The First Wikipedia Questions and Factoid Answers Corpus in the Thai Language

Kanokorn Trakultaweekoon
National Electronics and Computer Technology Center (NECTEC)
National Science and Technology Development Agency (NSTDA)
Pathumthani, Thailand
kanokorn.tra@nectec.or.th

Santipong Thaiprayoon
National Electronics and Computer Technology Center (NECTEC)
National Science and Technology Development Agency (NSTDA)
Pathumthani, Thailand
santipong.tha@nectec.or.th

Pornpimon Palingoon
National Electronics and Computer Technology Center (NECTEC)
National Science and Technology Development Agency (NSTDA)
Pathumthani, Thailand
pornpimon.pal@nectec.or.th

Anocha Rugchatjaroen
National Electronics and Computer Technology Center (NECTEC)
National Science and Technology Development Agency (NSTDA)
Pathumthani, Thailand
anocha.rugchatjaroen@nectec.or.th

***Abstract*— This article introduces a Thai questions-answers corpus for a question-answering task which was extracted from Thai Wikipedia which was downloaded on 17 December 2017. The answers comprise 5,000 annotated factoids. The corresponding questions are exact phrases/sentences that contain the answer, but are replaced by a question word, or synthetic questions acquired from phrases and/or sentences on the wiki page. A question must contain only one of a set of 7 specific question words and a complex question must be avoided. Fifteen annotators used an annotation system specifically designed for this task. Acceptance, rejection, and revision processes were monitored by a language specialist. The final set was divided into 4,000 pairs for a training set and 1,000 pairs for a validation set. A baseline evaluation was conducted and an F1 score of 27.25 was obtained from document readers and 71.24 from document retrievals.**

***Keywords*— Thai questions-answers corpus, Thai Question-Answering system**

I. Introduction

A Thai Question-Answering (QA) system is a challenging task especially in the Thai language. Nowadays, English QA corpuses are available for research and benchmarking. In 2016, Microsoft research introduced WikiQA, an English QA corpus gathered from Bing, Microsoft search engine, and query logs. It contains 3,047 pairs of questions and answers based on user clicks on the Wikipedia page [1]. In 2017, Joshiy M. et al. announced TriviaQA which is the largest QA corpus which contains 650K of question-answer evidence triples. The evidence documents are the third important element which were collected and checked for redundancy from a Web search of results and Wikipedia pages [2]. Recently, a Google research team introduced "Natural Questions: a Benchmark for Question Answering Research", in which the questions are from queries issued to the Google search engine. In total, it contains 307, 373 training examples with single annotations. Its development set contains 7,830 examples from 5-way annotations, which is a technique of answer collection that used a nonnull answer that is seen at least once in the 5 annotations (see Section 5 of this paper for more details) and a further 7,842 examples of sequestered test data [3].

In Thai, asking for a meaning of a word from Thai Wikipedia is the easiest type of question-answering task in NLP, because the system can search for its result by searching through a list of page titles. However, searching for a specific answer on a page is another challenging NLP task. The traditional Thai information retrieval system used a well-structured knowledge graph. This stores knowledge at the roots of knowledge trees that have well-designed paths to the root [4]. To construct a tree, the researcher has to work with Thai texts which do not show syllables, words, phrases, or punctuation at the end of sentences. Hence, the construction has always been done manually, which caused difficulties to create an open-domain QA system in the Thai language.

A process of information extraction from both questions and answers has to be well defined before establishing a corpus. Since this work planned to use the information from Thai Wikipedia, the first definition of this work was "the answer must be an exact word or number found in TH-Wiki". Then the corresponding question has to be a combination of phrases found in the same paragraph.

A web-based annotation system has been established. It has two modes, annotator mode and linguistic administration mode. For annotator mode, it allows a user to search for a topic, create, edit, revise, and delete QA pairs, whilst administration mode allows one to correct, comment, accept, or reject pairs. The content of this paper is organized as follows: Section 2 explains the difficulties of working with the Thai language for a QA task, Section 3 shows the proposed system which includes all the corpus design and the annotation restrictions, Section 4 analyzes the collected QA pairs, Section 5 describes the implementation of a QA baseline system with its accuracies and Section 6 concludes this paper.

II. Related works and Problems

In the English QA system, an unstructured knowledge-based approach has been developed using a deep learning approach such as memory networks from Jason Weston and Sainbayar Sukhabaatar (2015), which uses an attention mechanism for the information retrieval (IR) process. However, an IR implementation for an unstructured-knowledge source requires a good annotated corpus for training. Recent years, there are a few numbers of QA corpuses created from Wikipedia contents, WikiQA [1], SQuAD [5] and TriviaQA [2] mentioned in section 1, they also used Wikipedia as one of their major resources. WikiQA

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

from Microsoft research influenced our construction of the very first Thai Wiki-QA set.

Research in Thai information retrieval started in 1996 when Asst. Prof. Somchai Prasitjutrakul and his student Paramin Jindavimonlert established a Thai text retrieval system using PAT trees [4]. It used a hashed indexing tree to store and retrieve text. However, for the Thai language, Thai Wikipedia has also been a favorite resource for Thai NLP researchers, but Thai QAs corpuses has not been established yet. This work then tagged factoids and formed their corresponding questions, then published them for research use.

A factoid can be a unit of a word or a phrase or a date or a number or even an equation. This work focused on a unit of a single word only, hence single words were extracted manually by 7 annotators. Although the Thai writing system does not show syllables, words, phrases, or punctuation at the end of sentences, this corpus does not provide any of them. Therefore, the information extraction process needs to work automatically without them.

III. Corpus creation

This work proposed a set of Thai QAs extracted from Thai Wikipedia. The answers are factoids which are single words extracted from the source. The corresponding question is assembled from words and phrases from the same page with an additional question word.

There are three main components in the corpus creation of this work. They are a resource text which is from Thai Wikipedia, annotators, and an annotation system. Thai Wikipedia, the resource, was downloaded on 12 December 2017. It contained 120,764 articles from 304,693 registered users, some of which were authors. The annotators were 15 native Thai speakers with different kinds of expertise. They were undergraduate students, post-graduate students, and computer scientists. Fourteen of them are female, and one of them is male. They were 25 years of age on average.

The annotation system was a web application written in Javascript, PHP, and it used MySQL as a database. Three tables in a structure of a relational database system stored information of: created questions, answers, a character count of the position where the answer begins, a character count of the position where the answer ends, related Wiki content ID, a section of answers in the Wiki content ID, and annotator details. The frontend consisted of two modes, which were for the annotators and for the administrators.

When an annotator login to the system, the interface provides a topic search tool for choosing a Wiki article, then the user chooses an article and reads it. Afterwards, the user can annotate an exact answer (as found in the text), then creates a corresponding question. There were 2 types of question which are an exact question and a modified question, respectively. An exact question is a question formed by the phrase or sentence that contains the answer, but replaces the answer words with a question word and a modified question is a question formed by multiple phrases in the text plus an additional question word for the answer. All QA pairs were controlled by a conductor, who was a specialist in charge of setting the scope of the different questions and responsible for issuing annotation guidelines. The guidelines contain a preliminary scope of the Thai questions answers system. It was created by an experienced Thai semantics expert who is called a conductor. This corpus has been involved with only basic questions at present, therefore a question has to be appropriate in order to provide an answer. All constraints have to be carefully set. They are shown in the next few paragraphs.

The details of the annotation guidelines are divided into three parts: the definition of a question in this work, the constraints for question creation, and the language restrictions. This work defines the meaning of a question as "The simplest question one can ask for an annotated answer formed using phrases and words on the same page plus an additional question word." The interrogative words used are

- "อะไร" ('what')
- "ใคร" ('who')
- "ไหน" ('which or where')
- "เมื่อไหร่" ('when')
- "ใด" ('which or where')
- "กี่" ('how much/many or when')
- "เท่าไร" ('how much/many')

This corpus was created under 7 constraints when forming a question which are:

(1) It must not form a complex question or use "why" or "how" as the question word.

(2) It must use formal language with a formal question structure which means it must contain a subject + verb + object.

(3) Questions and answers must be semantically clear.

(4) Every question must contain a question word from the defined set only.

(5) A question must contain symbols, e.g. ":", ",", ";", which they must be used precisely as appears in the text, not from annotator insertion.

(6) The question length must be less than 2 lines.

(7) Typos from Wiki should be avoided in questions and answers.

In addition, there is a simple language restriction, which is that all questions must be polite, although they can be in informal language.

IV. Corpus characteristics

The corpus contains 5,000 QA pairs extracted from 2,923 Thai Wikipedia articles out of 120,764 which were downloaded. Table 4.1 shows the numbers of QA pairs separated by the two types of question as described in the previous section.

TABLE 1. Number of QA pairs in Training and Validation sets

Set of data	QA pairs	Number of Question type	
		Exact	Modified
Train	4,000	573	3,427
Validation	1,000	147	853

There was no control for balancing the numbers of each question type, because it was not easy to find a well-structured sentence or phrase for forming an “Exact” type of question. Therefore, the figures show that the annotators synthesized proper questions rather than editing a sentence/phrase.

The average position of answers on a page is 22.06% of page, which was calculated by finding the average location of all the answers then converting them to percentages based on overall page lengths, which we call a normalized location in this paper. Moreover, answer lengths are about 12.56 characters on average. The distributions of answers, normalized locations and lengths are shown in Fig.1.

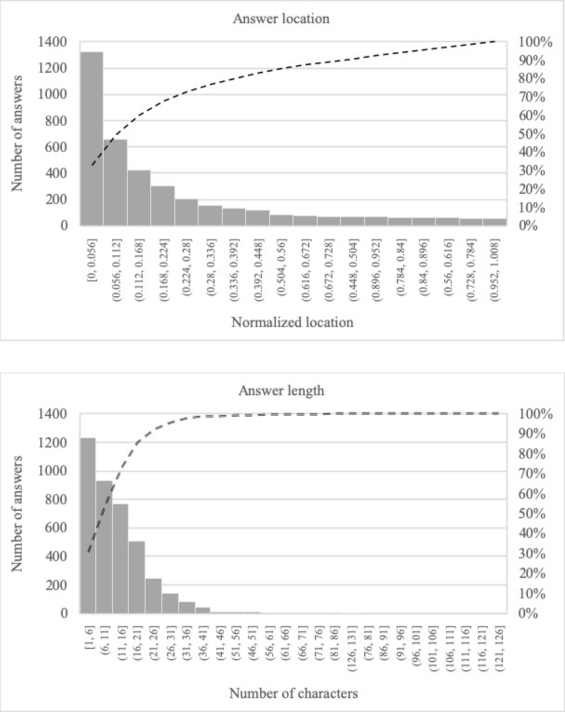

Fig. 1. Histograms of answer locations (top) and answer lengths (bottom). Dashed lines represent accumulative percentages of numbers in each graph.

The statistics for the use of each interrogative word are shown in Table 2. Some of the question words have common spelling variations, but they have the same meaning, so they are grouped in the table.

TABLE 2. NUMBER OF QA PAIRS SEPARATED BY AN INTERROGATIVE WORD.

Interrogative word		% of usage in corpus
Defined	Variation	
อะไร *(what)*	อะไร	31.10
	ว่าอย่างไร	0.04
Interrogative word		% of usage in corpus
Defined	Variation	
ใคร *(who)*	ใคร	16.11
ไหน *(which or where)*	ไหน	0.58
เมื่อไร *(when)*	เมื่อไหร่	n/a
	เมื่อไร	0.52
	เมื่อใด	0.92
กี่ *(how much/many or when)*	กี่	4.83
เท่าไร *(how much/many)*	เท่าไหร่	0.12
	เท่าไร	8.13
	เท่าใด	1.04
ใด *(which or where)*	ใด	36.65

A total of 5,000 factoids were annotated from the pages. The answers corresponding to the questions tended to be straightforward. However, this corpus contains noise data. Seven questions are noise. They contain typos or no question word, or neither. The questions with feigned noise are listed in Table 3.

TABLE 2. SEVEN NOISE QUESTIONS IN THE CORPUS.

Question	Error type(s)
ซิริล เฮย์คอค เกิดในตระกูลที่มีบิดาสืบเชื้อสายมาจากขุนนางเก่าแก่ บิดามีชื่อเรียกว่าอะไ Wha was Ceril Heycock's father name who was born under old hierarchies of nobility?	Spelling mistake
เอ ศุภชัย ได้ชักชวน เจมส์ มาร์ เข้าสู่วงการ ด้วยการเจอกันที่ไต Whare did A Supphachai persuate James Mars to join his angency?	Spelling mistake
นุติ เขมะโยธิน เกิดเมื่อวันที่ ~~When~~ was Nuti Khemayothin born?	NoQword
ชลาศัย ขวัญฐิติ หรือหม่อมลูกปลา เกิดเมื่อวันที่ ~~What~~ was Chalasai Kwanthiti, called Mhom Lookpla, birth date?	NoQword
ในปี 2012 เด็กอายุต่ำกว่า 15 ปี ที่ได้รับการวินิจฉัยว่า In 2012, Children who was under the age of 15 and diagnosed as having .... disease	NoQword + Spelling mistake
นวนิยายแนวลึกลับและสืบสวนชื่อเรื่องว่า รหัสลับดาวินชี ประพันธ ~~Who~~ wrot the Da Vinci Code, a mystery thriller novel?	NoQword + Spelling mistake
Question	Error type(s)
ใตรเป็นผู้ชนะเลิศการแข่งขันรายการเอเชียเน็กซต์ท็อปโมเดล ฤดูกาลที่ 3 Wno was the winner of Asian Next Top Model competition in season 3?	Spelling mistake

V. BASELINE ACCURACY

This research evaluated the corpus using a baseline system. The test flow is as shown in Fig. 2

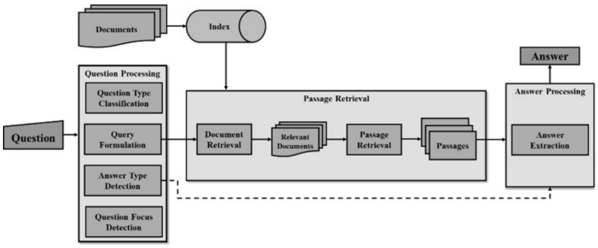

Fig. 2. Baseline system used for finding a preliminary accuracy one can implement using the proposed corpus

The objective of a QA system is to automatically generate a correct answer to a question by finding the relevant documents containing the answer strings. The baseline system consists of three main components: (1) question processing (2) passage retrieval, and (3) answer processing.

Question processing is the first process as shown on the left of Fig.2. It detects a question type, a question answer type, and a question focus word/phrase, then uses the components as query constraints. The second process in the middle of the figure is the passage retrieval. It aims to retrieve a set of text passages that might contain answer strings, so it tries to match a query with documents in the corpus and retrieves/selects only the top 5 passages which have the highest matched scores. Then the pas-sages are passed to the answer processing as shown on the right of Figure 5.1, which extract an answer by parsing all the passages to a name-entity tagger using a pattern extraction approach. Finally, the system re-turns an answer that matches the answer type detected at the beginning of the process.

The corpus was evaluated using the baseline system. The experimental setting was conducted for two tasks: (1) document retrieval task and (2) document reader task. For the task of document retrieval, we used an F-measure that is commonly used to evaluate performance in information retrieval. The experimental result shows that our technique for the document retrieval task achieved 71.24% at 1-best accuracy. For the task of the document reader, we use the exact match (EM) metric that computes common substrings at word level between the predicted answers and the gold answers. We can achieve 25.92% of F1 scores at 1-best accuracy.

VI. Summary

This paper presents a Thai Wikipedia QA Corpus whose answers are factoids extracted from Thai Wikipedia articles. There are 2 types of questions, which are Exact or Modified. Exact is a type of question that is formed by replacing the answer with a question word. Modified is a type that is synthesized from phrases around the answer plus a question word. In total, the corpus contains 4,000 training QA pairs, and 1,000 validation pairs. It should be noted that they can be mixed together and divided in different proportions for use as a training set, a validation set, or a testing set. A baseline system gives a preliminary baseline performance which can be applied to the proposed corpus. The experiment was conducted using another set of sequestered test data. The F1 score achieved 25.92% at 1-best accuracy. The corpus is free for research use at http://aiforthai.in.th/corpus.

Acknowledgment

The authors would like to thank all Thai Wikipedia authors and the Wikimedia Foundation, Inc. for supporting the Wikipedia platform in many languages.

References

[1] G Y. Yang, W.-t. Yih and C. Meek, "WikiQA: A challenge dataset for open-domain question answering," in Conference on Empirical Methods in Natural Language Processing, Lisbon, 2015.

[2] M. Joshi, E. Choi, D. Weld and L. Zettlemoyer, "Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension," in Proceedings of the 55th Annual Meeting of the Association for Compuational Linguistics (Volume 1: Long Papers), Vancouver, 207.

[3] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, M. Kelcey, J. Devlin, K. Lee, K. N. Toutanova and L. J. a. M, "Natural Questions: a Benchmark for Question Answering Research," Transactions of the Association of Computational Linguistics, 2019.

[4] P. Jindavimonlert, "A Thai Text Retrieval System using the PAT tree: M.Sc. Thesis," Department of Computer Engineering Chulalongkorn University, 1996, 1996.

[5] P. Rajpurkar, J. Zhang, K. Lopyrev and P. Liang, "SQuAD: 100,000+ Questions for Machine Comprehension of Text," in Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, 2016.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy