The objective of this approach is to handle the cognitive errors. All candidate words from both approaches are ranked based on their scores and suggested to the user. The experimental results showed that *ThaiQCor* 2.0 achieves the accuracy of 97.11% and 89.76% for place names and person names, respectively.**

Keywords-soundex; word approximation; grapheme-to-phoneme conversion; edit distance of phonemes; query correction

I. INTRODUCTION

Nowadays search engine is an important tool for facilitating users to find the information they need, especially on the Internet. Many websites often utilize web search to help their users seek for useful information such as products details, food prices, and movie promotions. Search engines find information from queries that users enter as input. The tool will then return search results to the user in less time.

One of the important processes of searching information on the Internet is typing query. Sometimes some Internet users waste their time finding the information they need due to inaccurate typing. The typographical mistakes typically occurring in two cases: (1) Typographical errors [6] and (2) Cognitive error or phonetic error [3]. The cause of typographical errors is adjacent letters on a keyboard layout that affects typological mistake of a user. For example, the word "ลาดพร้าว" is mistyped as the word "ลากพร้าว" because the character "ด" is close to the character "ก" on the keyboard. This example shows that the character "ด" is replaced by the character "ก". Generally, typing mistake is a result of insertion, substitution, deletion and transposition of characters. For the case of cognitive error, it is a result of misspelling query from a user who does not know how to spell search term correctly. More specifically, a person name or a specific name such as restaurant name, street name and product name is quite difficult to spell due to the uniqueness of the name. For this reason, Internet users try to spell by using similar pronunciation instead. For instance, the word "สาทร" is misspelled as the word "สาธร" because the character "ท" has similar sound with the character "ธ". With two examples mentioned above, typos are an important issue of the search engine tools. Moreover, these errors affect many Internet users to find the information to meet the results that they expect.

In our previous research studies, Angkawattanawit et. al. [1] developed a Thai query correction system called ThaiQCor version 1.0 that can verify an inaccurate query and correct it. This system was composed of two correction modules: word approximation version 1.0 and soundex version 1.0. For word approximation module, all initial words were reserved in a trie structure for fast retrieval of the candidate words. This module applied edit distance technique to calculate the distance between two words. This technique selected words from the trie structure that have a low distance to the word in question. Soundex module removed word occurring tone markers to reduce all tone levels to base tone level as a middle tone. Moreover, this module eliminated the Thanthakhat (์) which is a cancellation mark used to indicate silent final consonants in Thai writing system. because these marks have no significance in Thai word and are not pronounced as well. With the removal of tone marks and cancellation marks was sent to a grapheme-to-phoneme conversion (G2P) to generate a sequence of phones. Next, Chotimongkol et. al. [4] proposed an approximate sound search technique called soundex 1.5 version. Their proposed research was developed to support flexible searching the proper names. In addition, their research appended the edit distance of phonemes calculation with similar pronunciation. The objectives of their proposed research consist of two points. The first point is to deal with a limitation of the soundex 1.0 that cannot support

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

large vocabulary set because this version stored all words in trie structure allocated in the memory system. When increasing numerous words, the system crashes due to full memory. The second point is to apply a G2P conversion implemented by Thangthai et. al. [2]. This G2P version employed technique of collecting the collection of syllable pattern. However, soundex 1.0 still has some problems. The main weakness is that G2P has some errors in term of generating the most probable phones list. Another issue is weighting edit distance of phonemes that does not rely on the group of similar sound.

To overcome the problems, we designed and developed a new Thai query correction program called ThaiQCor 2.0 that can handle both typographical and cognitive errors. Our program consists of two main approaches: word approximation 2.0 and soundex 2.0 approaches. Word approximation 2.0 applies n-gram technique to resolve store and search problems using building inverted index and includes the substitution rules of consonants having similar sound and consonants normalization. Soundex 2.0 applies the latest G2P based on machine learning approach including Condition Random Fields (CRFs) technique. The objectives of this version are syllable segmentation and phoneme sequences prediction that achieved 96.09% accuracy. This version has no limit to the number of syllable forms and also yields more accuracy. More details about this technique can be found in Saychum [7]. In addition, this approach measures the distance of phonemes cost by defining weights following the group of similar sound and reducing the linking syllables in form of phonemes sequence. This approach tackles the limitations of Soundex 1.0 and can enhance the quality of accurately correcting the specific names and person names. Finally, both word approximation and soundex approaches are combined. All candidate words from both approaches are scored and ranked. The words with the high score will be suggested. As a result, our program can correct and suggest better query terms that help a user obtains satisfactory results.

The remainder of this paper is organized as follows. In next section, we review some related work on Thai phonology and groups of similar sounds. In Section 3, we describe details of two approaches of Thai query correction and explain the overview of ThaiQCor 2.0 version. The details of each approach are given with illustrations. In Section 4, we describe the experiments which evaluate the proposed approaches on the test sets of place names and personal names and discusses the results. Finally, the conclusions and suggestions for future works are presented in the last section.

II. Related Works

Several studies have been proposed to solve typographical issues that are a result of typing and spelling errors. Karoonboonyanan et al. [9] presented that one drawback encoding schemes is each character has the same encoding regardless of its position in a syllable except for the first character. This problem addressed by taking into account Thai syllable structure in their soundex coding scheme. Seung-Shik Kang [8] proposed a new method of a word similarity calculation method for syllable-structured languages. This metric was devised to improve the performance of the syllable-based and letter-based metrics. The accuracy of the word similarity calculation through the phonetic transformation by pronunciation rules and word-length normalization. Kenneth et. al. [5] described the development of a spelling correction system for medical text based on Shannon's noisy channel model. They applied spell checker to three different types of free-text data: clinical notes, allergy entries, and medication orders.

From Thai phonology perspective, this work groups Thai phonemes into four categories after the level of similarities of their characteristics:

1) *high-similarity phoneme group*: this level is a group of phonemes that can be written in many forms. For example, the phoneme /th/ is able to be written up to five forms, "ท", "ธ", "ถ", "ฒ" and "ฐ". In the meantime, a written form can be pronounced as one or more sounds. For instance, the consonant "ท" can be pronounced as either /d/ or /th/. Hence, this group is defined to have the shortest distance.

2) *middle-similarity phoneme group*: this level is a group of phonemes that are similar in terms of the aspiration (aspirated/unaspirated), the voicing (voice/voiceless), and the consecution (monophthong/diphthong). This group has members that are slightly differences. For example, the consonants "บ" and "ป" are voiced and unvoiced. Therefore, these phonemes and their written forms are grouped and marked as having middle-similarity.

3) *low-similarity phoneme group*: this level is a group of phonemes after their manner of articulation and their place of articulation. Then, plosive and affricate are in the same group, as same as nasal and approximant, fricative and affricate, alveolar and postalveolar.

4) *least-similarity phoneme group*: this level represents the similarities of final consonant sounds. Thai has conventionally eight final sounds namely, /k/, /d/, /b/, /ŋ/, /n/, /m/, /j/, and /w/. Therefore, these phonemes are grouped with the least-similarity.

Most of the previous studies have focused on correcting either edit distance of characters or phonemes. To correct typographical errors and cognitive, we combine phonemes similarity and the edit distance of characters approaches together and rank these words based on their score.

III. Processing Overview of ThaiQCor 2.0

In this section, we first describe the process overview of ThaiQCor 2.0 that consists of two approaches. The details of each approach are explained in the following subsections.

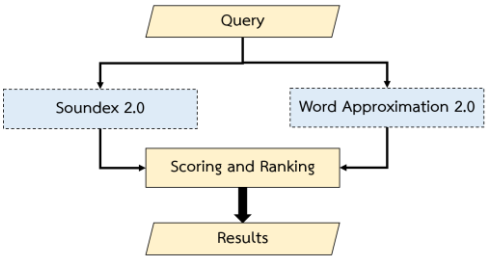

Figure 1. The processing overview of ThaiQCor 2.0

A. *The Process Overview of ThaiQCor 2.0*

ThaiQCor 2.0 consists of two approaches: word approximation 2.0 is utilized for correcting the typographical error and soundex 2.0 is applied for resolving cognitive errors. The overall process is illustrated in Figure 1. When use type a query, the query will be sent to word approximation 2.0 and soundex 2.0 approach to find a list of candidate words that are most likely to be matched to the search terms both similar sounds and similar words. Then the list of words obtained from each approach is calculated by considering the words order and co-occurring words. Next, the list of calculated words is sorted in descending order based on their score. The words with the high score will be suggested. Lastly, our program returns the suitable word results to a user.

B. *Word Approximation 2.0*

This approach is the process of word approximation search that can be divided into two main parts: index of characters generation and list of candidate words search. The details of each part are explained in the following subtopics.

1) Building index of characters: This part utilizes character n-gram method to indexing a sequence of characters. This module applies the n-gram of three characters size called trigram. For example, the word "สาทร" is sliced to {สาท, าทร}. This example shows that the characters sequence becomes a set of overlapping and continuous characters. For indexing the words, all initial words are built by the above process and saved them into the index of characters as an inverted index of characters.

2) Searching the list of candidate words: This part consists of three main modules. The details of each module are explained in the following subtopics. The process of word approximation search is illustrated in Figure 2.

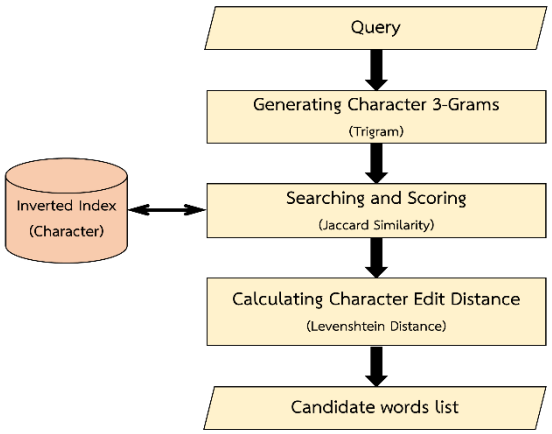

Figure 2. The process of word approximate search

a) Generating character 3-grams: The query is segmented by using the character 3-gram method. The segmented query transforms to a set of characters.

b) Searching and scoring: This module aims to find and select the list of candidate words from the inverted index of characters. In step of comparing between the set of characters as query term and the set of characters in the inverted index of characters, this step is applied the jaccard similarity coefficient to measure a similarity between two sets of characters. The candidate sets of characters as candidate words list having the suitable similarity score are selected.

c) Calculating edit distance of words list: This module is the measurement of edit distance calculated between the list of candidate words and the query by using levenshtein edit distance and consonant normalization. The candidate words similar to the query are selected. For example, the correct word in the index of characters is "บางปะอิน". The search term is "บางประอินท์". The characters distance between two words equals 3. This is, three characters are inserted to the search term namely, "ร", "ท", and "์".

C. *Soundex 2.0*

This approach is the process of soundex that can be divided into two major parts: index of phonemes generation and soundex search. The details of each part are explained in the following subtopics.

1) Building index of phonemes: This part utilizes character 3-gram method to index a sequence of phonemes. For instance, the word "พงษ์เทพ" is converted from a written form to a phonemes sequence as {ph-o-ng^, o-ng^-th, ng^-th-ee, th-ee-p^}. This example shows that the phonemes sequence becomes a set of overlapping and continuous phonemes. For indexing the words, all initial words are built by the above process and saved into the index of phonemes as the inverted index of phonemes.

2) Soundex search: This part consists of four major modules. The details of each module are explained in the following subtopics. The process of soundex search is illustrated in Figure 3.

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

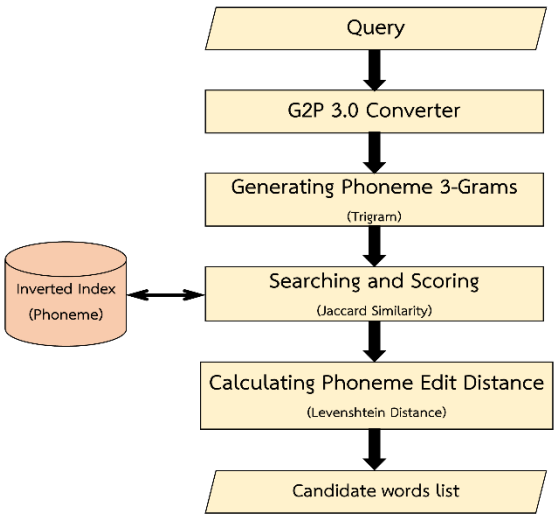

Figure 3. The process of soundex search

a) Grapheme-to-phoneme converter: This module is to convert a word or written form, as a series of characters or graphemes, to a pronunciation, as a series of phones. For instance, the word "พงษ์เทพ" is converted from a written form to a phones form as {ph-o-ng^0|th-ee-p^1}.

b) Generating phoneme 3-grams: This module aims to generate a set of phonemes that is overlapping and continuous.

c) Searching and scoring: This module is the process of finding and selecting the list of candidate phonemes sequence from the inverted index of phonemes. In step of comparing between the set of phonemes as query term and the inverted index of phonemes, we utilized the jaccard similarity coefficient to measure the similarity between two phonemes sequences. The candidate phonemes sequences having the proper similarity score are selected.

d) Calculating edit distance of phonemes: This module is to compare and measure the distance in the list of candidate phonemes sequences. In addition, this module utilizes the linguistic knowledge to improve the algorithm of edit distance calculation. To accurately improve how to find the phonemes sequences, this algorithm includes the weight values corresponding to the phonemes of Thai phonology. To accurately measure the distance between two phonemes sequences, the costs of edit operations, insertion, deletion, and substitution, can be adjusted according to the phoneme similarity analysis. A substitution of phonemes within the group of similar phonemes is assigned a lower cost, e.g. 0.5 instead of 1. But substitution between other phonemes, phoneme deletion, and phoneme insertion still have a cost of 1. For example, a substitution of phoneme /kr/ by phoneme /k/ has a cost of 0.5 to reflect that phoneme /r/ is often dropped from a consonant cluster, while a substitution of phoneme /k/ by phoneme /m/ still has a cost of 1. With a phoneme similarity edit distance, the search term as "บางประอินทร์" /b-aa-ng^-0|**pr**-a-z^-1|z-u-n^-0|/ and the correct spelling in the inverted index of phonemes as "บางปะอิน" /b-aa-ng^-0|**p**-a-z^-1|z-i-n^-0/ has a distance of 0.5 instead of 1 because phoneme /p/ and phoneme /pr/ are consonant clusters that are in the same group of similar phonemes.

D. Words Scoring and Ranking

To efficiently enhance query correction of ThaiQCor 2.0, we proposed a formula for calculating the score of each word obtaining from both word approximation 2.0 and soundex 2.0 approaches. The calculation considers the words order and co-occurring words from both approaches. By definition, $W_{approx} = \{w_{a1}, w_{a2}, \dots, w_{a3}\}$ is a set of words obtaining from word approximation 2.0. $W_{sound} = \{w_{s1}, w_{s2}, \dots, w_{s3}\}$ is a set of words obtaining from soundex 2.0. $T_{all} = W_{approx} \cup W_{sound} = \{t, t \in W_{approx} \vee W_{sound}\}$ is a set of terms combined to a set of T_{all}. After that, each term in T_{all} is given a score such score is calculated based on the formula shown in the equation 1.

$$Score_{(t,T_{all})} = \begin{Bmatrix} N_{approx} - rank_t, \{t | t \in W_{approx} \wedge t \in W_{sound}\} \\ N_{sound} - rank_t, \{t | t \notin W_{approx} \wedge t \in W_{sound}\} \\ \frac{(N_{approx} - rank_t) + (N_{sound} - rank_t)}{2} + 1, \ \{t | t \in W_{approx} \wedge t \in W_{sound}\} \end{Bmatrix} \quad (1)$$

Where

$Score_{(t,T_{all})}$ is the score of each term in T_{all}.

N_{approx} is the number of all words in word approximation 2.0 approach.

N_{sound} is the number of all words in soundex 2.0 approach.

$Rank_t$ is the position of the term.

After completion in the calculation of the scores, the calculated terms are sorted by descending based on their score. The terms with a high score will be suggested.

IV. Experimental Results and Discussions

The accuracy evaluation of ThaiQCor 2.0 can be described in the following subtopics.

A. Accuracy Evaluation of ThaiQCor 2.0

To evaluate the performance of ThaiQCor 2.0, we prepared the words corpus that consists of place names 61,738 words and person names 2,890 words. We constructed the test sets to evaluate our approaches. The test sets, which is the list of wrong words from wrong typing and spelling, consists of place names 868 words and person names 254 words.

For the evaluation, we use accuracy as a metric. The equation for calculating accuracy is presented in the equation 2.

$$N - Best\ Accuracy = \frac{Number\ of\ words\ where\ approach\ can\ search\ correctly}{Number\ of\ words\ in\ the\ test\ set} \quad (2)$$

The methods of accuracy evaluation of word approximation consist of two methods: (1) ThaiQCor 1.0 applied the word approximation and soundex techniques. (2) ThaiQCor 2.0 applied the techniques that are similar to ThaiQCor 1.0 but include approximate string search technique including the edit

distance of characters and phonemes calculation and consonant normalization. The evaluation results are summarized in Table 1.

TABLE I. EVALUATION RESULTS OF THAIQCOR 2.0

Approaches	Place names		Person names	
	nBest=1	*nBest=5*	*nBest=1*	*nBest=5*
ThaiQCor 1.0	52.07	89.97	12.59	51.57
ThaiQCor 2.0	88.82	97.11	75.19	89.76

From the Table 1, we can conclude that ThaiQCor 2.0 yielded an average accuracy performance up to 88.82% for place names and 75.19% on person names at 1-best candidate. With the list of 5-best candidates, the average accuracy climbed up to 97.11% on the place names and 89.76% on the personal names. The experimental results showed that ThaiQCor 2.0 yielded better query correction than ThaiQCor 1.0 in the all test sets because ThaiQCor 2.0 was adopted the n-gram technique to find the most similar words and applied distance calculation of phonemes cost by defining the different weight value following the group of similar sound and reducing the linking syllables in form of phonemes sequence.

The experimental results indicated that the word approximation approach has some drawbacks when evaluating on person names because the most person names often consist of many cancellation marks in a word such as the words "ศักดิ์สิทธิ์" and "พันธ์ศักดิ์". However, the word approximation approach is more appropriate for words from inaccurate typing rather than spellings and it also is appropriate in the domain of place names. For soundex approach, the experimental results showed that this approach yielded the good accuracy for searching in the domain of person names because it can handle misspelling well.

V. CONCLUSION AND FUTURE WORKS

In this paper, we designed and developed a new Thai query correction program called ThaiQCor 2.0. The main purpose of this program is to handle both typographical and cognitive errors. Our program consists of two main approaches: word approximation and soundex approaches. Word approximation approach employs approximate string retrieval technique including character edit distance calculation. Soundex approach is applied with a grapheme-to-phoneme conversion for performing approximate string matching on phonemes sequences. The distance between these sequences is calculated based on edit distance of weighted phonemes. With the combination of word approximation and soundex approaches, ThaiQCor 2.0 yielded the average accuracy with 1-best candidate up to 88.82% for place names and 75.19% on person names. With the list of 5-best candidates, the average accuracy climbed up to 97.11% on the place names and 89.76% on the personal names. The experimental results showed that ThaiQCor 2.0 yielded better query correction than ThaiQCor 1.0 in all test sets. In the future to improve our program, we plan to calculate the proximity between Thai keyboard keys by adjusting the cost of edit operations. In soundex approach, we will improve the performance of generating phones on G2P conversion based on a Long Short-Term Memory (LSTM) recurrent neural network (RNN).

REFERENCES

[1] Angkawattanawit, N., Haruechaiyasak, C., & Marukatat, S. (2008). Thai Q-Cor: Integrating Word Approximation and Soundex for Thai Query Correction. In Proceedings of ECTI-CON 2008, 121-124.

[2] A. Thangthai, C. Hansakunbuntheung, R. Siricharoenchai, and C. Wutiwiwatchai, "Automatic Syllable-patternInduction in Statistical Thai Text-to-phone Transcription," in Proc. of Interspeech, Pittsburgh, Pennsylvania, USA, Sep. 2006.

[3] Bonaventura, Patrizia et al. "Phonetic Rules for Diagnosis of Pronunciation Errors," KONVENS (2000).

[4] Chotimongkol, A., Thaiprayoon, S., & Thatphithakkul, S. (2014). Flexible Proper Name Search using Sound Approximation Approach. In Proceedings of Oriental-COCOSDA 2014.

[5] Kenneth H. Lai, Maxim Topaz, Foster R. Goss and Li Zhou, "Automated misspelling detection and correction in clinical free-text records," Journal of Biomedical Informatics, pp. 188-195, 2015.

[6] R. Morris and L. L. Cherry, "Computer detection of typographical errors," in IEEE Transactions on Professional Communication, vol. PC-18, no. 1, pp. 54-56, March 1975.

[7] Saychum, S., Kongyoung, S., Rugchatjaroen, A., Chootrakool, P., Kasuriya, S., & Wutiwiwatchai, C. (2016). "Efficient Thai Grapheme-to-Phoneme Conversion Using CRF-Based Joint Sequence Modeling," In Proceedings of Interspeech 2016, 1462-1466.

[8] Seung-Shik Kang, "Word Similarity Calculation by Using the Edit Distance Metrics with Consonant Normalization," Journal of Information Processing Systems, vol. 11, no. 4, pp. 573-582, 2015. DOI: 10.3745/JIPS.04.0018.

[9] T. Karoonboonyanan, V. Sornlertlamvanich, and S.Meknavin, "A Thai soundex system for spelling correction," in Proc. of the National Language Processing Pacific RimSymposium 1997 (NLPRS-97), Phuket, Thailand, pp. 633-636, Dec. 1997.

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