WALAILAK JOURNAL
http://wjst.wu.ac.th Information Technology

Improving Answer Retrieval from Web Forums with Topic Model and Ontology

Kanda Runapongsa SAIKAEW[1,*], Seksan POLTREE[1], Kornchawal CHAIPAH[1] and Choochart HARUECAIYASAK[2]

[1]*Department of Computer Engineering, Faculty of Engineering, Khon Kaen University, Khon Kaen 40002, Thailand*
[2]*National Electronics and Computer Technology Center, Pathum Thani 12120, Thailand*

([*]Corresponding author's e-mail: krunapon@kku.ac.th)

Received: 30 October 2014, Revised: 11 March 2015, Accepted: 28 April 2015

Abstract

Searching for information online has become an essential part of modern living. For a general domain, one can use a tool such as a search engine to find information. However, for domain-specific questions, other means, such as a web forum, are preferable. Common problems with web forums are post duplication and poor search results for long query strings, such as sentences. To overcome these issues, we propose a system based on a language model and ontology. Given a question, the system performs language processing to analyze the syntactic structure of the sentence and its category. The question sentence is classified using regular expressions, keyword matching, and query templates. Based on knowledge constructed from the existing information in the web forum, we can retrieve the suggested links to the answers for the given question. To the best of our knowledge, our paper is the first article that attempts to understand questions and to suggest existing sources of answers in a Thai web forum. We compared our 2 proposed subsystems, language-model-based and ontology-based, with the Google custom web search engine. We evaluated the systems by using a Thai breastfeeding web forum containing 6,823 threads with 75,906 messages. The evaluation results show that the proposed systems have fewer duplicated suggestions compared with the Google search engine. Moreover, for the input with some ambiguous keywords, our proposed system, based on a language model, outperforms the Google search engine, because the system based on a language model is better at finding related terms. For a question with no frequently found keywords, our proposed system, based on ontology, suggests answers which are more appropriate than answers from the Google search engine, because it contains related knowledge defined by experts.

Keywords: Latent dirichlet allocation, ontology construction, question answering system, natural language processing

Introduction

Searching for information online has become an essential task that most people perform daily. Using search engines based on keywords is a popular method of finding information. However, some people may not be familiar with how to search using keywords. A more natural method is "Question and Answer" (Q&A), which has been in use for many years. Examples of such applications are Google Guru and Yahoo Answers. In addition, a web forum is a common Q&A platform for obtaining knowledge and information. However, common problems encountered when using web forums are post duplication and poor results from searches based on sentences. To solve these issues, questions must be studied and

analyzed, by addressing matters such as relevancy, question complexity, question domains, question ambiguity, and answer quality.

An example of a web forum is one for Thai parents who intend to breastfeed their children. There are many obstacles that can occur. The parents want to receive advice as fast as possible. They often type questions without searching for answers that may already exist in the forum. Therefore, many questions are repeated, and the web forum data is increased unnecessarily. In addition, users need to wait for the responses of newly-asked questions instead of finding the answers to questions that have been previously posted and answered by others.

Suggesting answers for the new question before it is posted will reduce the amount of time that web forum users need to wait for the answers, and also decrease the amount of data posted to the forum. Currently, in a Thai web forum, there is no such tool that suggests answers for a question to be posted. Although the web forum admin can install Google search for the forum, there are still some types of questions that Google search cannot answer well.

Although there have been several research works about Natural Language Processing (NLP) for Thai documents, there are still many open questions remaining. The questions that are in our focus are how to find the similarities between a new question and existing information, and how to rank the quality of several questions as responses to the given question.

In this paper, we propose an algorithm for searching a web forum that automatically suggests answers to a question that a user asks. Our goal is to reduce topic duplication and improve information searching. NLP is applied to understand existing posts and the input question. The system preprocesses posts by using information retrieval techniques to index transformed text, in order to determine the relevancy and similarity between the question and the existing posts. Ontology is created to collect specific-domain knowledge, and system answers are displayed to users as possible answers.

To the best of our knowledge, this is the first paper that presents an approach to understand questions and to suggest existing sources of answers in a Thai web forum. In addition, the process of ontology creation to collect knowledge and to suggest answers for a given question can be applied to question answering systems in other languages. Such processes consist of 1) question type classification using a regular expression parser, 2) ontology term selection using word segmentation programs and WordNet, 3) ontology term meaning ambiguity resolved by using WordNet, and 4) ontology mapping extraction using a regular expression parser. In our work, we assume that the existing questions are the post topic. We do not consider the content of the post topic to be the existing questions that are used to compare with a question that a user is about to ask.

Next, we discuss related work in web knowledge mining and question answering systems. We then present interesting techniques in applying semantic web in answering questions.

Web knowledge mining

Inui *et al*. [1] created a database that stored information about personal experiences and opinions using data from user-generated content (UGCs), such as personal web blogs and posts. However, their database could not be used to imply any answer for a given question, whereas the database presented in this paper can perform such a task.

Ting *et al*. [2] proposed the architecture of a social recommendation system based on the data from microblogs. From the analysis results, they found significant differences in the Social Network Analysis (SNA) measurement between different products. Although both their work and our work analyzed the existing real data, they proposed a system architecture to recommend different products to target customers, whereas we implemented a system to suggest the answers to the input questions.

Question answering systems

Chen and Wen [3] introduced a question answering service using NLP. The system consisted of 3 modules. The first module was a question analyzer. When a user entered a question into the system, the question was then tagged as part of a speech using a NLP corpus. The second module was a dialog theme recognition that determined the domain of the entered question. The third module was a semantic recognition and data extractor that tagged a word as a verb, noun, or adverb. Those tags were then passed

into a semantic recognition formula. After that, the query was sent to a web service to generate a message based on predefined human-readable response templates using NLP semantically.

Al-Rajebah and Al-Khalifa [4] presented a solution to extract ontologies from Arabic Wikipedia, using a linguistic approach based on the semantic field theory introduced by Jost Trier. The system consisted of 3 main phases: (1) filtration, (2) extraction, and (3) ontology generation. They conducted 3 empirical experiments to evaluate the results of the proposed system, and achieved an average precision of 65 %. They suggested several enhancements to improve the precision by existing linguistic relation such as WordNet.

Shaheen and Ezzeldin [5] proposed a survey paper about different Arabic question answering systems, resources, tools, and future trends. Regular NLP systems cannot handle Arabic language because of its richness, such as high derivations, high inflections, high ambiguity, and difficult named entity recognition. They reviewed the 3 main subtasks of Arabic QA (question analysis, passage retrieval, and answer extraction). They made an analysis comparison between Arabic QA systems and the state-of-the-art English QA systems by discussing the approaches, presenting the performance metrics, and pointing out some drawbacks. They suggested that Arabic QA researchers should investigate more on inference-based and logic-based approaches, as well as semantics-based approaches. They suggested several enhancements to improve the precision of the proposed system, such as extracting more semantic relations and adding hyperonym or synonymy relations.

Chen and Chiu [6] used sequential logic regression and structural equations that analyzed user messages based on content, social clues, and personal information. This related work used the message type concept to categorize question type and to suggest the answer if messages to be posted were based on earlier messages. Our approach also categorizes user messages into types, but the information is retrieved from the web forum posts instead of from a search of external resources.

Semantic web in question answering

Some systems have used semantic web technology for question answering. Lamberti *et al*. [7] proposed a relational-based ranking strategy in conjunction with an existing semantic web search. The algorithm employed annotation and underlying ontology within a web page to create a relevancy score for that page based on a user query. The prototype system applied a sample travel ontology written in the Web Ontology Language (OWL). Keywords were entered by the user, who manually selected a keyword concept class from a hierarchical pull-down menu. After that, a graph-based algorithm was applied to create a user query sub-graph. The relevancy score of a potential answer page was computed from the number of edges in that sub-graph. Finally, candidate pages were reordered by their relevancy scores and were displayed to the users based on their rankings.

A semantic web for question answering has been applied to Thai language. Kongthon *et al*. [8] introduced a method for mapping Thai natural language to the SPARQL Protocol and RDF Query Language (SPARQL). The tourism domain was used as a case study. They created a “tour” ontology and constructed a query pattern to map a question to a SPARQL query. The results showed that these input query patterns could be mapped to SPARQL, and that the mapping queries could return the query results to the user. However, in some cases, the system could not extract a query and derive an answer. Moreover, no systematic evaluation was conducted.

Another system that is closely related to our proposed work is QAST (Question Answering System for Thai Wikipedia) [9]. However, there are some differences between the two. QAST is considered a data-intensive Q&A system, in that it focuses on extracting answers from both unstructured and structured data. Our system, however, also suggests answers derived from a knowledge base. QAST does not use a language model and ontology, while our proposed work uses these features to improve the performance of the system.

Suktarachan *et al*. [10] proposed the development of a question answering services system for farmers through SMS query analysis. The similarity between this related work and our proposed work is that both works apply semantic web and natural processing techniques to attempt to tackle 2 Q&A analysis problems: 1) interpretation of a question word, and 2) answer identification. The main difference

is that their system has not been tested and evaluated, while our proposed system has been evaluated and compared against existing search engines.

Automatic ontology learning

The goal of automatic ontology learning is to support the semi-automatic construction of ontologies starting from available digital resources (e.g., a corpus, web pages, dictionaries, semi-structured and structured sources) in order to reduce the time and effort in the ontology development process. Gharib *et al*. [11] proposed an enhanced methodology for enriching WordNet. Their experimental results showed that using Coarse-Grained word senses provides higher precision than Fine-Grained word senses in the Word Sense Disambiguation task. Our proposed system does not focus on enriching WordNet, but uses it to process the suggested answers for a given question.

Materials and methods

In this section, we describe the proposed system for suggesting effective input questions and for analyzing web forum input data and structure. Later, we discuss question preprocessing, question type classification, and a content-indexed database for suggestions using Latent Dirichlet Allocation (LDA), a generative probabilistic model for automatically discovering topics in a collection of data. We also consider how to apply an ontology knowledge database and SPARQL to give answers to an input query.

System design overview

The proposed system is composed of 2 major parts, as illustrated in **Figure 1**. The first part is question pre-processing, which aims to prepare the question for which the answers are to be found. The second part consists of 2 question-answering subsystems.

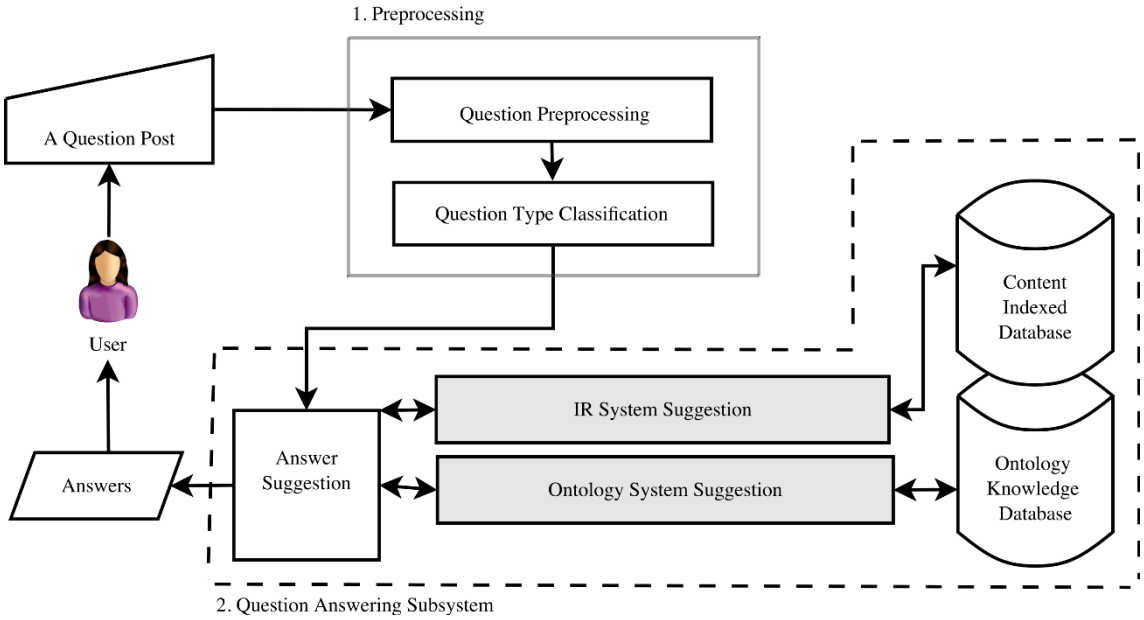

Figure 1 System overview.

Briefly, the first subsystem is an LDA-based subsystem. A content-ranking database is created using indexes on existing forum posts. The answers are computed from this content-indexed database

subsystem. The second subsystem is an ontology subsystem based on Frequently Asked Questions (FAQs). The input question is classified using a message type classifier, and the answers will be processed from an ontology knowledge database. The suggested answers will be put into templates and then displayed to users.

Preprocessing

In this section, we explain how the system selects, analyzes, and processes input data.

Analyzing web forum input data and structure

In this work, we access the database of a Thai breastfeeding web forum (http://www.thaibreastfeeding.org) as a case study. The domain of the data is about how to help parents successfully breastfeed their babies, and also gives advice about other topics related to raising babies and children. The web site consists of a web forum, articles, and a collection of FAQs. There are 6,823 threads that contain 75,906 messages posted to the web forum. We assume that this first post in a thread is the question or topic that the user is talking about. Due to time and space limitations, to give suggestions, we index and process only the first post of each thread. There are some questions that users frequently post to the forum about particular topics. These FAQs are the short articles manually extracted from the forum or the related content in the posts.

Question preprocessing

The system processing uses the word tokens as input data. Three segmentation programs, SWATH [12], LibThai [13], and Thai Word Segmentation web service [14] are tested to segment the posts to find out the possible tokens. SWATH and LibThai apply the same maximal matching segmentation algorithm, but employ different built-in dictionaries and implementations. We have used the BEST corpus [15] to evaluate these segmentation programs in the terms of precision, recall, and f-score. It is found out that SWATH performed the best. Thus, we use SWATH to segment all first posts for each thread in the web forum. The segmentation results in 62,155 words, and the occurrences of each word are counted in the posts.

In word segmentation, stop words do not express the meanings of a sentence. Thus, these words are removed to reduce the processing time. Using regular expressions may be inaccurate, especially in Thai, because a word can have more than one sense or meaning. Thus, an ORCHID part-of-speech tagged corpus using a Natural Language Toolkit (NLTK) is applied for tagging functions. The parts of speech are tagged by an N-Gram, a train model which looks for N words around that position. The system starts with trigram tagging and back off by using bigram and unigram, respectively. To solve the problem of unknown words that occur in about 40 % of all words, only words that can only be nouns or verbs from Thai WordNet are extracted. These word lists are used as back of tagging if there are unknown words from the N-gram tagger. Using Thai WordNet and N-gram tagger decreases the number of unknown words to around 10 - 20 % of all words.

Question answering system

The question answering system is composed of 2 subsystems. Each subsystem prepares the databases for preprocessed input questions. We first discuss the question answering system based on LDA and the one based on ontology.

Content-indexed database and question answering subsystem using LDA

In the first question answering subsystem, a language model is used to create an index corpus for similar queries. To analyze user input, the corpus query data is created. The main purpose for this is to find the most similar topics between the existing data and a new question. Our assumption is that the first post in each forum thread represents a question and a topic for that thread. **Figure 2** illustrates the process summary using an LDA language model to transform posts in forum threads. As shown in this figure, there are 2 types of input: the first posts in each thread, and the new question. Both types of input need to

be preprocessed and computed using LDA models. Then, the similarity between the existing first posts and the new question is computed.

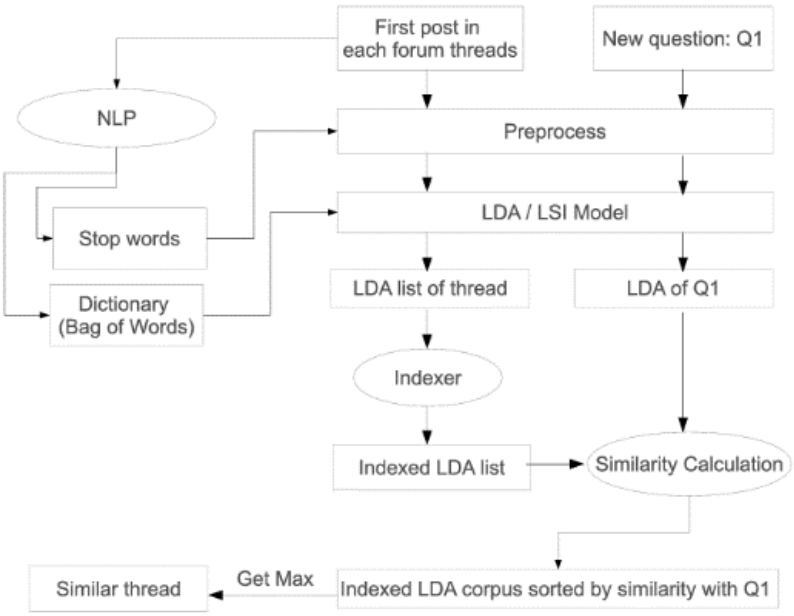

Figure 2 Information retrieval process summary using a language model.

Creating a dictionary corpus

It is necessary to create a domain-specific dictionary corpus and a bag of words to transform each topic into a model. The topics are segmented and then counted for topic frequency. Out of 6,823 forum threads, about 43,499 words are found. The word occurring most frequently appears 48,843 times. To reduce the number of words in our dictionary, the words that are found only once are removed. Finally, 22,606 unique tokens are left in the bag of words dictionary. Using Gensim framework, a dictionary corpus is compiled to use with this specific domain.

Preparing a corpus of first post messages in threads

After creating a dictionary, the corpus is transformed to word vectors. Next, in the preprocess part of each topic, the first post message is segmented and stop words are removed.

Transforming the thread corpus into an LDA model, and indexing

Each topic of total 200 segmented topics is transformed into a word vector. Then, the obtained vectors are combined as a topic corpus. This corpus is transformed into an LDA model using a transform function in Gensim. **Figure 3** expresses the output of transformed topics from the corpus. In this figure, we randomly choose topics to consider the words that are most frequently found. In many topics, the word "นม" (mean in English is milk) is one of the top 5 most frequently found.

Topic # 10		Topic # 60		Topic # 80		Topic # 117	
Terms	Prob.	Terms	Prob.	Terms	Prob.	Terms	Prob.
แช่(freeze)	0.060	นมแม่(mother milk)	0.048	แพ้(allergic)	0.047	นม(milk)	0.050
ช่อง(compartment)	0.048	นม(milk)	0.044	อาหาร(food)	0.041	ปั๊ม(pump)	0.042
ละลาย(melt)	0.043	เลิก(wean)	0.029	ชนิด(type)	0.039	ลูก(child)	0.024
นม(milk)	0.033	กิน(eat)	0.027	โปรตีน(protein)	0.030	เต้า(breast)	0.022
ตู้เย็น(refrigerator)	0.028	น้อง(child)	0.017	เนื้อสัตว์(meat)	0.015	ดูด(suck)	0.017
ตู้(cabinet)	0.026	ครบ(complete)	0.017			เดือน(month)	0.016
แช่แข็ง(freeze)	0.023	เดือน(month)	0.016			กิน(eat)	0.016
เก็บ(keep)	0.019	ลูก(child)	0.015			ตอนนี้(now)	0.014

Figure 3 Example of topic transformation using an LDA algorithm.

Calculating query similarity

The similarity of LDA topics is calculated by comparing matrix similarity between each transform topic calculation and the incoming new query topic. The similarity value is a vector multiplication calculation using the dot product of the matrix. The number at row *x* column *y* in the matrix represents the word ID and its transformed value. Then, we sort the list of matrix similarity values.

Next, we introduce the second question answering subsystem, based on an ontology knowledge database.

Ontology knowledge database and question answering subsystem

A semantic web is a large and open technique to suggest answers for a given question. We apply ontology to create a meaningful database for the second question answering subsystem. Next, we describe how to apply the ontology in the proposed system.

Question type classification

It is necessary to know the input question type in order to suggest answers to users. Regular expressions for each question type are used, as shown in **Table 1**. We assume that there is only one keyword for each input question type because most queries have at most one keyword question type. Some queries do not even have any keyword question type.

Table 1 Regular expression rules to determine input query types.

Type	Regex
Yes/no	(.*)(ไหม|มั้ย|หรือยัง|หรือเปล่า|รึเปล่า|รึไม่|รียัง|หรือไม่|ใช่ไหม|ใช่มั้ย)
What	(.*)(อะไร)(.*)
Where	(.*)(ที่ไหน)(.*)
When	(เมื่อไหร่|เมื่อไร|เมื่อใด|ตอนไหน)
Which	(.*)(แบบไหน|อย่างไหน|อันไหน)(.*)
How	(.*)(อย่างไร|ยังไง|ทำงัย|งัย)(.*)

Ontology development

Ontology is a conceptualization of things. A Resource Description Framework (RDF) is used to represent the linked data as an RDF graph. Ontology expresses the meaning of the RDF data. The standard representation of ontology is an OWL language. RDF/XML format is used to represent the data. In this work, FAQ content is chosen to represent specific domain information.

Ontology term selection

There are 56 FAQ topics. First, SWATH, a Thai word segmentation program, is used to segment each topic. The part-of-speech tagging is applied by using ORCHID and Thai WordNet words. The occurrence of the words is then counted for each part of speech.

Ontology class and data type properties are usually concepts from nouns, while object property or relation is usually a verb concept. Note that a noun or a verb may be an individual or an instance of a class, rather than being an entire class. It depends on what users want to conceptualize or how they wish to express their data.

Ontology term meaning ambiguity

There is much word ambiguity in Thai language. WordNet synset is used to decide whether a given word should be a class or an individual in the ontology.

WordNet has hyperonyms and hyponyms, in which hyponyms are a subclass of hyperonyms. Meronyms and holonyms are related in the terms of parts of relation. WordNet is employed to select the terms for class and its relation to the ontology. This initial ontology is loaded into the system using RDFlib framework.

Extracting ontology individuals and mapping

Knowledge information consists of ontology individuals and their property relations. Regular expression rules are created to extract the individuals from the FAQs section of the Thai breastfeeding website. An NLTK framework function called RegexParser is used to chunk the tagged text [14]. **Figure 4** expresses 3 regular expression patterns used to extract the triples from the FAQs.

Regex 1: {(⟨NNN⟩|⟨NCMN⟩|⟨NPRP⟩)+(⟨PREL⟩|⟨JSBR⟩)*((⟨XVBM⟩)*⟨NEG⟩(⟨XVAM⟩)*)* (⟨VVV⟩|⟨VACT⟩|⟨VSTA⟩)+⟨XVAE⟩*(⟨NNN⟩|⟨NCMN⟩|⟨NPRP⟩)+(⟨VATT⟩)*}
Regex 2: {(⟨NNN⟩|⟨NCMN⟩|⟨NPRP⟩)+(⟨UNK⟩)*(⟨RPRE⟩)+(PDMN)*(⟨UNK⟩)*((⟨NNN⟩|⟨NCMN⟩|⟨NPRP⟩)+|(⟨FIXN⟩)*(⟨VVV⟩|⟨VACT⟩|⟨VSTA⟩)+)}
Regex 3: {((⟨NNN⟩|⟨NCMN⟩|⟨NPRP⟩)+(⟨UNK⟩)*(⟨FIXN⟩)*(⟨VVV⟩|⟨VACT⟩|⟨VSTA⟩)*)(⟨UNK⟩)*(⟨PUNC⟩*(⟨NLBL⟩|⟨NCNM⟩)+⟨PUNC⟩*)+(⟨UNK⟩)*(⟨CMTR⟩|⟨CLTV⟩|⟨CNIT⟩|⟨CFQC⟩|⟨CVBL⟩)}

Figure 4 Regular expression parser rules to extract individuals from FAQs.

Information retrieval technique is employed to find the most relevant topic to a user. The technique uses the probability and statistical calculation in terms of a matrix. It can find the relevant results, but it may not have semantic knowledge. Thus, SPARQL queries are employed to extract a question from an input query.

SPARQL template

Regular expression rules are created to extract the triple. **Figure 5** shows the regular expression parser rules that are used to extract a question from an input query.

Regex 4: {(⟨NNN⟩|⟨NCMN⟩|⟨NPRP⟩|((⟨FIXN⟩)⟨VVV⟩|⟨VACT⟩|⟨VSTA⟩))+(⟨PREL⟩|⟨JSBR⟩)*((⟨XVBM⟩)*⟨NEG⟩(⟨XVAM⟩)*)*(⟨VVV⟩|⟨VACT⟩|⟨VSTA⟩)+⟨XVAE⟩*(⟨QUES⟩|⟨PNTR⟩|⟨QYN⟩)+(⟨VATT⟩)*}

Figure 5 Regular expression parser rules to extract a question from an input query.

At this point, we have explained how the system suggests the answers for an input question in a given domain. Next, we evaluate the proposed system to find out the precision of suggested answers.

Results and discussion

In this section, first we discuss the evaluation method. Then, we explain how random questions were selected for a user survey. Later, we present about evaluation parameters and analyze evaluation results.

Evaluation method

Since the Google web search engine has 10 suggestions per page, we set the same number for the question answering system based on LDA. However, the question answering system based on ontology has a number of suggested answers that depend on the availability of the related knowledge. **Figure 6** represents a sample of the system input/output interface. The output consists of LDA results listed on the left side and ontology outputs appearing on the right side. On the left-hand side, the system selects the 10 LDA topics most similar to the input query. On the right-hand side, the system selects the first 10 outputs from the ontology-based question answering subsystem. If there is no existing knowledge in the ontology, nothing is displayed.

We created an online survey to collect the results from users. Nineteen users who were active participants of the Thai breastfeeding web forum, from individual internet sessions and with unique IP addresses, filled out the 29-question survey. The survey listed the questions, as well as the suggested answers from the 3 systems.

- **Google** (the Google web search engine) is chosen as a comparative evaluation tool, because it is widely considered to be the top search engine for suggesting answers to a given question.
- **LDA** is our question answering subsystem based on LDA.
- **Ontology** is our question answering subsystem based on ontology knowledge.

After the users were presented with the questions, they marked whether each suggested answer was related to the given question. We invited those Thai breastfeeding mothers who were experts in the given domain to answer the questions based on http://www.thaibreastfeeding.org forum in the survey.

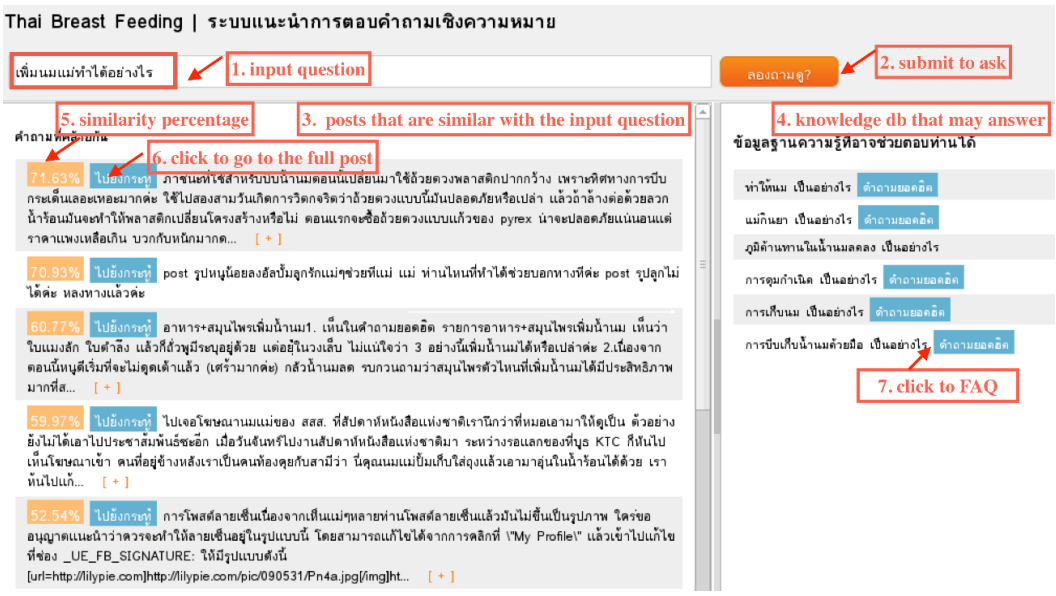

Figure 6 System user interface and example output.

Random question selection for the survey

We randomly selected questions from the subject in the first post, using statistical information based on the number of words in the post. We employed a straight sampling method to randomly choose the questions for the survey. There were 6,823 existing questions. From the statistics of the subject, the median of these questions was 11 words and the mean was 10.8. We thus used these median and mean values to define the questions that were short (1 - 8 words), medium (9 - 12 words), and long (13 - 21 words). There were 2,050 short questions (about 30 % of the total questions), 3,143 medium questions (about 46 % of the total questions), and 1,630 long questions (about 24 % of total questions). We randomly chose the number of sampled questions in each group proportionally with the number of total questions in each group. We randomly selected 9 out of total 2,050 short questions, 13 out of 3,143 medium questions, and 7 out of 1,630 long questions.

Table 2 lists the randomly chosen questions with their characteristics that are question length, whether the question contains frequently found keywords, whether the question contains an ontology vocabulary, and whether the question contains a misspelled word. An underlined word is one of frequently found keywords in the questions in the web forum. A word with parentheses is an ontology vocabulary. A word with a star is a misspelled word.

Table 2 Random questions in the evaluation survey.

ID	Length	Question
1	Short	(ลูก)เท้าเย็นแล้วก็สั่น (The (baby) has cold feet and is shivering)
2	Short	ภูมิแพ้ผิวหนังกับ(อาหาร)ค่ะ (Skin allergies with (food))
3	Short	แย่แล้วน้ำ(นม)หายไป (Breast (milk) is dried up)
4	Short	บทความ ช่วย(ลูก)ด้วย ของติดคอ (Please help! Something is stuck in my (baby)'s throat)
5	Medium	อยากรู้ผลเสียการนำ(นม)แช่แข็งละลายแล้วไปแช่แข็งอีก (Can breast (milk) be refrozen after being defrosted?)
6	Medium	แม่ ๆ[นิ่ง]*ที่ปั๊ม(นม)กันถึงเมื่อไรคะ (How long do moms pump (milk)?)
7	Medium	ต้องกลับไปทำงานต่างจังหวัด.....หย่า(นม)? (I have to work out of town. Do I need to stop giving (milk)?)
8	Long	อยู่ดี ๆ น้ำ(นม)ก็ลดวูบ ทั้ง ๆที่ปั๊มเหมือนเดิม นมหมด? (I have been pumping my (milk) regularly but there seems to be decreased amount of milk now. Has my milk dried up?)
9	Long	เบบี๋แต่ละบ้าน หย่า(นม)กันตอนอายุเท่าไหร่คะ (How old was your baby when you stopped giving (milk)?)
10	Long	(ลูก)อายุ 3 เดือนปั๊มได้ครั้งละ 4 ออนซ์น้อยไปไหมค่ะ (My (baby) is 3 months old and I can pump 4 ounces, is it enough?)

Evaluation metrics

We evaluated the statistical relevancy of suggested answers across questions using Mean Reciprocal Rank (MRR), which is defined by;

Reciprocal rank $RR_i = \frac{1}{rank_i}$

Mean reciprocal rank over n questions $MRR = \frac{1}{n}\sum_{i=1}^{n} RR_i$

MRR is chosen as an evaluation metric since it is the official measurement used for the Q&A system in TREC [16]. To evaluate the system, a question is correctly answered only when the user chooses at least one of the suggested answers. If no user chooses any of the candidate answers, the score for that question is equal to zero.

Evaluation results

We select the questions of interest depicted in **Table 2** for discussion.

Relevancy results comparison between systems

Table 3 expresses the MRR comparison between the 3 systems. From this table, we conclude that question length does not have any effect on the MRR values of each system. As shown in **Table 3**, for all ranges of length, all systems have the highest MRRs (underlined values) for some questions.

Table 3 Average MRR of users for each question (categorized by length) in survey.

Question	Length	Google	LDA	Ontology
1	Short	0.037	0.019	0.018
2	Short	0.000	0.000	0.019
3	Short	0.007	0.110	0.022
4	Short	0.133	0.015	0.023
5	Medium	0.047	0.099	0.031
6	Medium	0.037	0.127	0.101
7	Medium	0.000	0.019	0.049
8	Long	0.029	0.009	0.000
9	Long	0.080	0.109	0.155
10	Long	0.055	0.130	0.000

It has been found that the query length is not a parameter that affects the MRR value of any of the systems. Next, we consider other 3 factors: 1) whether the question contains the frequently found keywords, 2) whether the question contains an ontology vocabulary, and 3) whether the question contains an ambiguous or misspelled word, as shown in **Table 4**.

Table 4 Average MRR of users for each question in survey.

Question	Length	Ambiguous	Has frequently found keywords	Has ontology vocabs	Google	LDA	Ontology
3	Short	-	นม (milk)	นม (milk)	0.007	0.110	0.022
5	Medium	-	นม (milk)	นม (milk)	0.047	0.099	0.031
6	Medium	นิ่ง (still) /นึ่ง (stream)	นม (milk)	นม (milk)	0.037	0.127	0.101
10	Long	-	ลูก (baby)	ลูก (baby)	0.055	0.130	0.000
Average					0.037	0.116	0.039
1	Short	-	ลูก (baby)	ลูก (baby)	0.037	0.019	0.018
4	Short	-	ลูก (baby)	ลูก (baby)	0.133	0.015	0.023
8	Long	-	นม (milk)	นม (milk)	0.029	0.009	0.000
Average					0.066	0.014	0.014
2	Short	-	-	อาหาร (food)	0.000	0.000	0.019
7	Medium	-	นม (milk)	นม (milk)	0.000	0.019	0.049
9	Long	-	นม (milk)	นม (milk)	0.080	0.109	0.155
Average					0.027	0.042	0.074
Overall Average					0.043	0.058	0.042

From **Table 4**, with the ontology approach, the MRR values of question 8 and question 10 are 0, because the questions are in the form of spoken language rather than written language. With the LDA-based system and the Google search engine, the MRR value of question 2 is 0, because the question does not contain any top frequently found keyword.

Both the Google search engine and the LDA-based system provide higher MRR values, with a high frequency of found keywords in the domain. The Google search engine has the distinct performance of full-text-search, which is able to answer questions with top keywords, such as question 8.

However, the LDA-based system and the ontology-based system have higher MRR values than the Google search engine in the case where there are some ambiguous or misspelled words, such as question 6. The LDA-based system results in higher MRR for the questions that have a complex structure or contain ambiguous or misspelled words, because it automatically searches for related terms using vector similarity. On the other hand, in an ontology-based system, related terms are found using WordNet.

The ontology-based system is superior to both the Google search engine and the LDA-based system for questions that do not contain frequently found words but contain an ontology vocabulary, such as question 2.

The LDA-based system outperforms the Google search engine and the ontology-based system for more numbers of questions, because it can respond well to the questions with frequent keywords as well as the questions with some ambiguous words.

Conclusions

We have created a system to suggest answers for a given question. Our system uses both information retrieval technique and semantic web technology. We have also created an ontology for semantic suggestions. The knowledge is based on frequently asked questions that are assumed to relate with specific knowledge in the domain. We evaluated our system using a survey to ask domain experts to compare our 2 subsystems, language-model-based and ontology-based, with the Google custom web search.

We have found that no system performs the best for all kinds of questions. For questions with top keywords, the Google search engine, which is a full-text search-based system, or our proposed LDA-based system should be applied. However, for ambiguous keywords, the LDA-based system, which understands language structure, should be employed. For questions without frequently found keywords, using ontology is more likely to help users retrieve more desirable answers than by using other approaches. Nevertheless, the LDA-based system performs well for many kinds of questions, such as those questions with frequently found keywords and those questions with ambiguous words.

In the future, we need to use context-free grammar to analyze sentence structure for higher precision. The system should automate or semi-automate ontology construction and triplestore knowledge extraction from documents. It should have a description logic programming system to manage description logic inference rules in ontology.

Acknowledgements

We wish to thank to Yuqing Melanie Wu, Nuwee Wiwatwattana, and Antony Harfield for the comments and suggestions that have helped to improve this work. This work has been supported by the Faculty of Engineering, Khon Kaen University, and it has also been funded by Khon Kaen University.

References

[1] K Inui, S Abe, K Hara, H Morita, C Sao, M Eguchi, A Sumida, K Murakami and S Matsuyoshi. Experience mining: Building a large-scale database of personal experiences and opinions from web documents. *In*: Proceeding of the IEEE/WIC/ACM International Conference Web Intelligence and Intelligent Agent Technology 2008. IEEE Computer Society, Sydney, Australia, 2008, p. 314-21.

[2] IH Ting, PS Chang and SL Wang. Understanding microblog users for social recommendation based on social networks analysis. *J. Univ. Comput. Sci.* 2012; **18**, 554-76.

[3] Z Chen and D Wen. A new web-service-based architecture for question answering. *In*: Proceedings of 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering, IEEE Computer Society, Beijing, China, 2005, p. 179-83.
[4] NI Al-Rajebah and HS Al-Khalifa. Extracting ontologies from Arabic Wikipedia: A linguistic approach. *Arabian J. Sci. Eng.* 2014; **39**, 2749-71.
[5] M Shaheen and AM Ezzeldin. Arabic question answering: Systems, resources, tools, and future trends. *Arabian J. Sci. Eng.* 2014; **39**, 4541-64.
[6] G Chen and MM Chiu. Online discussion processes: Effects of earlier messages' evaluations, knowledge content, social cues and personal information on later messages. *Comput. Educ.* 2008; **50**, 678-92.
[7] F Lamberti, A Sanna and C Demartini. A relation-based page rank algorithm for semantic web search engines. *IEEE Trans. Knowl. Data Eng.* 2009; **21**, 123-36.
[8] A Kongthon, S Kongyoung, C Haruechaiyasak and P Palingoon. A semantic based question answering system for thailand tourism information. *In*: Proceedings of the Knowledge and Reasoning for Answering Questions 2011. Chiang Mai, Thailand, 2011, p. 38-42.
[9] W Jitkrittum, C Haruechaiyasak and T Theeramunkong. Qast: Question answering system for thaiwikipedia. *In*: Proceedings of the 2009 Workshop on Knowledge and Reasoning for Answering Questions, Association for Computational Linguistics, Suntec, Singapore, 2009, p. 11-4.
[10] M Suktarachan, P Rattanamanee and A Kawtrakul. The development of a question-answering services system for the farmer through sms: Query analysis. *In*: Proceeding of the 2009 Workshop on Knowledge and Reasoning for Answering Questions, ACL and AFNLP, Suntec, Singapore, 2009, p. 3-10.
[11] TF Gharib, N Badr, S Haridy and A Abraham. Enriching ontology concepts based on texts from www and corpus. *J. Univ. Comput. Sci.* 2012; **16**, 2234-51.
[12] P Charoenpornsawat. Software: Swath - Thai Word Segmentation, Available at: http://www.cs.cmu.edu/~paisarn/software.html, accessed May 2014.
[13] Linux.thai.net: "libthai library", Available at: http://linux.thai.net/projects/libthai, accessed May 2014.
[14] S Poltree and KR Saikaew. Thai word segmentation web service. *In*: Proceedings of the Joint International Symposium on Natural Language Processing and Agricultural Ontology Service, King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand, 2011, p. 115-9.
[15] M Boriboon, K Kriengket, P Chootrakool, S Phaholphinyo, S Purodakananda, T Thanakulwarapas and K Kosawat. Best corpus development and analysis. *In*: Proceedings of the 2009 International Conference on Asian Language Processing 2009, IEEE Computer Society, Washington DC, USA, 2009, p. 322-7.
[16] EM Voorhees and DM Tice. Building a question answering test collection. *In*: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Athens, Greece, 2000, p. 200-7.