NATURAL LANGUAGE DATABASE MANAGEMENT SYSTEM FOR DATA MANIPULATION.

ChalermpolTapsai[*] &PhayungMeesad[**]
[*]College of Innovation and Management, SuanSunandha Rajabhat University, Thailand:
[**]King Mongkut's University of Technology North Bangkok, Thailand
Email: [*]chalermpol.ta@ssru.ac.th, [**]phayung.m@it.kmutnb.ac.th

ABSTRACT

At present, Natural Language Processing is one of the most popular research topics applied in various fields. In the case of data management, Natural Language Processing plays an essential role in the interfacing between human and database, which help general users who lack the technical knowledge to be able to perform many operations on databases more conveniently. However, most of the studies in this topic focused on retrieving data from the database, and still lacking studies related to data manipulation in other ways. In this research, we propose a new algorithm that learned Natural Language sentences from a Learning dataset and created Semantic Patterns used to analyze the meaning as well as execute the data manipulation commands covering, insertion, editing, and deletion. The performance evaluation was performed by 500 Natural Language Commands from a sample of 50 users. The results show very high performance with the values of Precision, Recall, and F-measure more than 0.9.

Keywords: Natural Language Processing, data manipulation, database, Thai language.

INTRODUCTION

Natural Language Processing (NLP) has become a major role in helping humans to use computers more conveniently. In general, when humans want to use computers, they have to create a program in a computer language to instruct the computer to process the data for the expected results. In the case of data management by computer, a database is an effective system that widely used in most organizations due to flexibility in operating and retrieving data. However, as mention above, the data management in a database needs a program written by SQL language; this is a major obstacle for general users who lack knowledge about database and SQL language. Until now, though many studies have presented NLP algorithms for human-database interfaces, most of them mainly focused on retrieving data from the database. Examples of these studies are [1,] [2,] [3], [4]. For our knowledge, there is only a study related to the database creation using natural language [5], while database manipulation, including data insertion, editing, and deleting, are still not presented. For this reason, we are interested in database manipulation with natural language, covering all three types of commands as above, in which the details will be presented in the next section

The SQL commands [6] used to manage data in the database divided into three types, which are:

1) Data Definition Language (DDL), the commands used to define the structure of the database, including CREATE, DROP, ALTER, and TRUNCATE.

2)Data Manipulation Language (DML), the commands used to insert, edit, delete, and query data, including INSERT, UPDATE, DELETE, and SELECT.

3) Data Control Language (DCL), the commands used to grant or revoke privileges to access data, including GRANT and REVOKE.

In this research, we focused on only three DML commands, including INSERT, UPDATE, and DELETE, which are never presented in other studies.

METHODOLOGY

This research is to create an NLP model named "Natural Language Processing for Database Manipulation (NLP-DM)"to manipulate data in the database covering four commands, including INSERT, UPDATE, DELETE, and SELECT. The research process divided into four steps, which are Data Collection,Model development, and Model evaluation.

1. Data Collection

In this step, the data used in this research divided into two parts, which are Experimental data, and the Natural Language Command dataset.

The experimental dataset is an hourly rainfall quantity measured in 73 provinces disseminated by the Digital Government Development Office [7] covering three years of data, including 2012, 2013, and 2014. This dataset was taken through the pre-processing for data cleaning and transform into a database by the CSV Data Processing Model (CSVDPM) model [5]. This database consists of three tables, namely RainQuantity, Station, and Province, with each detail structure, as shown in Tables1, 2, and 3.

Table 1. Detail structure of the table RainQuantity

Field name	Type (length)	Detail
rain_date	Date	Date
station_id	Text (7)	Station ID number
rain_quantity	Float	Rain amount
rain_time	Integer	Rain time (value 0 - 23)

Table 2. Detail structure of the table Station

Field name	Type (length)	Detail
station_id	Text (7)	Station ID number
station_name	Text (50)	Station name
station_address	Text (150)	Station address
province_id	Text (3)	Province ID number

Table 3. Detail structure of the table Province

Field name	Type (length)	Detail
province_id	Text (3)	Province ID number
province_name	Text (50)	Province name

Natural Language Command (NLC) dataset consists of NLCs, which collected from a sample of 50 people, with each person writing 20 sentences, covering commands to Select, Insert, Update, and Delete data based on the experimental database. Each command can be written in the form of Simple, Compound, or Complex sentences. Examples of each type of NLC shown in Figure 1.

- Add the amount of rainfall at Doi Tao Station on September 5, 2014, at 9.00 o'clock, with 18 millimeters of rainfall.
- Add the amount of rain at Doi Tao Station at 10 and 11 o'clock on September 5, 2014, with the amount of rain 20 and 25 millimeters.
- Edit the amount of rain at Doi Tao Station that fell on September 5, 2014, at 9.00 o'clock to 22.5 millimeters.
- Delete the rain data at Doi Tao Station on September 5, 2014, at 9.00 o'clock.

Figure 1. Example of Natural Language Commands

With a total of 1000 commands, this dataset divided into two subsets, which are:

- Learning set that consists of 500 sentences used for developing the model
- Test set that consists of 500 sentences used model evaluation.

2. Model development

This stepis to develop theNLP-DM model, which divided into four modules, namely Lexical Analysis, Semantic Analysis. User Interface, and Command Operation, as shown in Figure 2.

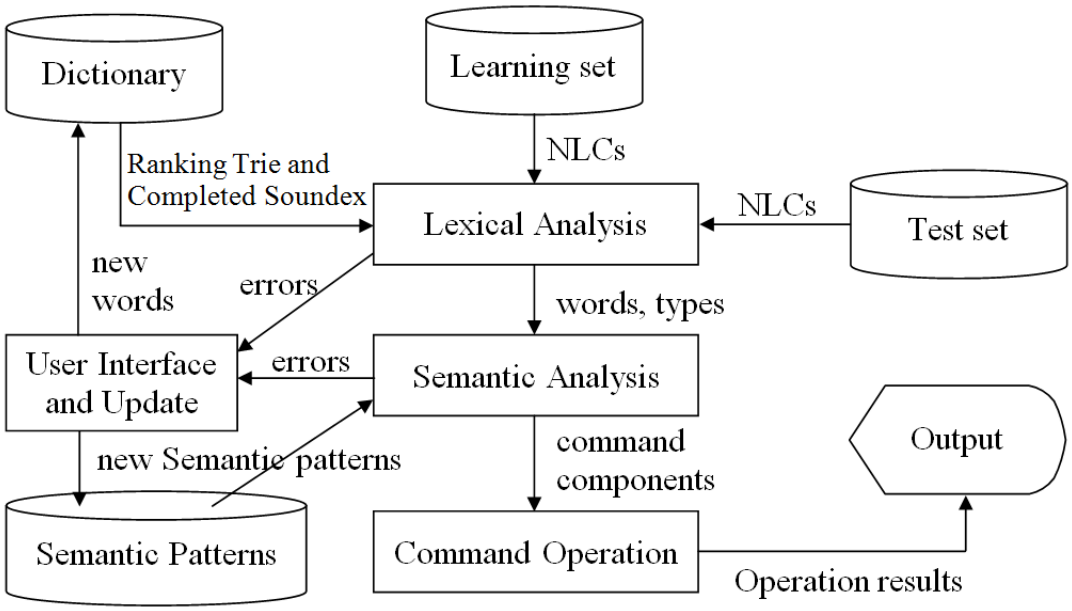

Figure 2. The NLP-DM modules

The first module, Lexical Analysis,is a module that received NLCs to segmented into words as well as specify the type of each word. In this research, we use TLS-ART-MC[8], a high-performance word segmentation model that uses Ranking Trie techniques to increase the segmentation efficiency and use Completed Soundex to correct misspelling words. The results of this module, which are words with words type, will be sent to process in the Semantic Analysis.

The second module, Semantic Analysis,used Pattern-matching technique [9] by parsing all words and words type derived from Lexical analysis with Semantic Patterns to define the meaning of the inputted NLCs and create essential components output to the Command Operation module for further process.

The Command Operation module used essential components to create SQL Data Manipulation Commands and execute to provide the output for users.

In the case of errors, such as unknown words or sentence patterns, all unknown items will be sent to the User Interface module to create a message, which guide users to add new words to the dictionary or edit the NLCs for Lexical Analysis again.

After finish creates all modules, all NLCs from the Learning set were inputted into the model for sentence learning with experts helping to analyze sentences and create the Semantic Pattern related to each sentence pattern.

3. Model evaluation

The evaluation of model performance was conducted by inputting all NLCs from the Test set to process and collect all outputs for the calculation of performance values, including Precision, Recall, and F-measure.

RESULTS

The test results of the model evaluation with 500 NLCs shown in Table 4.

Table 4. Test results of the model evaluation

		Actual output	
		True	**False**
Predicted output	**True**	True Positive 436	False Positive 5
	False	False Negative 13	True Negative 46

The efficiency values of the model were calculated and shown that Precision, Recall, and F-measure are 0.99, 0.97, and 0.98, respectively.

CONCLUSION AND DISCUSSION

According to the research results, although the results are high performance, however, due to the small amount of data used in the experiment and the limited scope of data related to only one field. Therefore, the application of this model to other fields may found different specific vocabularies and sentence patterns that may cause incorrect semantics. The solution to this problem should develop an algorithm, which allows the model to learn a large number of sentences covering many areas without having to rely on experts to make the model be able to support most sentence patterns and vocabularies for more widely use.

ACKNOWLEDGEMENTS

The author would like to thank the Research and Development Institute, SuanSunandhaRajabhat University, Bangkok, Thailand, for financial support.

REFERENCES

[1] Llopis, M. and Ferrández, A.(2013). How to make a natural language interface to query databases accessible to everyone: An example. *Computer Standards & Interfaces*, pp. 470-481.

[2] Wudaru, V., et al. (2019). Question Answering on Structured Data using NLIDB Approach. *2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS).*

[3] Srirampur, S., et al. (2014). Concepts identification of an NL query in NLIDB systems. *2014 International Conference on Asian Language Processing (IALP).*

[4] Reinaldha, F. and Widagdo, T. E. (2014). Natural Language Interfaces to Database (NLIDB): Question handling and unit conversion. *2014 International Conference on Data and Software Engineering (ICODSE).*

[5] Tapsai, C. (2018). Information Processing and Retrieval from CSV File by Natural Language:*2018 IEEE 3rd International Conference on Communication and Information Systems, Singapore*, pp. 212-216.

[6] Date, C. J. (2000).*An Introduction to Database Systems* (7th ed.). Massachusetts, USA: Addison-Wesley, pp. 83–99.

[7] Digital Government Development Agency. (2019). *High Value Dataset: Hourly rainfall data 2012-2014.*Retrieved from https://data.go.th/Datasets.aspx?kw=ฝน

[8] Tapsai, C., Meesad, P., &Haruechaiyasak, C. (2019). Thai Language Segmentation by Automatic Ranking Trie with Misspelling Correction. *Paper presented at The Autonomous Systems 2019, Cala Millor, Spain*, pp. 121-134.

[9] Androutsopoulos, I., et al. (1995). Natural Language Interfaces to Databases - An Introduction. *Journal of Natural Language Engineering.* N.P.: Cambridge University Press, 1, pp. 29-81.