A Hybrid Approach for Natural Language Querying Segmentation for Tourism Ontology

A. Salaiwarkul[1] and S. Khruakong[2]
[1] Department of Computer Science and Information Technology, Naresuan University, Thailand.
[2] Centre for Real-time Information Networks, University of Technology Sydney, Australia.
anongporns@nu.ac.th

***Abstract*—We propose an approach to the interpretation of natural language queries, particularly relevant to the tourist industry in Thailand, by applying natural language queries against a tourism ontology containing tourism information specific to Thailand, and in the Thai language. Queries in Thai language are difficult to segment into words and meaningful phrases given that Thai has no word separation, such as in European languages which have a space between words, meaning specific Thai natural language processing is required. This paper demonstrates the identification and comparison of various methodologies currently available for segmenting natural language phrases, which allowed us to develop a hybrid approach based on aspects of natural language processing drawn from the methods analyzed. Our primary contribution is the hybrid approach, which was applied particularly to the queries and questions likely to be posed by Thai tourists in Thai language. Our discussion presents and describes the framework of the proposed methodology, which applies Thai Word Segmentation together with Trigram and the Name Entity method, together with our evaluation of the query response accuracy achieved, which collectively was 99%. We are confident that the proposed approach can be applied for developing any semantic searching application that allows natural language query, otherwise than for tourism in Thailand, our immediate concern.**

***Index Terms*—Name Entity; Natural Language Query; Tourism Ontology; Trigram.**

I. INTRODUCTION

The amount of tourism information provided on the Internet is comprehensive yet may be overwhelming. Searching the Internet using keywords may not provide all the desired information a user needs Semantic inquiry is not supported. Therefore, Natural language querying is a desirable, if not essential, approach to finding the required information, especially tourism information. Tourism is a major industry in most countries and providing information to potential tourists is thus an important aspect of tourism promotion.

Internal tourism is a major part of the tourism industry. Under the term 'E-tourism', easy and convenient access to a wide variety of information for the tourist is essential, and natural language processing of queries is an important approach to providing such information, specific to the questioner [1, 2]. Natural language processing (NLP) [3] should be adopted to support searching by a user.

A Thai tourist looking for information on the internet may not find information which meets their needs because of the complications inherent in the Thai language. The complexity of the Thai language includes its script, syntax, grammar; natural language processing [4] has particular problems to overcome. Solving these problems can help a user to find the correct tourism. Word segmentation is crucial in natural language processing of the Thai language which does not use delimiters to separate words, as occurs in European languages with punctuation marks and spaces.

Ontology technology is designed to be a part of the semantic search [5-7] which can support the searching system. We have applied ontology technology to facilitate semantic searching of the huge amount of information available in the Thai language, based on users' queries. The perceived advantage of using ontology as the core knowledge base for an information retrieval system is that it enables the user to query the system with a semantic based question and allows the system to provide answers that are related to the question even if the keywords are not found or matched. However, the challenge of such a query system that interacts with people is to be able to interpret a user's natural language queries and to submit the query to the ontology semantic domain.

This paper presents the E-tourism system applicable to the Thai tourism industry that applies an ontology as the core concept together with natural language processing of the Thai written language for discovering information about tourist attractions. The system provides necessary information to assist and advise tourists in their excursions and visits in Thailand as shown in Figure 1. To increase the accuracy of the retrieved information achieved from the semantic searching, the design of the tourism ontology is shown and the technique considered most suitable for processing natural language queries with correct word segmentation is proposed and discussed. The algorithm of Natural language processing that applies the name entity for matching to the tourism ontology supports more accurate selection of the category of the tourism ontology domain. Finally, we illustrate the design of the framework, and evaluate and discuss our search for the best algorithm for our framework for achieving accuracy in understanding Thai language query text. We believe that our hybrid approach can support efficient semantic searching which can be applied to E-tourism.

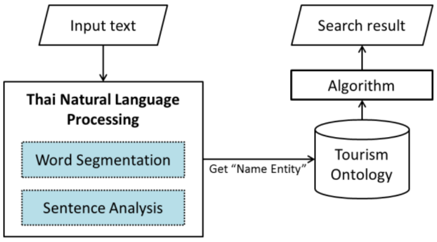

Figure 1: Overview of the Proposed Hybrid Approach.

II. RELATED WORKS

In this section we describe works that are related to our approach which uses the Thai natural language processing and ontology for segmenting the query and semantic search from a user.

An ontology is a specification of conceptualization [8] which may be input or output as knowledge agents applying to applications or software. It will be used for communicating to the agents in the ontology language with the ontology's vocabulary. An ontology can be built to support many diverse applications, such as a travel ontology for E-tourism [9], Siricharoen proposed using an ontology as the key concept for exchanging tourism information. An ontology was developed as an intelligent tour planning application for tourists but has the drawback that it only matches the travel information with the supplied information from the user with a simple keyword matching [10]. Our literature review identified research showing an ontology that has contributed to the Thailand travel industry [11] but this ontology was designed for solving queries on specific aspects of tourism such as Thai culture which is insufficient to support the semantic searching and presentation of general tourism information to the user.

Guo and Zhang proposed a Chinese language question and answer system that uses an ontology as a knowledge base to enhance the question and answer system (FAQ) [12]. This system can invoke answers to a question, based on the semantics of the question, rather than just on keyword matching.

In the aspect of natural language processing reseach review, Thai word segmentation is one of the techniques of NLP. Chaonithi proposed a Thai language based system that applied a hybrid approach [13], which utilized a dictionary based algorithm accessing a crowd-sourced dictionary, to use the bi-gram model, to extract basic and compound word segments from the dictionary. The method proposed in this research improved on other methods by recognising abbreviations. However specific names of events and attractions cannot be extracted correctly. An attempt to understand users' opinions that are published on social networks in natural language was proposed by [14], which uses a rule-based model together with the n-gram approach to understand and analyze the semantics inherent in the language of the texts.

An existing Thai word segmentation software product that is widely used and referenced is TLex [15-17] which uses a Conditional Random Fields (CRFs) algorithm to train a Thai word segmentation model based on a given corpus (Thai word segmentation corpus), which contains 5 million words. This algorithm solves the problem of unknown and ambiguous words in a dictionary based segmentation process. The experimental result showed increased performance in word segmentation.

Another Thai word segmentation software product, LextoPlus, is open-source software created by NECTEC [18]. It uses a longest-matching algorithm in order to segment the given sentence [18]. The longest-matching algorithm is a technique of finding the longest word or phrase in the database for the word segmentation process.

To the best of our knowledge, the primary purpose of the previous works focused on an ontology design as a knowledge base for retieving a semantic search and use keywords to find the appropriate travel information. We propose the development of an E-Tourism system that utilizes an ontology and NLP approach as a core concept for semantic search and natural language query. Additionally, we will improve the efficiency of the searching technique which enables the system to understand the natural language query, thereby allowing a more precise response to a user query and retrieving the tourism information that they require.

III. METHODOLOGY

In this section, we present a system that can extract and present tourism information in a way that allows a user to ask questions in natural language. As current methodologies based on a rule-based model, or the n-gram or Thai Word Segmentation (TWS) engine cannot be directly used as they are limited to extracting specific names of events or attractions with direct keyword matching, we propose our framework to assist the user to be able to query in their normal language. The system is specific to Thailand and supports the Thai language only. To enhance the information retrieval performance, specific names of attractions, hotels, well-known personalities, events and places of interest are stored and natural language queries that indicate these data entities are processed.

The framework of the system, entitled the Thai Tourism Natural Language Ontology System. The system processes include the word segmentation task to segment the sentence entered by a user into words, based on the dictionary. However, this process does not comprehend the semantics of the user request hence sentence semantic analysis is then required. Ontology is applied as a core concept that allows the semantics of the query to be understood.

A. Word Segmentation

In our framework, the TWS engine, TWH, is used to process the natural language user's queries by extracting text segments (word segments) in the Thai language. Each sentence is segmented using a grammatical and dictionary based approach. The separated words from this process could misinterpret the meaning of the sentence. Figure 2 illustrates the patterns that are possible with the different segmentation approaches.

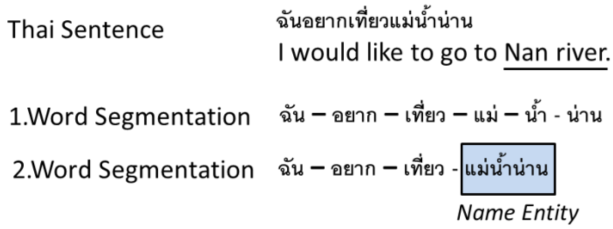
Thai Sentence
ฉันอยากเที่ยวแม่น้ำน่าน
I would like to go to Nan river.
1.Word Segmentation
ฉัน – อยาก – เที่ยว – แม่ – น้ำ - น่าน
2.Word Segmentation
ฉัน – อยาก – เที่ยว - แม่น้ำน่าน
Name Entity

Figure 2: Examples of Word Segmentation

As shown in Figure 2, for example, when the sentence "ฉันอยากเที่ยวแม่น้ำน่าน" is segmented it may result in two word-patterns. The first pattern extracts the sentence into six words. the Thai word for 'Nan River' is (phonetically) 'mai naam Nan'. The first two words have individual meaning; 'mother' and 'water' and the last word could be seen as the name of a Thai Province which is not relevant to the query. The second pattern which uses the longest matching algorithm with the name entity approach recognises the name of an attraction which is the desirable outcome in our research purpose. There may be other patterns derivable from this sentence, but they are not relevant to our immediate purpose.

Our objective is to find the specific name of attractions where the focus is on the tourism information by applying the name entity technique. Identifying each word and failing to find them in the corpus does not allow correct interpretation of the sentence to identify that it is the specific name of attraction in the sentence. It is only when combination of the words, or the phrase, is constructed and that is recognised in the ontology that the query can be precisely serviced.

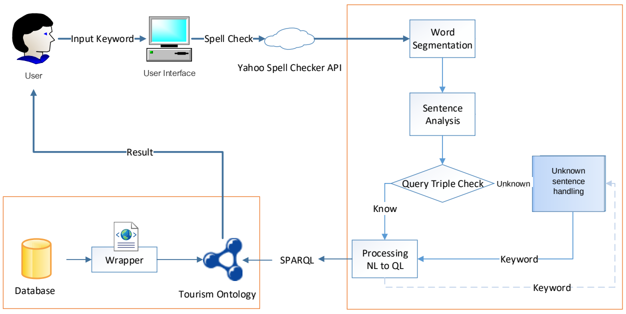

Figure 3: Framework of our Ontology-based System

Our proposed system, diagrammed in Figure 3, overcomes this problem. When a user inputs a query into the system, we use Yahoo Spell Checker as a spelling checking tool. Correct queries are segmented into words using the word segmentation engine. The method applied in our ontology-based system was used to test the accuracy of the published methods. However, as discussed earlier, segmenting single words does not allow identification of events and attractions identified by phrases. To solve this problem, a more sophisticated sentence analysis is required.

B. *Sentence Analysis*

If the algorithms can’t identify some words after the word segmentation process, they are sent to the sentence analysis. The sentence analysis will predict those words from the sentence.

The individual word segments are checked to see if they exist in the corpus with their own meaning. Such words may be, for example, the name of an event, attraction or geographical feature, which are often named in phrases, rather than a single word. Words that are not specifically or individually located in the corpus are ignored in further processing.

When the users search the travel information on the Internet, they may input text as whole sentences, not only keywords. The NLP uses word analysis and an assumption is made that the segmentation result will be specific names of attractions, hotels, or well-known personalities, even though these names are not exact occurrences in the query.

If the segmented words or phrases identified in the sentence analysis are in the tourism corpus, SPARQL (an SQL-based query language) [19, 20] queries are created and processed against the tourism ontology. Alternatively, the unknown sentence handling action is invoked; this process is shown in Figure 4. The unknown information will be checked against an archive of previously asked questions from the user and assumptions and guesses are made to form a keyword tuple to be processed by an SPARQL-based engine.

However, sentence analysis will contribute to learning the new words or new place information in the future which needs other intelligence methods for support.

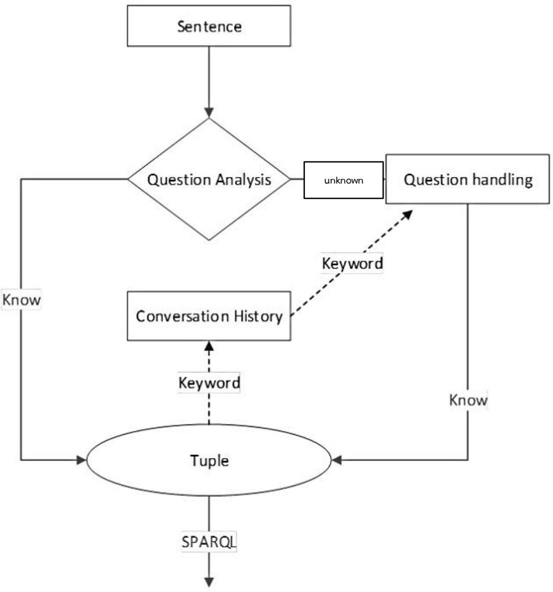

Figure 4: Unknown Sentence Handling Algorithm

C. *Tourism Ontology*

This process is applying the named entity with the tourism ontology after the word segmentation algorithm gets the result of segmentation as words or place name etc. That word will be used as a keyword or name entity for searching in the tourism ontology. The tourism ontology is already established in which the Thailand travel data is classified and all information is categorized to allow the semantic search. The percentage accuracy for the search is followed by the keyword of NLP.

IV. Experiment and Discussion

Our experimental comparison of various natural language processing engines or methods, testing the accuracy of the response to a query, is shown in Table 1. We evaluated the performance of each approach from 100 questions collected from three websites that are popular in Thailand for searching for tourism information. For this purpose, accuracy means the percentage of times that the system could appropriately and correctly segment the query. What was appropriate and correct was decided by the researcher.

The results shown in Table 1 show that the accuracies achieved ranged from 12% to 95%.

Table 1
Comparison Result of Research Techniques

Technique	Precision
Thai Word Segmentation Hybrid (TWH) [13]	12%
LextoPlus	10%
TLex	82%
TWH + Bigram	40%
TWH + Trigram	95%
TWH + Quadrigram	91%

The low scores attained by the TWH approach and the Lexto software are the result of these two methods segmenting the sentence into words without attempting to identify the semantics inherent in the sentence.

The TWH engine together with Bigram, Trigram, and Quadrigram show results of 40%, 95% and 91% respectively. The lower precision achieved by Quadrigram, is lower than Trigram, and is due to the size of the corpus. The greater the value of N in N-gram, the larger the required corpus. The tourism Corpus was not big enough to support Quadrigram, although it was sufficient to support Trigram. We acknowledge the problematic situation where the size of the corpus was a limitation, resulting, as stated, in a lower accuracy achieved by Quadrigram. At this point we do not have any information or thoughts on the necessary size of the corpus, or whether that size is attainable conveniently. This is a variable that we intend to explore in further work.

Table 1 clearly indicates that the best engine for recognising specific words as names of attractions and related knowledge in our tourism ontology in Thailand is the hybrid approach of TWH + Trigram which showed an accuracy of 95%. Even so, this does not allow for proper natural language processing where the meaning and intent of the user’s query is not correctly being interpreted

In our research, we attempted to improve the accuracy of the approach by combining the name entity technique to increase the precision value. In brief, the name entity recognition technique is an approach that tries to locate the specific name in the text by searching the pre-defined corpus which specifies the name of attractions, festival or event. (From our previous example, ‘Mother’ + ‘water’ is identified as a meaningful phrase meaning ‘River,' with the next word probably the name of the river, which, in our example, it is). This method reduces the time required to predict the next item in the sequence. It also reduces the false prediction of the N-gram model. The modified approach shown in Figure 5 illustrates the process of hybrid approach for recognising the particular name in the tourism information ontology. The result of the experiment indicates that our ontology-based approach increased the query-response precision from 95% to 99%.

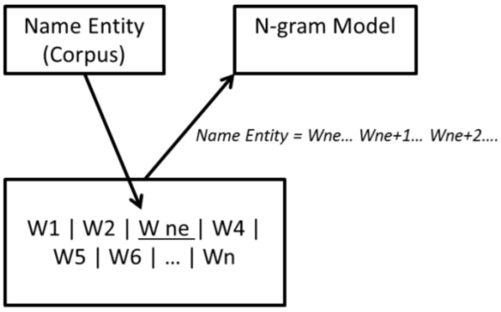
Name Entity (Corpus)
N-gram Model
Name Entity = Wne… Wne+1… Wne+2….
W1 | W2 | W ne | W4 | W5 | W6 | … | Wn

Figure 5: Our Hybrid Approach of Name Entity Approach with N-gram Model

V. Conclusion and Future Work

An enormous amount of useful yet potentially confusing information on tourist attractions and events is available on the Internet. The further significant problem is that the domain appropriate to the query is not identifiable (searching for ‘jaguar’ as a nature-based query will return many responses about motor cars). Natural language processing is a critical approach to allowing user queries to be given appropriate and useful responses. We have applied Ontology technology to facilitate the user in searching for information, enabling semantic searching rather than simple keyword identification and matching, in a bounded data domain, but as yet natural language queries are not supported. The ontology-based semantic searching enhances the performance of the query processing system and better enables natural query processing. Our approach allows the segmentation of words in Thai script, with the ontology based on words and phrases particular to tourism information in Thailand. The TWH together with Trigram and the Name Entity approach, both presented in this paper, achieved an accuracy of 99% which guarantees high quality and relevance of the information returned to the user.

In future research, we will apply this model to other techniques for improving the efficiency of the tourism ontology and allowing an intelligent semantic search. This approach could well be adapted to process dynamic information or real-time travel information such as information on road and traffic conditions, or weather reports, or any other domain-specific information available in Thai language.

References

[1] D. Buhalis and S. H. Jun, "E-tourism," *Contemporary tourism reviews,* pp. 2-38, 2011.

[2] L. Sebastia, I. Garcia, E. Onaindia, and C. Guzman, "e-Tourism: a tourist recommendation and planning application," *International Journal on Artificial Intelligence Tools,* vol. 18, pp. 717-738, 2009.

[3] G. G. Chowdhury, "Natural language processing," *Annual review of information science and technology,* vol. 37, pp. 51-89, 2003.

[4] S. Iwasaki and I. P. Horie, *A reference grammar of Thai*: Cambridge University Press, 2005.

[5] T. Tran, P. Cimiano, S. Rudolph, and R. Studer, "Ontology-based interpretation of keywords for semantic search," in *The Semantic Web,* ed: Springer, 2007, pp. 523-536.

[6] D. Bonino, F. Corno, L. Farinetti, and A. Bosca, "Ontology driven semantic search," *WSEAS Transaction on Information Science and Application,* vol. 1, pp. 1597-1605, 2004.

[7] R. Guha, R. McCool, and E. Miller, "Semantic search," in *Proceedings of the 12th international conference on World Wide Web*, 2003, pp. 700-709.

[8] S. Bechhofer, "OWL: Web ontology language," in *Encyclopedia of Database Systems,* ed: Springer, 2009, pp. 2008-2009.

[9] W. V. Siricharoen, "Using Ontologies for E-tourism," in *The 4th WSEAS/IASME International Conference on Engineering Education (EE 2007) Proceeding*, 2007, pp. 113-118.

[10] R. Jakkilinki, M. Georgievski, and N. Sharda, "Connecting destinations with an ontology-based e-tourism planner," *Information and communication technologies in tourism 2007,* pp. 21-32, 2007.

[11] S. Khruahong, X. Kong, and D. Hoang, "Ontology Design for Thailand Travel Industry," *International Journal of Knowledge Engineering,* vol. vol. 1, no. 3, pp. 191-196, 2015.

[12] Q. Guo and M. Zhang, "Question answering system based on ontology and semantic web," in *International Conference on Rough Sets and Knowledge Technology,* 2008, pp. 652-659.

[13] K. Chaonithi and S. Prom-on, "A hybrid approach for Thai word segmentation with crowdsourcing feedback system," in *Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 2016 13th International Conference on,* 2016, pp. 1-6.

[14] S. Poria, E. Cambria, G. Winterstein, and G.-B. Huang, "Sentic patterns: Dependency-based rules for concept-level sentiment analysis," *Knowledge-Based Systems,* vol. 69, pp. 45-63, 2014.

[15] C. Haruechaiyasak and S. Kongyoung, "TLex: Thai lexeme analyser based on the conditional random fields," in *Proceedings of 8th International Symposium on Natural Language Processing*, 2009.

[16] C. Haruechaiyasak, S. Kongyoung, and M. Dailey, "A comparative study on thai word segmentation approaches," in *Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, 2008. ECTI-CON 2008. 5th International Conference on*, 2008, pp. 125-128.

[17] N. Wanichayapong, W. Pruthipunyaskul, W. Pattara-Atikom, and P. Chaovalit, "Social-based traffic information extraction and classification," in *ITS Telecommunications (ITST), 2011 11th International Conference on*, 2011, pp. 107-112.

[18] C. Haruechaiyasak and A. Kongthon, "LexToPlus: A thai lexeme tokenization and normalization tool," *WSSANLP-2013*, p. 9, 2013.

[19] E. Prud and A. Seaborne. (2006, 15 Jan 2017). *SPARQL query language for RDF*. Available: http://www.w3.org/TR/rdf-sparql-query

[20] B. Quilitz and U. Leser, "Querying distributed RDF data sources with SPARQL," in *European Semantic Web Conference*, 2008, pp. 524-538.