Global Journals LaTeX JournalKaleidoscope™
Artificial Intelligence formulated this projection for compatibility purposes from the original article published at Global Journals. However, this technology is currently in beta. *Therefore, kindly ignore odd layouts, missed formulae, text, tables, or figures.*

Thai to Khmer Rule-Based Machine Translation using Reordering Word to Phrase

Sukchatri Prasomsuk[1] and Sukchatri Prasomsuk[2]

[1] University of Phayao

Received: 8 December 2016 Accepted: 5 January 2017 Published: 15 January 2017

Abstract

In this paper, an effective machine translation system from Thai to Khmer language on a website is proposed. To create a web application for a high performance Thai- Khmer machine translation (ThKh-MT), the principles and methods of translation involve with lexical base. Word reordering is applied by considering the previous word, the next word and subject-verb agreement. The word adjustment is also required to attain acceptable outputs. Additional steps related to structure patterns are added in a combination with the classical methods to deal with translation issues. PHP is implemented to build the application with MySQL as a tool to create lexical databases. For testing, 5,100 phrases and sentences are selected to evaluate the system. The result shows 89.25 percent of accuracy and 0.84 for F-Measure which infers to a higher efficiency than that of Google and other systems.

Index terms— thai khmer translation, machine translation (MT), rule based, pattern-based.

1 Introduction

ssociation of Southeast Asian Nations (ASEAN) consists of ten countries with various cultures and languages. Thailand and Cambodia are included in ASEAN, and the eastern border of Thailand is adjacent to Cambodia. Therefore, efficient communication is significant for international relations between these two countries. Cambodian natives have Khmer as a national language while formal language in Thailand is Thai. The linguistic differences of Thai and Khmer in both writing and speaking contribute to a translation barrier. For instance, since Thai language has been adapted partly from Pali, Sanskrit and Old Khmer, Thai vocabulary is relatively diverse. Thai language also contains complex orthography and relational markers. Furthermore, standard written Thai is complicated due to various combinations of syllabic alphabets, which consists of 44 basic consonants, 21 vowel symbols and 4 tone diacritics, applied under the rule that all diacritics appear in front of, above or below the consonants. Furthermore, Thai syntax has a noun classifier system as well as conforms to a basic sentence structure called subject-verb-object (SVO) Khmer contains 33 consonants, 23 dependent vowels and 15 independent vowels; however, no tone is presented. Due to the linguistic differences, current Thai-Khmer translation systems have scarcely achieved complete and accurate outputs. Moreover, the existent systems have rarely been created and developed. There is also a shortage of intellectuals who are competent in both languages and able to convey knowledge for creating a system of translation. As a result, the improvement of the Thai-Khmer translation system has been disrupted. Document translation between Thai and Khmer which requires high accuracy has consequently encountered difficulties. To solve the issues, machine translation (MT) from Thai to Khmer language requires development.

The proposed system in this paper implements translation techniques including rule-based algorithm with verification of sentence patterns to improve translation quality. The overview operation of the translation system is to input a Thai language text in a web application and then convert it into a desired output in Khmer. A lexical analyzer is first applied in the process to divide Thai sentences or phrases into individual syllabic words so that the separated words are analyzed and processed in the following steps resulting in Khmer sentences.

2 II. Related and Previous Works

There have been many attempts to research on machine translation between Thai and other languages. English-Thai machine translation was developed in 1998 with regard to the sentence-based technique which combines the rule-based and the example-based method to establish a system for English to Thai sentence translation [1]. However, the research result of performance evaluation and comparison was not indicated. In 2012, a technique called generalized patterns is presented to improve machine translation from Japanese to Thai language [2]. The method was compared to the others implemented in Google and Bing translators by executing 3,107 Japanese sentences in testing. F-Measure score was applied to assess performance of the translator.

Machine translation between Khmer and other language has also been researched. One of the studies selected Moses DoMY CE, which is statistical machine Abstract-In this paper, an effective machine translation system from Thai to Khmer language on a website is proposed. To create a web application for a high performance Thai-Khmer machine translation (ThKh-MT), the principles and methods of translation involve with lexical base. Word reordering is applied by considering the previous word, the next word and subject-verb agreement. The word adjustment is also required to attain acceptable outputs. Additional steps related to structure patterns are added in a combination with the classical methods to deal with translation issues. PHP is implemented to build the application with MySQL as a tool to create lexical databases. For testing, 5,100 phrases and sentences are selected to evaluate the system. The result shows 89.25 percent of accuracy and 0.84 for F-Measure which infers to a higher efficiency than that of Google and other systems.

with a horizontal and vertical writing direction from left to right and from top to bottom, respectively. Similarly, translation (SMT), as a tool to create an online system for English -Khmer translation based on Python, XML and HTML language in 2013 [3]. There is also research in 2014 on developing a French-Khmer dictionary called 'MotàMot' [4]. In 2015, an automatic machine translation was created to provide translation between Khmer and other 20 languages by using three statistical methods: the phrase-based approach, the hierarchical phrasebased approach and the operation sequence model (OSM) as well as selecting BLEU and RIBES to evaluate translation quality [5].

There is, furthermore, research specifically on Thai-Khmer machine translation. For example, Thai -Khmer machine translation on a website has been developed based on Java (JSP) and SQL (Appserv) with 4,000 words from a Thai -Khmer dictionary as a database [6]. In testing, 212 sample sentences have been processed, and the result has shown 72.16% of accuracy which is higher than that of Google translator. In 2014, the rule-based machine translation (RBMT) combined with statistical methods was recognized to be widely applied in automated translation [7]. The technique has shown the potential to improve translation between Thai and Khmer. Even though such classical technique is applied, the research has rarely a result with high performance.

3 III.

4 Background of thai to Khmer Translation

Sentences in Thai and Khmer language are similarly formed; on the other hand, ordering and semantic structure are different. With regard to the existent methods, the newly presented one for the proposed system is expected to balance between advantages and disadvantages of the classical techniques and be straightforward for implementation. In this paper, a process to translate Thai to Khmer language is composed of six main steps including 1) Input process: reading Thai text into the system from a website screen, 2) Word segmentation: applying Lex To and the longest matching approach to divide Thai sentences into words, 3) Word search: retrieving data from the database of Thai-Khmer dictionary to find a matched-meaning word in Khmer for each Thai word, 4) Boundary check: considering a boundary of each Thai word such as conjunction, verb, adjective and surrounding nouns to inspect parts of speech, 5) Pattern verification: examining Thai sentence patterns by using the rule-based algorithm, and 6) Khmer word rearrangement: reordering Khmer words in phrases or sentences.

To build a Thai-Khmer dictionary for testing in this paper, approximately 37,052 Thai words from the Royal Institute Dictionary (RID, 1999) are translated according to the existent Thai-Khmer dictionary [8]. In the process of examining word boundaries, patterns and various conditions of grammar rules are taken into account to solve translation mistakes.

The classification of machine translation architecture which is regularly implemented on clientweb server for online translation is the direct model shown in Figure 1. The direct machine translation architecture transforms a source language sentence (Thai) into a target language sentence (Khmer). Besides, the proposed system applies the indirect architecture which is demonstrated as a diagram in Figure **??**. A sample screen of the program is also provided Figures 3. IV.

5 Web Page

6 Methodology and Proposed Algorithm of Reordering

In general, Thai and Khmer sentences are sorted verbatim. Regarding to the verbatim characteristics of these two languages, the classical algorithm of word reordering could appear to be a proper tool to cope with phrase and sentence arrangement. On the other hand, the reordering method is unable to suit all cases of input phrases

and sentences since the reordering could cause translation mistakes. The analysis to deal with the issue has consequently become essential. The verification process is included in the proposed system to investigate errors; in addition, simple approaches to examine the previous word, the next word, and noun and verb positions are used to attain accurate outputs. Implementing pattern-based machine translation also alleviates the translation issue although the method is not novel. The technique based on patterns is also applied to assist translation due to its reputation of promoting translation performance. In this paper, patterns for the method are designed according to Thai and Khmer grammatical structures. The process to deal with the translation issue is arranged into four steps as follows.

1. Morphological analysis 2. Concept of pattern matching 3. Search for proposed patterns 4. Word rearrangement and translation

In the first step, LexTo software is applied to separate each word in a sentence from the others so that the morphology of an input sentence is analyzed. Next, positions of noun (n.), verb (v.), adverb (adv.), adjective (adj.), conjunction (con.) and interjection (int) are considered to acquire a concept of pattern matching with regard to SVO sentence structure. Sample sentences are demonstrated as follows. A simple sentence: Example:I eat rice.

A sample sentence with color term or color perception [9]: A sample sentence: The counter unit without subject (S): There are five eggs.

In the third step is to search for the proposed patterns in the input Thai sentence so that the output words are appropriately used in the Khmer sentence. Mapping between grammatical structures of Thai and Khmer language is undertaken. Khmer sentence patterns are then converted into proper forms through mapping algorithms.

To explain the mapping process, let Thwd{x} be a Thai word and Khwd{x} be a Khmer word where x is an index of a word in the sentence. Then the pattern mapping is defined as follows. : : :

7 Example1.1.1:

To verify a Thai word ”??” (k?on) = person, people, human, man If the word ”??” (k?on) is in a position following any other words in the sentence, the Khmer word ”?á??” ? ? ? ? ” (mÉ?”nuh) is replaced by ”á??”? ? ? ” (neak). The example sentences are provided below.

8 Example1.1.2:

To verify a Thai word ” ” (na?) = sit If the word ” ” (na?) is in a position before a noun (n.), the Khmer word ” kuy) is replaced by ”? ? ?” (cih). The example sentences are demonstrated as follows.

The algorithm is used in a case that a Thai phrase contains words (Thwd{3}) such as ??? ? , ??? , ???, ????, ??? ?, ??? ?, etc.

9 Example1.2.1:

To verify a Thai word ”??? ? ” (yuù) = is, am, are, was, were, be If the word ”??? ? ” (yuù) follows another Thai word ”?? ??? ?” (kam?la?) = ?ing , the Khmer word ”á??”? ” (n?v) is removed from the sentence. The example sentences are shown below.

The algorithm is implemented when a Thai phrase contains words (Thwd{3}) such as ?? ???, ?? ? , ??, ???, etc.

Example1.3.1: To verify a Thai word ”?? ???” (tham ?may) = why, for what If Thwd{3} = ” ?? ??? ” (tham ?may) is in a position after a verb (v.), Khwd{3-1} = ”?? ? ? ? ? ?? ” (haet Ê?”v?y) is replaced by Khwd{3-2} = ”? ? ?? ? ? ?? ” (tv?? Ê?”v?y). The example sentences are explained as follows.

10 Algorithm2.1:

IF(Thwd{3}==”number || noun”) THEN Khwd{3} and Khwd{4} is swapped position ELSE Khwd{3} and Khwd{4} is not swapped

The algorithm is for the case that a Thai phrase contains words (Thwd{4}) such as ??? and ??. ”???? ?? ? ?? ???????” (I have three brothers.) (p?om-mii-p?îî-nÉ?”?É?”?-sa?m-k?on), and Khmer: ?? ? ? ?á??”???? ?á??”?? ?á??”? ?? ? (k?om -mien -b???-pÊ?”oon -b?y -mÉ?”nuh) ?? ? ? ?á??”???? ?á??”?? ??á??” ? (k?om -mien -b???-pÊ?”oon -b?y -neak) Thai: ”???????? ? ????????? ????” (He will go to work by train.) (k?aw -ja?-na? -rot ?fay -pay -t?am ??aan), and Khmer:??á??” á??”? ???? ? ?????? ? ?????? ? ?? ? (koat -n?? -Ê?”??kuy -r??t pl??? -t?v -tv??- ??á??” á??”? ??? ?????? ? ?????? ? ?? ? (koat -n?? -cih -r??t pl??? -t?v -tv??-kaa) Algorithm1.2: IF((Thwd{1}=”?? ? || ?? || ?? ??? || ??? || ?? ??? || ???”) || (Thwd{2}=”?? ? || ?? || ?? ??? || ??? || ?? ??? || ???”) || ((Thwd{1}=”?? ??? ?” || Thwd{2} =”?? ??? ?”) && Thwd{4}!=”noun”)) THEN Thwd{3} is translated as Khwd{3-2} ELSE Thwd3 is translated as Khwd{3-1}

Thai: ”???? ???? ???à¸?”??????? ? ??? ?” (writing a letter sir.) (kam?la?-k?i?n-jot ?ma?y-yuù-k?ra?)Khmer: ?? ?? ???????? ?? ?????? ? (k?mpu?-s?? se?-s?mbot-n?v-baat) ?? ?? ???????? ?? ???? ? (k?mpu?-s?? se?-s?mbot-baat)

Algorithm1.3: IF(((Thwd{1}=”verb”) || (Thwd{2}=”verb”)) || ((Thwd{1}=”verb || noun || adv”) || (Thwd{2}=”verb || noun || adv”) && Thwd{4} !=”noun || verb”)) THEN Thwd{3} is translated as Khwd{3-2} ELSE Thwd{3} is translated as Khwd{3-1} Thai: ”?? à¸?”???? ? ?? ? ?? ???” (Why did you come here?)

(k?un-maa-t?iî ?niî-tham ?may)Khmer: ?? ????? ?á??"????? ?? ? ? (neak-mÉ?"É?"k-tii nih-haet Ê?"v?y) ?? ????? ?á??"???? ? ?? ? ? (neak-mÉ?"É?"k-tii nih-tv?? Ê?"v?y) B. Sample 2: [ThPattern 1] [KhPattern 12] Thwd{1} + Thwd{2} + Thwd{3} + ? + Thwd{ x n } Khwd{1}+Khwd{2}+Khwd{4}+Khwd{3}+Khwd{5} + ? + Khwd{ x n }

If the word "???" (moo?) follows a number (of time indications), the Khmer word "??? ?á??" á??"?? ? ?" (pram bu?n-mao?) is replaced by drow eht "?? ? ???? ?á??" á??"" or swap the position with that of the word "?? ? ?-????á??" á??"". In this case, the example sentences are provided below. Reordering words and translating are in the final step to diminish the translation issue. After the pattern each word which is then rearranged to be in a proper position. As a result, a Khmer sentence is attained as the output.

V.

11 Performance Evaluation

The proposed system is assessed for translation performance from Thai to Khmer by sentences from various sample documents as the input. In the testing process, the total phrases and sentences The algorithm is applied for a Thai phrase consisting of a word (Thwd{2}) ?? ?.

Thai: "?? ?????à¸?"? ???? ??????? ??? ??"

(iìk -sÉ?"?É?"?-d?an-c?a?-jaÊ?"-pay-kam.p?uu?c?aa) (Next two month I will go to Cambodia.) Khmer: ?????? ????? ? ? á??"? ?????? ? ? ? (ti?t-pii-k?ae-k?om-n??-t?v-kampuÊ?"cie) ?? ????????? ? ? á??"? ?????? ? ? ? (pii-k?ae-ti?t-k?om-n??-t?v-kampuÊ?"cie) mapping is completed, around 37,000 words from Thai -Khmer dictionary database are retrieved to match Table I: Sample of Phrase/Sentences for testing Three translation systems including Google translator [11], Chhun's translation system and the proposed system in this paper are assessed through translating the sample phrases and sentences. The translated outputs of each translation system are categorized into three groups consisting of accuracy (correct), acceptance (acceptable) and mistake (wrong). According to 5,100 Thai sentences selected for testing, the proposed system is able to translate 4,083 words correctly (80.06%), reach the acceptable level of translation for 469 sentences (9.189%) and produce errors only in 548 sentences (10.75%). The total translation accuracy of the proposed system becomes 89.25 % which is a sum of its accuracy and acceptance value. On the other hand, Chhun's translation contributes to 3,590 correct sentences (70.38%) which is less than those of the proposed system, 658 acceptable sentences (12.9%) and 857 mistakes (16.81%). Google translation also achieve less accuracy compared to the proposed system: 1,067 correct sentences (20.9%). whereas it acquires 798 acceptable sentences (15.64%) and 3,230 mistakes (63.34%), respectively, higher than those of the proposed one.

Moreover, performances of all systems are compared with regard to system precision, recall and efficiency by implementing F-measure as shown in Table **??** The result in Tables 2 reveals that the proposed system attains the highest score in all evaluations: the precision is 0.89, the recall is 0.80 and the efficiency (F-Measure) is 0.84.

12 VI.

13 Conclusion

The methodology in this paper is presented for creating Thai to Khmer machine translation system by using syntactic and semantic analysis to transform and structure patterns as well as implementing the rulebased translation. The presented processes can also simplify compound sentences into simple ones based on predefined sentence structures. The previous word, the next word and the subject-verb agreement are also considered. In addition, switching with more suitable words, reordering words and adjusting output sentences are also performed with regard to Thai and Khmer grammar. As a result, the proposed system is apparently able to improve the quality of source texts and translated outputs as well as assist Thai-Khmer language learners. Nevertheless, a larger amount of sample sentences in the corpus than that which is currently applied in the proposed system is necessary to achieve higher performance in Thai-Khmer translation. Furthermore, the larger dictionary database as well as the higher diversity of sample sources would be added to the process. Other methods or tools would also be considered to develop Thai-Khmer translation in future research.

VII. [1] [2]

[1] © 20 7 Global Journa ls Inc. (US)

[2] © 20 7 Global Journa ls Inc. (US) 1

23

Figure 1: Fig. 2 :Fig. 3 :

	Thai Phrase/ Sentence		Sentence Segmentation	Searching Word by Word
Year 2017 30 Volume XVII Issue III Version I) (H	Khmer Phrase /Sentence			
Global Journal of Computer Science and Technology	Thai-Khmer Pattern & Rule-Base	Thai -Khmer PHP Code (Dictionary) Lexical Base MySQL Web Se		

[Note: Fig. 1: Architecture of Thai-Khmer Translation on Web]

Figure 2: Pattern using Algorithm Grammars Patterns & Rules List Check Conj./Adj./ Verb/Noun Re-arranging sentence

This car is red.

Year 2017
31
)
(H

Thai sentence	Khmer sentence
S + V + O	S + V + O
?? ? + ?? ? + ?? ?? c?a? + kin+ k?aâw	= ?? ? ? + ?? ? +?? k?om + ?am + bay

Figure 3:

Example:
Thai sentence S + (V) + O
Khmer sentence S + (V) + O

?????? + ?? ??? ? + ?? ?à¸?"? = ???á??"? + ?á??"? + ??? ???? rot ?yon + k?an-nií+ si? ?d??? r

Example:
[ThPattern x]

Year 2017
Thai sentence S + V + O + unit(s)
Khmer sentence S + V + O + unit(s)

32
+ ?? + ??? + 5 + ??? = +?á??"+?? ? ?+??? +???á??" + mii + k?a ?y + haâ + fÉ?"É?"? + mier

Volume XVII
Note: ? = incorrect ordering words, ? = Correct A.

Issue III
Sample 1:
[ThPattern 1]
[KhPattern 11]
Thwd{1} + Thwd{2} + Thwd{3} + ? + Thwd{x n }

Ver-sion I
)
(H
Khwd{1}+Khwd{2}+(non, Khwd{3-1} or Khwd{3-2})

Global Journal of Com-puter Sci-ence and Tech-nol-ogy
[ThPattern 1]
[KhPattern 11] : : : [KhPattern 1n] [KhPattern 21] [KhPattern 22] [KhPatt

Translation Methods	Precision Recall		F-Measure
Google	0.57	0.21	0.31
Chhun	0.84	0.70	0.76
Proposed System	0.89	0.80	0.84

Figure 5:

.1 Acknowledgment

This research was funded by School of Information and Communication Technology, University of Phayao. We would like to show our appreciation to Cambodian students who have assisted the research by correcting Khmer words, phrases and sentences. We are also immensely grateful to other researchers for their supports in this research project.

.2 Global Journals Inc. (US) Guidelines Handbook 2017

www.GlobalJournals.org

[Vanavong et al. ()] , Y Vanavong , R Saravut , Thai-Khmer Dictionary . 2012. Cambodia: Nokothom Book Shop. p. .

[Thu et al. (2015)] 'A Large-scale Study of Statistical Machine Translation Methods for Khhmer Language'. Y K Thu , V Chea , A Finch , M Utiyama , E Sumita . *The 29th Pacific Asia Conference on Language, Information and Computation*, (Shanghai, China) October 30 -November 1, 2015. p. .

[Available ()] `http://translate.google.com/` *Available*, 2016.

[S ()] 'Color in Khmer : Perceprion and Grammatical Construction'. S . *Thesis, Silpakorn Univ* 2005. p. .

[Loy] *Fundamental Khmer 2*, P S Loy . p. . Ramkhamhaeng University

[Google web site, Google Translater] *Google web site, Google Translater*,

[Jabin et al. (2013)] 'How to Translate from English to Khmer using Moses'. S Jabin , S Samak , K Sokphyrum . *International Journal of Engineering Inventions* Sep.2013. p. .

[Sae-Tang and Prayote (2012)] 'Japanese-Thai Machine Translation with Generalized Patterns'. P Sae-Tang , A Prayote . *Computing and Convergence Technology (ICCCT), 7th International Conference*, (Seoul, Korea) Dec. 2012. IEEE. p. .

[Mangeot] 'MotàMot project: conversion of a French-Khmer publ-ished Dictionary for building a multilingual lexical system'. M Mangeot . `http://arxiv.org/abs/1405.5674` *Languag-es Resources and Evaluation Conference*, p. .

[Malézieux et al. ()] 'Rule-Based machine translation (RBMT) with statistical knowledge'. G Malézieux , A Bosc , V Berment . *Proceedings of the 5th Workshop on South and Southeast Asian NLP, 25th International Conference on Computational Linguistics*, (the 5th Workshop on South and Southeast Asian NLP, 25th International Conference on Computational LinguisticsDublin, Ireland) August 23-29, 2014. p. .

[Chancharoen et al. (1998)] 'Sentence-based machine translation for English-Thai'. K Chancharoen , N Tannin , B Sirinaovakul . *The 1998 IEEE Asia-Pacific Conference on Circuits and Systems, Chiangmai Plaza Hotel*, (Chiangmai, Thailand) Nov.1998. p. .

[Table II: F-Measure Results for Thai into Khmer] *Table II: F-Measure Results for Thai into Khmer*,

[Chhun ()] 'Thai-Khmer simple sentence translation on Web'. C Chhun . *the 1st RMUTL Chiangrai National Conference, (RCCON)*, (Thailand) 2015.