Research Article

Natural Language Interface to Database for Data Retrieval and Processing

Chalermpol Tapsai* and Phayung Meesad
Department of Information Technology Management, Faculty of Information Technology, King Mongkut's University of Technology North Bangkok, Bangkok, Thailand

Choochart Haruechaiyasak
National Electronics and Computer Technology Center, Pathum Thani, Thailand

* Corresponding author. E-mail: chalermpol.t@email.kmutnb.ac.th DOI: 10.14416/j.asep.2020.05.003
Received: 5 September 2019; Revised: 9 April 2020; Accepted: 29 April 2020; Published online: 21 May 2020

Abstract
Though many studies related to natural language interface to a database have been conducted for many years, the results of these studies are not covered in many used cases such as the use of negative sentences, processing functions, and variety of sentence patterns with various types of query specification. To solve these problems, a model called "Natural Language Processing for Data Retrieval and Processing (NLP-DRP)" was proposed. A new algorithm named 'Ranking Trie' was implemented with the combination of Pattern Parsing, Ontology, and Fuzzy system to improve Lexical analysis, Semantic analysis, and Output transformation processes to allow users to retrieve and process data with various patterns of sentences and conditions. The model was incrementally tested and updated by a Learning dataset collected from users with a total of 3,868 Natural Language Query Sentences (NLQSs) then finally evaluated by the test dataset with a total of 500 NLQSs. The results showed that the NLP-DRP could retrieve data, processed, and generated outputs which consistent with user requirements with all values of Accuracy, Precision, Recall, and F-measure higher than 0.9.

Keywords: Data retrieval, Natural language, Pattern parsing, Ranking Trie, Ontology, Fuzzy system

1 Introduction

Currently, the database is an essential source of information that is popular and widely used. Retrieving information from a database needs SQL language, which is a computer language with a specific format, and users must be trained and learned until they have enough skills. Moreover, in order to make a correct SQL command, users must have knowledge and understanding of both the storage structure and the relationship between data items within the database. These factors are major obstacles that make general users not be able to retrieve the information from a database themselves. For many years, many researchers have conducted in Natural Language Processing (NLP) in order to allow users to communicate with computers by human languages. Natural Language Interface to Database (NLIDB) is one of the most popular topics which allows humans to retrieve information from a database by natural language without having additional training. However, natural language is a highly flexible language that can be written in many styles. With a variety of words and sentence patterns, errors are easily occurred in the NLP research and make the existing results not cover in many issues of actual usage, including the use of negative sentences, processing function, and variety of retrieval conditions and sentence patterns.

Please cite this article as: C. Tapsai, P. Meesad, and C. Haruechaiyasak, "Natural language interface to database for data retrieval and processing," *Applied Science and Engineering Progress*, vol. 14, no. 3, pp. 435–446, Jul.–Sep. 2021, doi: 10.14416/j.asep.2020.05.003.

Thai language is a non-segmentation natural language that all words are written continuously without any spaces or special characters separated between words. This written style is found in many languages such as Chinese, Japanese, Laos, Arabic, etc., and easily causes many errors in the segmentation process. Therefore, most studies related to natural language processing of Thai language mainly focus on word segmentation, and just only a few studies were conducted in more advanced topics. For this reason, the researcher is interested in developing an information retrieval model from the database focusing on solving the above problems to make users easily retrieve information by using numerous sentence patterns and query conditions. The remainder of this article divided into seven parts: 1) Background 2) Conceptual framework 3) Natural Language Processing for Data Retrieval and Processing 4) Functional testing and model improvement 5) Performance evaluation of the model 6) Results, and 7) Conclusions.

2 Background

2.1 *Natural language processing*

Natural language processing consists of 4 steps, including Lexical Analysis, Syntactic Analysis, Semantic Analysis, and Output Transformation [1].

2.1.1 Lexical analysis

This process aims to analyze and separate the natural language sentences into words with its' type, which will be used for further analysis in the following step. Word segmentation is a very important process in Lexical analysis, especially for a non-segmentation language such as Thai, Chinese, and Japanese, etc. For the Thai language, there are many studies focus on this topic, which divided into four techniques, including rule-base word segmentation (RB-WS), dictionary-based word segmentation (DB-WS), and learning-based word segmentation (LB-WS).

RB-WS is the first phase of the Thai word segmentation algorithm that was used to analyze the experimental texts to define the boundary of syllables which is an elementary component of each word. [2], [3]. By using linguistic spelling principles which relate to the characteristics of each type of Thai alphabets, the specific rules were set up to identify the front boundary and the rear boundary of each syllable. Although this method can identify the boundaries of the syllables with very high accuracy, most words in Thai are consist of more than one syllable. Therefore, it does not work well for word segmentation.

DB-WS was firstly used for Thai word segmentation by Pooworawan [4]. This algorithm firstly designed for defining the boundary of each syllable by parsing the input text with a dictionary that consists of Thai syllables with the longest match strategy. Unfortunately, although this method has a high degree of correctness in syllable scoping, it does not work well for WS. This problem is later improved by Raruenrom [5], who changed from the syllable dictionary to be a dictionary of words and reduce the time for parsing by using the Trie structure to organize words in the dictionary. At a later time, dictionary-based WS has studied and conducted by many techniques, such as Maximal Matching [6], which will define all possible word boundaries and select the best result, which improved ambiguous word segmentation. Until now, the most popular Thai DB-WS program named LexTo [7] is developed and distributed by NECTEC.

MLB-WS is a non-dictionary word segmentation algorithm that required the training dataset that will be analyzed and extracted for some features and used to defined the boundaries of words. The examples of MLB-WS studies are "Feature-based Thai Words Segmentation" [8], [9].

Haruechaiyasak [10] compared the efficiency of DB-WS with four techniques of MLB-WS. The result showed that the DB-WS provided the highest accuracy in word segmentation.

2.1.2 Syntax analysis

This step is intended to verify the accuracy and completeness of the sentence to ensure that no important parts of the sentence are missing. Many studies used two techniques, which are Syntax Tree [11], [12], and Grammar rules [13], [14], for this process. Syntax Tree is a tree structure created reliably on each sentence pattern of natural language. The example of Syntax Tree of the English sentence "What country is in ASIA" showed in Figure 1. In the same way, the Grammar rules are sets of rules, which are created

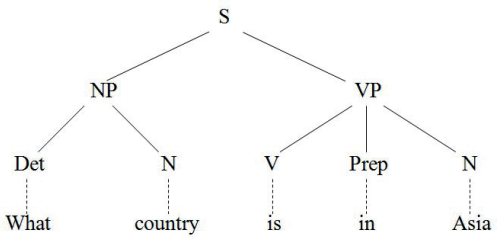

Figure 1: Syntax tree.

S (Sentence)	→ NP+VP
NP (Noun Phrase)	→ Det+N
VP (Verb Phrase)	→ V + NP | V+Prep+N
Det (Determiner)	→ What|Which
Prep (Preposition)	→ in|on|at
N (Noun)	→ country:continent| Thailand:Asia|USA:America
V (Verb)	→ is|am|are

Figure 2: Grammar rules.

reliably on each sentence pattern of natural language. The example of Grammar rules shown in Figure 2.

To check the correctness and completeness of sentences, all words that result from the Lexical Analysis will be parsed with the Syntax tree or Grammar rules to review each sentence for missing components. If an error is encountered, the process may terminate or report errors to the user. On the other hand, the correct sentences will be forwarded to the process in the next step.

2.1.3 Semantic analysis

This process analyzes the meaning of sentences and produces a meaning representation that will be converted into results according to the purpose of each job. Many techniques have been studied and presented for Semantic analysis, such as:

1) Syntactic-based Semantic Rules mapping: Semantic Rules are sets of pre-defined rules for each sentence structure, used to map sentences to the specific outputs. This technique compares each word of the inputted sentence with syntax structure (Syntax Tree or Grammar rules), then mapped each matched pattern to the outputs according to the Semantic Rules. The example of Semantic Rules used to map the sentence "What country is in ASIA" given by Grammar rules in Figure 2, is shown in Figure 3.

(Det N) will be converted to (for_every country (is_country, X) and (V Prep N) will be converted to (show country (is_country, X) (equal_:Asia, X)).

$Det+N_1$	→ for every N_1 (is_N_1, X)	
$V+Prep+N_2$	→ show N_1 (is_N_1, X) (equal_:N_2, X)	
X	→ data-list of (N)	
Note:	Det	is a determiner
	N_1, N_2	is a Noun
	V	is a Verb
	Prep	is a Preposition
	X	is a data in data-list of Nouns

Figure 3: Semantic rules.

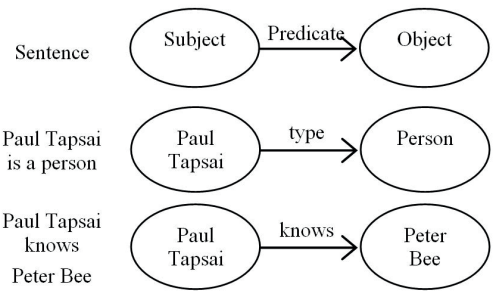

Figure 4: RDF triple.

The sentence "What country is in Asia" will be mapped to (for_every country (is_country, X) (show country) (equal_:Asia, X)) where X is a data-list in (N) that are Thailand:Asia|USA:America. Therefore, the result will be "Thailand."

The examples of studies that implement this technique are [15]–[18].

2) Ontology, a definition system of terms (words) and the relationship between the terms for meaning representation. Semantic analysis of words and sentences by Ontology is based on rules which covered globalization, aggregation, synonym, symmetry, and transmission characteristics, etc. Each ontology rule consists of 3 important parts called triple, including object, predicate, and subject, as shown in Figure 4. For coding, each ontology rule can be written in many XML standard formats defined by The World Wide Web Consortium (W3C), such as:

- RDF (Resource Description Framework) [19]
- RDFS (Resource Description Framework Schema)
- RDFa (Resource Description Framework in Attributes) [20]

```
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-
rdf-syntax-ns#"
  xmlns:p="http://example.org/pers-schema#">
  <rdf:Description rdf:about="http://example.
org/~sname#james">
    <p:name>Paul Tapsai</p:name>
    <rdf:type  rdf:resource=”http://example.org/
pers-schema#Person”/>
  </rdf:Description>
</rdf:RDF>
```

Figure 5: RDF graph of the sentence “Paul Tapsai is a person”.

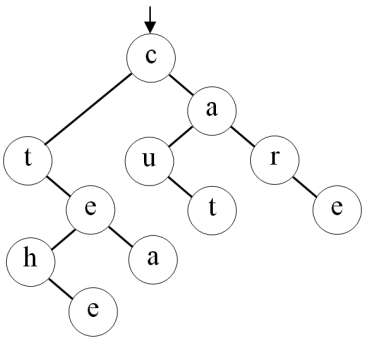

Figure 6: Trie structure.

- SPARQL (Specifications of Protocol and RDF Query Language) [21]
- OWL (Web Ontology Language) [22], [23].

Figure 5 shows the RDF and RDFS definition of the sentence, “Paul Tapsai is a person”. Examples of studies that use ontology to analyze the meaning of natural language sentences are Ontology-based semantic information retrieval for enterprise management information system [24], Ontology-based information retrieval for historical documents [25], An Ontology-based Dialog Interface to Database [26].

2.1.4 Output transformation

This process transforms the meaning representations obtained from the semantic analysis process into the results that meet the user’s requirements such as a command to control robots, message for Question-Answer system, or SQL command for retrieving information, etc.

2.2 *Trie and ranking trie*

Trie is a Tree-like data structure stored words of a dictionary for word segmentation. As shown in Figure 6, each node of Trie contains only one character, and words with the same initial characters will share the nodes together. [27] Such the structure in this way make Trie structure very small and used less time than Tree to compare characters for word boundaries. However, although most Thai word segmentation algorithms used Trie as a standard technique, it was found that the efficiency of these algorithms is not good enough due to lacking a suitable arrangement of words in Trie.

Ranking Trie is a new technique, which arranges words in Trie structure by word usage frequency to improve the word segmentation efficiency. The research of Ranking Trie was conducted in 2016 [28]. In the first step, 1,320 text files were collected from popular websites and chat messages covering all major fields, including economics, social, political, entertainment, and others. Then, 1200 files are randomly selected, segmented into words, and counted the frequency of each word. After that, all words were sorted in descending order on word usage frequency and used to create Trie by placing the words with higher usage frequency at the beginning of the Trie structure before the lower usage frequency. The Ranking Trie was implemented in a new word segmentation algorithm named “Thai Language Segmentation by Automatic Ranking Trie (TLS-ART)” that will update the frequency of words with each segmentation result automatically. In the final step, TLS-ART was tested by the remainder 120 text files to evaluate performance compared to the most popular segmentation algorithm named “LexTo” The results show that TLS-ART has better performance that reduces the number of parsing tasks, and words in the dictionary with the percentage of 12.73 and 86.07, respectively, while the values of precision, recall, and F1-scores are slightly better.

2.3 *Fuzzy system*

In general, many data which cannot be classified by a clear-cut criteria values are usually used in daily life. For example, to specify the human height level, even if the values could be divided into three values: “tall”, “medium”, and “short”, but the criteria may vary

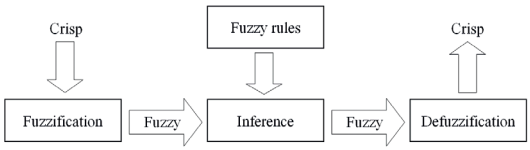

Figure 7: Fuzzy system working process.

according to the knowledge and experience of each person. For some people, if the "tall" value is defined as 180 cm and above, these criteria will result in height 179.9 being considered as "medium" value, although having only 0.1 cm difference. On the other hand, the more reasonable meaning should be considered 179.9 cm as the value that has a degree of "tall" slightly less than 180 cm. The data in this manner called "Fuzzy data", and a Fuzzy system is required for processing the data of this type [29], [30]. As shown in Figure 7, Fuzzification is the first process that converts each input data to a linguistic variable suitable for the nature of each data by using the membership function. Then, all outputs will be inferred by the appropriate fuzzy rules to provide the results in the form of fuzzy values. Finally, all results will pass to the Defuzzification process that transforms the values back into the classical values that meet user requirements.

3 Conceptual Framework

This research presents a new algorithm in retrieving information from a database with Thai natural language. As shown in Figure 8, the researcher developed a model called "Natural Language Processing for Data Retrieval and Processing (NLP-DRP)" that has improved the performance of natural language processing in various steps including, Lexical analysis, Syntax analysis, Semantic analysis, and Output transformation. In Lexical analysis, Ranking Trie was presented to overcome the word segmentation problems and improve segmentation efficiency. In the Syntax analysis, Semantic analysis, and Output transformation, several techniques, such as Ranking Trie, Semantic patterns, Ontology, and Fuzzy system, have been applied to solve problems that occurred with previous studies. In the case of problems related to the variety of sentence patterns and data retrieval conditions, we used Semantic patterns in conjunction with Ontology rules to infer the meaning. The data used for this research are Natural Language Query

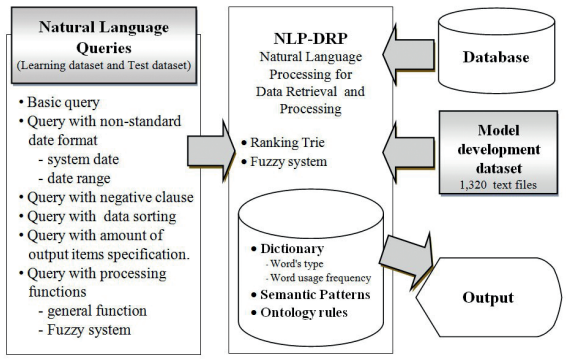

Figure 8: Research's conceptual framework.

(NLQ) sentences with predefined expected results collected from a sample of 473 users covered most sentence types and retrieval conditions in everyday sentences. With a total of 4,368 NLQ sentences, 3,868 are randomly selected and used as the Learning dataset to improve the model. The remainder NLQ sentences were used as the Test dataset for the model evaluation. The Learning dataset with a total of 3,868 NLQ was analyzed and used to create Semantic patterns. The outputs of this process are 406 Semantic patterns of various clauses, phrases, and sentences, which cover most sentence patterns and retrieval conditions based on real-life usage. The cooperation of the Ontology technique with Semantic patterns by applying some Ontology words in Semantic patterns allows the model to be able to analyze words that need to be interpreted, such as synonyms, antonyms, symmetric meaning, and processing functions, etc.

The NLQ sentences used in this research covered most common sentence types, including simple sentence, compound sentence, and complex sentence with various forms of query condition that is consistent with the following Thai language principles [31], [32].

1) The standard terms related to rainfall that can be used in an NLQ statement are defined by the Meteorological Department and divided into two types, which are "Rain distribution criteria" and "Rain quantity criteria" [33]. The Rain distribution criteria is a measure based on the percentage of rainfall area within 24 h that are divided into five levels: "บางพื้นที่ [baːŋ-pʰúːn-tʰîː] (Isolated)", กระจายเป็นแห่งๆ [krà-tɕaːj-pen-hɛ̀ː-hɛ̀ː] (Widely Scattered)", "กระจาย [krà-tɕaːj] (Scattered)", "เกือบทั่วไป [kɯaːp̄-tʰuâː-pai] (Almost Widespread)", and "ทั่วไป [tʰuâː-pai] (Widespread)".

The Rain quantity criteria are the amount of rain that falls within 24 h that are divided into five levels: "วัดจำนวนไม่ได้ [wát-tɕam-nuaːn-mâi-dâi] (Trace)", "เล็กน้อย [lék̄-nɔ́ːj] (Light rain)", "ปานกลาง [paːn-klaːŋ] (Moderate rain)", "หนัก [nàk̄] (Heavy rain)", "หนักมาก [nàk̄-mâːk̄] (Very heavy rain)".

2) Numeral identification can be specified using Arabic numbers, Thai numbers, and/or text. As shown in Table 1.

Table 1: Example of numeral identification

Arabic Numeral	Thai Numeral	Thai Text Numeral	Phonetic Annotation
18	๑๘	สิบแปด	[sīp̄-pɛːt̄]
25	๒๕	ยี่สิบห้า	[jiː-sīp̄-hâː]
451	๔๕๑	สี่ร้อยห้าสิบเอ็ด	[sìː-rɔ́ːj-hâː-sīp̄-ʔèt̄]
2557	๒๕๕๗	สองพันห้าร้อยห้าสิบเจ็ด	[sɔ̌ːŋ-pʰan-hâː-rɔ́ːj-hâː-sīp̄-tɕèt̄]

3) Date, month, and year can specify in many forms. Either a standard format or system-date related format with Thai, Arabic, or text numerals for a specific date or date range, as shown in Table 2.

Table 2: Example of date identification

Date Specification Terms [Phonetic Annotation]	Meaning (Date)
15 มกราคม 2556 [sīp̄-hâː-má-kà-raː-kʰom-sɔ̌ːŋ-hâː-hâː-hòk̄]	15/1/2013
๑๕ มกราคม พ.ศ. ๒๕๕๖ [sīp̄-hâː-má-kà-raː-kʰom-pʰɔː-sɔ̌ː-sɔ̌ːŋ-hâː-hâː-hòk̄]	15/1/2013
15-20 มกราคม 2557 [sīp̄-hâː-tʰŭŋ-jiː-sīp̄-má-kà-raː-kʰom-sɔ̌ːŋ-hâː-hâː-tɕèt]	15/1/2013–20/1/2013
วันนี้ [wan-níː]	Today
เดือนนี้ [duaːn-níː] ปีนี้ [piː-níː]	This month This year
เมื่อวานนี้ [mɯ̀aː-waːn-níː] เดือนที่แล้ว [duaːn-tʰîː-léːw] ปีที่แล้ว [piː-tʰîː-léːw]	Yesterday Last month Last year

4) Identification of time can be specified either at a specific time, time period, or Thai terms according to the definition of meaning in the Royal Institute Dictionary 2011 [34], [35], as shown in Table 3.

5) Sorting can be specified by the terms "เรียง [riaːŋ]" or "เรียงลำดับ" which could be added with "จากมากไปหาน้อย [tɕàːk̄-mâːk̄-pai-hăːn-nɔ́ːj]" for descending order, or "จากน้อยไปหามาก [tɕàːk̄-nɔ́ːj-pai-hăː-mâːk̄]" for ascending order.

Table 3: Example of time identification

Thai Terms	Meaning (Time)
9 น. [kâu-nɔː]	9 o'clock
๙ นาฬิกา [kâu-naː-lí-kaː]	9 o'clock
เช้า [tɕʰáu]	6.00–9.00
สาย [săːj]	9.00–10.00
บ่าย [bàːj]	13.00–15.00
เย็น [jen]	16.00–18.00
กลางวัน [klaːŋ-wan]	6.00–18.00
กลางคืน [klaːŋ-kʰɯːn]	19.00–5.00

6) Output display can be specified for the maximum or minimum values as well as the number of output items needed. For example: "มากที่สุด [mâːk̄-tʰîː-sùt] (maximum)", "น้อยที่สุดสามอันดับ [nɔ́ːj-tʰîː-sùt-săːm-ʔan-dàp] (the three most minimum)", etc.

7) Specifying of the processing function such as : "รวม [ruaːm] (sum)", "ผลรวม [pʰŏn-ruaːm] (total result)", "ยอดรวม [jɔ̂ːt-ruaːm] (grand total)", "นับ [náp] (count)", "นับจำนวน [náp-tɕam-nuaːn] (count for amount)", "กี่ [kìː] (how many)", etc.

The experimental dataset is an hourly rainfall quantity that is measured in 73 provinces prepared by the Meteorological Department and published by Digital Government Development Office [36] with all data for three years, including 2012, 2013, and 2014. This dataset was taken through the pre-processing to check and correct errors and transform them into a database. Moreover, in order to reduce processing time during the retrieval process, the data were summarized for the total quantity of rainfall on each province and time period, such as daily, monthly, yearly, and the Thai time period according to Thai terms in Table 3.

4 Natural Language Processing for Data Retrieval and Processing

The Natural Language Processing for Data Retrieval and Processing (NLP-DRP) model created in this research consists of 4 main processes: Lexical Analysis, Syntactic and Semantic Analysis, SQL Transformation, and SQL Processing as shown in Figure 9.

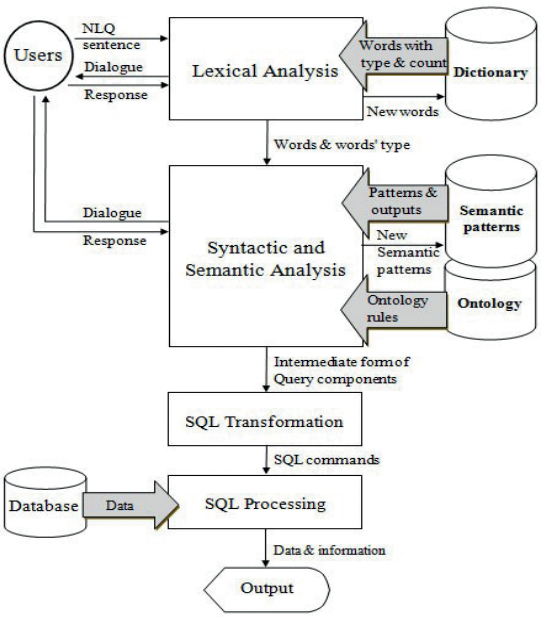

Figure 9: The NLP-DRP model.

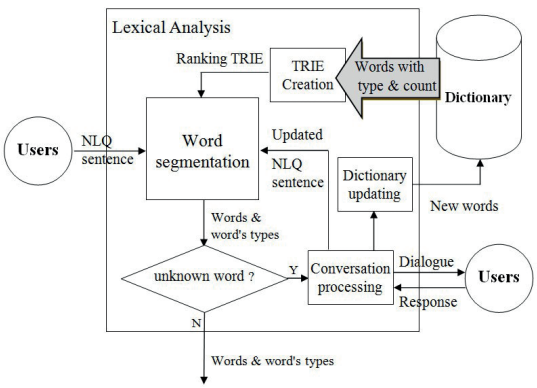

Figure 10: Lexical analysis process.

4.1 *Lexical analysis*

This process analyzes the input that is an NLQ statement for words and word types. The definition of word types in this research complied with the ORCHID corpus [37] guidelines, as some examples shown in Table 4.

Moreover, the researchers have defined eight additional word types: PFK, LK, LSW, TK, TSW, VK, DK, ODW as shown in Table 5.

As shown in Figure 10, the NLQ sentence will be segmented into words and defined each word's type by TLS-ART, the word segmentation module that uses the Ranking Trie created from data in the dictionary. In the case of unknown characters are found, these unknown characters will be passed to the Conversation processing module to generate some dialogue messages which help the user to edit the NLQ statement or add new words to the dictionary.

Table 4: Examples of word types

Type of Words	ORCHID's POS tag	Example (Words)
Common Noun	NPRP	นก [nók] (bird), รถ [rót] (car)
Action verb	VACT	วิ่ง [wîŋ] (run), กัด [kàt] (bite)
Relative pronoun	PREL	ซึ่ง [sûŋ] (that), อัน [ʔan] (which)
Adverb with normal form	ADVN	หนัก [nàk] (heavy), มาก [mâ:k] (much)
Preposition	RPRE	ใต้ [tâi] (under), บน [bon] (on)
Subordinating conjunction	JSBR	เนื่องจาก [nuâ:ŋ-tɕà:k] (cause by), เมื่อ [muâ:] (when)
Cardinal number	NCNM	๑๘ [sip-hâ:] (15), สิบห้า [sip-hâ:] (15)
Negator	NEG	ไม่ [mâi] (new), ยกเว้น [jók-wé:n] (except)

Table 5: Additional word types

Type of Words	New Tag	Example (Words)
Verb keyword	VK	คำนวณ [kʰam-nua:n] (calculate), แสดง [sà-dɛ:ŋ] (show)
Data keyword	DK	ปริมาณ [pà-rí-ma:n] (quantity), จำนวน [tɕam-nua:n] (amount)
Processing function keyword	PFK	รวม [rua:m] (sum), กี่ [kì:] (how many)
Location keyword	LK	ภาค [pʰâ:k] (part), จังหวัด [tɕaŋ-wàt] (province)
Time keyword	TK	ปี [pi:] (year), วันที่ [wan-tʰî:] (date)
Location specification word	LSW	ตรัง [traŋ] (Trang), เชียงราย [tɕʰia:ŋ-ra:j] (Chiengrai)
Time specification word	TSW	เย็น [jen] (evening), สิงหาคม [sǐŋ-hă:-kʰom] (August)
Ontology defined word	ODW	ติดกัน [tìt-kan] (adjacent), ปริมณฑล [pà-rí-mon-tʰon] (perimeter)

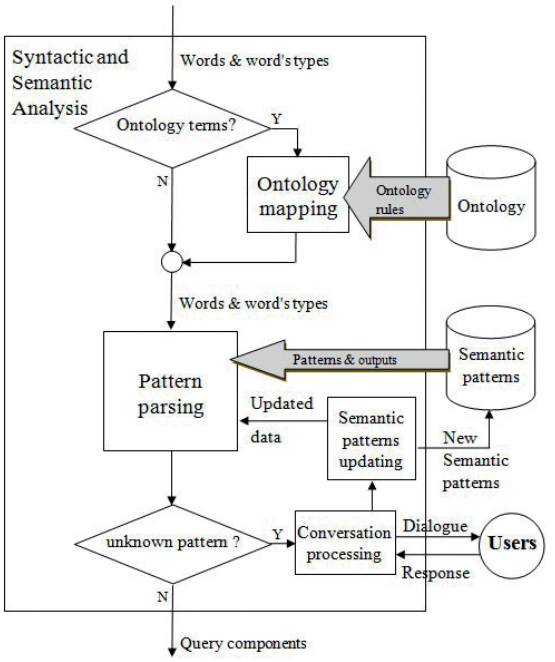

Figure 11: Syntactic and Semantic Analysis process.

4.2 *Syntactic analysis and semantic analysis*

As shown in Figure 11, firstly, all inputted words were checked for Ontology terms and send to the Ontology mapping module to infer the meaning of words in various forms of relationships, such as synonyms, symmetric meaning, inherited meaning, etc. The Ontology mapping module will process each Ontology term by Ontology rules and provide the target words that match with the Semantic Patterns.

Due to differences in individual language usage, synonyms are often found in natural language sentences. Some examples of synonyms found in this research are shown in Table 6. The examples of Ontology processing shown in Figures 12 and 13.

To solved the problem of using synonyms, an Ontology rule is defined by the "sameAs" relationship between both synonyms. For example, Figure 8 shown the annotation and Ontology definitions of the words "กรุงเทพมหานคร" and "กรุงเทพฯ".

Figure 13 shown how to infer the meaning of "กรุงเทพฯและปริมณฑล [kruŋ-tʰêːp-lɛ́-pà-rí-mon-tʰon] (Bangkok and perimeter)" through Ontology rules. Firstly the word "ปริมณฑล [pà-rí-mon-tʰon] (perimeter)" that is an ontology-word will be converted to "ติดกัน

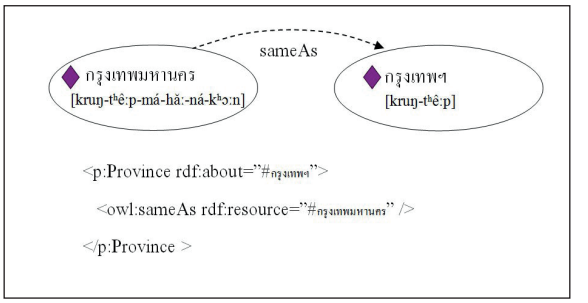

Figure 12: The annotation and the ontology definitions of the "sameAs" relationship.

[tìt-kan] (adjacent)" by the "SameAs" relationship. Then, the ontology-word "ติดกัน [tìt-kan]" that defined by a SymmetricProperty relationship between many provinces and "กรุงเทพฯ [kruŋ-tʰêːp] (Bangkok)" are analyzed and finally output the target province's name: "นนทบุรี [non-tʰá-bù-riː] (Nonthaburi)", "ปทุมธานี [pà-tʰum-tʰaː-niː] (Pathumthani)", "สมุทรปราการ [sà-mùt-praː-kaːn] (Samutprakarn)", "สมุทรสาคร [sà-mùt-săː-kʰɔːn] (Samutsakorn)", and "นครปฐม [ná-kʰɔːn-pà-tʰŏm] (Nakhonprathom)".

Table 6: Examples of synonyms words

List of Synonyms Words	Main Word
ผลรวม [pʰŏn-ruaːm], ผลบวก [pʰŏn-buàːk̄], ยอดรวม [jɔ̂ːt-ruaːm]	ผลรวม [pʰŏn-ruaːm] (sum)
กรุงเทพฯ [kruŋ-tʰêːp], กรุงเทพมหานคร [kruŋ-tʰêːp-má-hăː-ná-kʰɔːn], กทม. [kɔː-tʰɔː-mɔː], บางกอก [baːŋ-kɔ̀ːk̄]	กรุงเทพฯ [kruŋ-tʰêːp] (Bangkok)
พื้นที่ [pʰúːn-tʰîː], บริเวณ [bɔː-rí-weːn], เขต [kʰèːt]	พื้นที่ [pʰúːn-tʰîː] (area)
ติดกัน [tìt-kan], ข้างเคียง [kʰâːŋ-kʰiaːŋ], ปริมณฑล [pà-rí-mon-tʰon]	ติดกัน [tìt-kan] (adjacent)

Then, all words and types were parsed to the Semantic patterns to checked for the integrity of the sentence and identify query components, which corresponding to the sentence's meaning, such as data field-names, location, date, time, and processing function, etc. The example of Semantic pattern parsing shown in Figure 14.

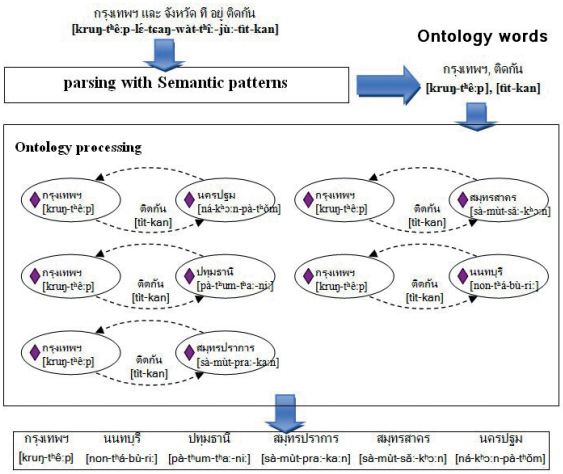

Figure 13: Ontology rules processing.

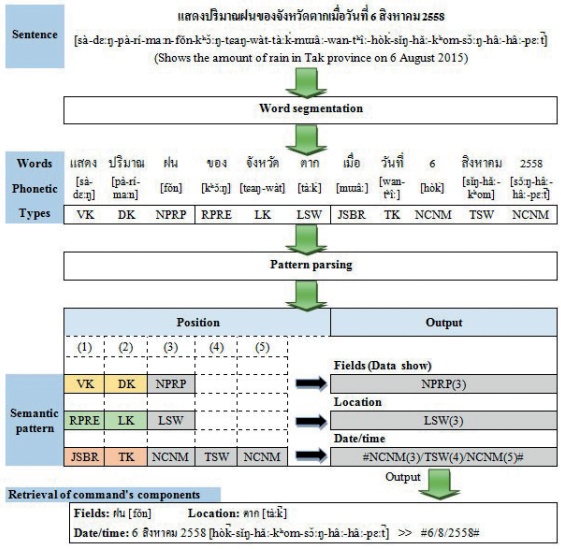

Figure 14: Analysis of a natural language query statement for retrieval components.

In the case of an unknown pattern or lack of some essential words, the model will create a dialogue message along with recommendations that help users to correct the sentence or create new Semantic patterns to improve the model.

4.3 *SQL transformation*

This step transforms the retrieval command's components from the previous step into SQL commands which divided into 5 categories as follow:

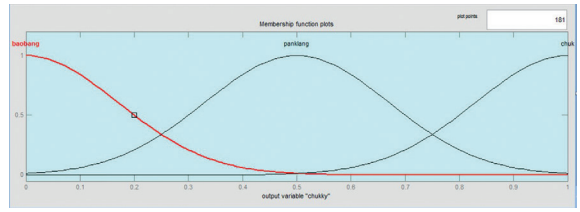

Figure 15: Linguistic variables named "chukky", the output of fuzzy system.

- Commands with basic conditions
- Commands with system date or date range
- Commands with negative clause
- Commands with output specification such as sorting and amount of data displayed
- Commands with processing functions

4.4 *SQL processing*

This step executes the SQL commands for data retrieval and process data for the expected output.

4.5 *Fuzzy data processing*

The example of fuzzy data processing in this research is the word "ชุก [tɕʰúk̄] (dense)", which is a fuzzy value that can be inferred from two inputted data, including quantitative density and spatial density. To achieve this result, we created a fuzzy system with the following specification:

- A linguistic variable named "chukky", with the Gaussian membership function to represent the word "ชุก [tɕʰúk̄] (dense)" as shown in Figure 15.
- The linguistic variables used as inputs are 2 variables. The first variable, "Rain distribution", is the percentage of the area that has rained in the specific area zone that can be divided into 5 levels, including Isolated, Widely scattered, Scattered, Almost widespread, and Widespread. The second variable, "Rain quantity", is the amount of rainfall within 24 h that can be divided into 5 levels, including Trace, Light rain, Moderate rain, Heavy rain, and Very heavy rain. The values of both variables in the first 4 levels are characteristic of Gaussian membership function, and the final value has the characteristic of Sigmoidal membership function as shown in Figures 16 and 17, respectively.
- The inferential Processing of the fuzzy system

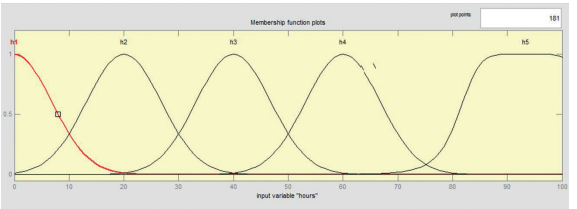

Figure 16: Linguistic variables named "area" for Rain distribution values.

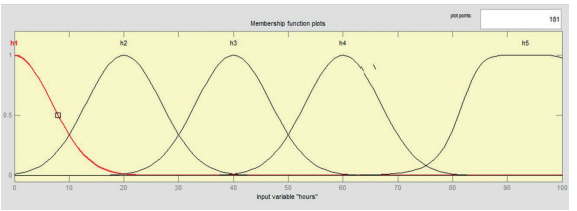

Figure 17: Linguistic variables named "hours" for Rain quantity values.

in this research is Mamdani's fuzzy rules system. The output (chukky) of the fuzzy system is a value between 0 and 1 and the value that will be defined as "ชุก" is 0.8 and above.

5 Functional Testing and Model Improvement

After finish creating the NLP-DRP model, the model was incrementally tested and updated by Learning dataset, which consists of 3,868 natural language query sentences. Then, the results of the testing were used to update all model's components, including Dictionary, Semantic patterns, and Ontology rules for better performance.

6 Performance Evaluation of the Model

The NLP-DRP model was tested by a Test dataset, which consists of 500 natural language query sentences collected from users with various sentence patterns and query specifications consistent to real used cases. All results were collected and calculated for Accuracy, Precision, Recall, and F-Measure values to evaluate the performance of the model.

7 Results

Test results of the NLP-DRP models on retrieving data and processing, including True Positive, False Positive, False Negative, and True Negative, are shown in Table 7. The Accuracy, Precision, Recall, and F-measure values are shown in Table 8.

Table 7: Testing results of the NLP-DRP models

		Actual Output	
		True	**False**
Predited Output	**True**	True Positive 481	False Positive 5
	False	False Negative 14	True Negative 0

Table 8: Performance evaluation values of the NLP-DRP model

Measurements	Values
Accuracy	0.96
Precision	0.99
Recall	0.97
F-measure	0.98

8 Conclusions

From the results in the previous section, it was found that the performance of the NLP-DRP model is very good and covers the use of words, sentence patterns, processing functions, and various retrieval conditions. The operation of the Semantic patterns, Ontology, and Fuzzy system allows NLP-DRP to analyze the meaning of complex query sentences in all levels, including words, sentences, and query conditions that have never been presented in any previous studies. New features, such as inferring of the word "ติดกัน [tìt-kan] (adjacent)" to provide all neighbor provinces around a given name, processing retrieved data items by non-SQL functions, and Fuzzy system, etc. Outstanding results show, especially in Thai, a non-segmentation language, in which the difficulty in word segmentation may easily cause errors in the Lexical analysis process. Such good performances are the result of the improvement of the word segmentation by TLS-ART, in which the outputs showed all performance values, including Accuracy, Precision, Recall, and F-measure higher than 0.9. However, we also found some issues that should be addressed as follows:

In terms of word segmentation, although the Ranking Trie algorithms can reduce the size of the dictionary and the number of parsing tasks, there still

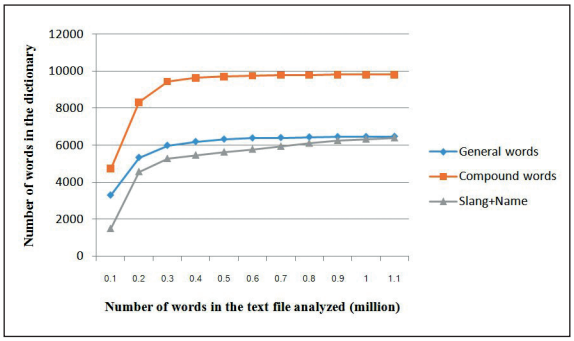

Figure 18: Trend of new words occurrence.

have an interesting question: how many suitable and sufficient words should be stored in the dictionary for use with other jobs? To get the answer for further research, the researcher expanded the scope of data to cover more content types, including economy, society, politics, health, education, agriculture, entertainment, sports, technology and IT, and others. With a total of 6,579 files from a variety of sources of data with both online and offline. These text files were analyzed to find the characteristics of new words (words that do not exist in the dictionary) and trends of its' occurrence. The result showed that most of the new words are compound words, abbreviations, specific names, slang, and words spelled for foreign terms. These words have a relatively high rate of increase and are not suitable to be stored in the dictionary due to the number of occurrences is too high and cannot be predicted as the results shown in Figure 18.

In the case of sentence analysis, although the NLP-DRP model supports most sentence patterns used in reality, some new sentence patterns are always found when tested by a new group of users. The researcher found that, in some cases, although the sentence lack grammar integrity, it is still read and analyzed the correct meaning by humans. Therefore, this issue should be further researched to developed the natural language processing algorithms that can analyze sentences in this manner by machine learning with no longer analyzed by humans.

Acknowledgments

Thank you to the National Electronics and Computer Technology Center for the Orchid Corpus that has been distributed and used in the research.

References

[1] A. Shah, J. Pareek, H. Patel, and N. Panchal, "NLKBIDB - Natural language and keyword based interface to database," in *Proceedings of the International Conference on Advances in Computing, Communications and Informatics*, 2013, pp. 1569–1576.

[2] Y. Thairatananond, "Towards the design of a Thai text syllable analyzer," M.S. thesis, Asian Institute of Technology, Bangkok, 1981.

[3] S. Chanyapornpong, "A Thai syllable separation algorithm," M.S. thesis, Asean Institute of Technology, Bangkok, 1983.

[4] Y. Poowarawan, "Dictionary-based Thai syllable separation," in *Proceeding of the 9th Electrical Engineering Conference*, 1986, pp. 167–175.

[5] S. Raruenrom, "Word segmentation by dictionary," Senior Project Report, Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok, Thailand, 1991 (in Thai).

[6] V. Sornlertlamvanich, "Word segmentation for Thai in machine translation system," National Electronics and Computer Technology Center, Bangkok, Thailand, 1993 (in Thai).

[7] National Electronics and Computer Technology Center, "Thai Lexeme Tokenizer : LexTo," 2019. [Online]. Available: http://www.sansarn.com/lexto/

[8] P. Chaloenpomsawat, "Feature-Based Thai word segmentation," M.S. thesis, Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok, Thailand, 1998 (in Thai).

[9] A. Kawtrakul, C. Thumkanon, and S. Seriburi "A statistical approach to Thai word filtering," in *Proceedings of the 2nd Symposium on Natural Language Processing*, 1997, pp. 398–406.

[10] C. Haruechaiyasak, S. Kongyoung, and M. Dailey, "A comparative study on Thai word segmentation approaches," in *Proceedings of the 5th International Conference ECTI-CON*, 2008, pp. 125–128.

[11] W. A. Woods, R. M. Kaplan, and B. N. Webber, "The lunar sciences natural language information system: Final report," BBN Report 2378, 1972.

[12] E. D. Sacerdoti, "Language access to distributed data with error recovery, knowledge," in *Proceedings of the 5th International Joint Conferences on Artificial Intelligence*, 1977, pp. 196–202.

[13] J. K. Jia, Y. B. Shao, H. Long, and Q. Z. Du, "A natural language sentence analysis algorithm based on word order modifier syntax rules," *Procedia Computer Science*, vol. 166, pp. 496–500, Jan. 2020.

[14] A. Rajendra and J. Manish, "Natural language interface using shallow parsing," *International Journal of Computer Science and Application*, vol. 5, no. 3, pp. 70–90, 2008.

[15] F. B. Thompson, P. C. Lockemann, B. Dostert, and R. S. Deverill, "REL: A rapidly extensible language system," in *Proceedings of the International Conference on Computational Linguistic*, 1969, pp. 399–417.

[16] D. L. Waltz, "An english language question answering system for a large relational database," *Communications of the ACM*, vol. 21, no. 7, pp. 526–539, 1978.

[17] G. G. Hendrix, E. D. Sacerdoti, D. Sagalowicz, and J. Slocum, "Developing a natural language interface to complex data," *ACM Transactions on Database Systems*, vol. 3, no. 2, pp. 105–147, 1978.

[18] A. Gupta, A. Akula, D. Malladi, P. Kukkadapu, V. Ainavolu, and R. Sangal, "A novel approach towards building a portable nlidb system using the computational paninian grammar framework," in *Proceedings of the International Conference on Asian Language Processing*, 2012, pp. 93–96.

[19] W3C, "Resource Description Framework (RDF) : Concepts and Abstract Syntax: 2014," 2019. [Online]. Available: http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/

[20] W3C, "RDFa Core 1.1 - Third Edition: 2015," 2019. [Online]. Available: https://www. w3.org/TR/2015/REC-rdfa-core-20150317/

[21] W3C, "SPARQL 1.1 Update: 2013," 2019. [Online]. Available: https://www.w3.org/TR/2013/REC-sparql11-update-20130321/

[22] W3C, "Web Ontology Language (OWL): Overview:2012," 2019. [Online]. Available: https://www.w3.org/OWL/

[23] W3C, "OWL 2 Web Ontology Language: Document Overview (Second Edition):2012," 2019. [Online]. Available: https://www.w3.org/TR/owl2-overview/

[24] S. Jinxing, "Ontology-based semantic information retrieval for enterprise management information system," in *Proceedings of the 6th IEEE Joint International Information Technology and Artificial Intelligence Conference*, 2011, pp. 409–412.

[25] A. Mittal, J. Sen, D. Saha, and K. Sankaranarayanan, "An Ontology based dialog interface to database," in *Proceedings of the International Conference on Management of Data*, 2018, pp. 1749–1752.

[26] F. Ramli, S. A. Noah, and T. B. Kurniawan, "Ontology-based information retrieval for historical documents," in *Proceedings of the Third International Conference on Information Retrieval and Knowledge Management (CAMP)*, 2016, pp. 55–59.

[27] P. Smith, *Applied Data Structures with C++*. Massachusetts: Jones and Bartlett publisher, 2004, pp. 253–273.

[28] C. Tapsai, P. Meesad, and C. Haruechaiyasak, "TLS-ART: Thai language segmentation by automatic ranking trie," presented at the 9th International Conference Autonomous Systems, Cala Millor, Spain, Oct. 23–28, 2016.

[29] H. J. Zimmermann, *Fuzzy Set Theory and Its Applications*, 4th ed. New York: Springer Seience+Business Media, 2001.

[30] P. Meesad, *Fuzzy System and Artificial Neural Network*. Bangkok, Thailand: KMUTNB Textbook Publishing Center, 2012 (in Thai).

[31] U. Sillapasarn, *Thai Language Principle*. Bangkok, Thailand: Thai Wattanapanich Publishing, 1990 (in Thai).

[32] K. Thonglor, *Thai Language Principle*. Bangkok, Thailand: Amorn Printing, 1996 (in Thai).

[33] Meteorological Department of Thailand, "Meteorological knowledge: Distribution criteria of rain," 2019. [Online]. Available: https://www.tmd.go.th/info/info.php?FileID=29

[34] Royal Society of Thailand, *Royal Institute Dictionary 2011*. Bangkok, Thailand: Nanmeebooks Publication, 2013.

[35] Royal Society of Thailand, "Royal Institute Dictionary 2011," 2019. [Online]. Available: http://www.royin.go.th/dictionary/

[36] Digital Government Development Agency, "High Value Dataset: Hourly rainfall data 2012–2014," 2019. [Online]. Available: https://data.go.th/Datasets.aspx?kw=ฝน

[37] V. Sornlertlamvanich, N. Takahashi, and H. Isahara, "Thai Part-of-speech Tagged Corpus: ORCHID," 2019. [Online]. Available: https://www.researchgate.net/publication/243783378_Thai_Part-of-speech_Tagged_Corpus_ ORCHID