Database Creation by Natural Language Processing

Chalermpol Tapsai
Faculty of Information Technology
King Mongkut's University of Technology North Bangkok
Bangkok, Thailand
chalermpol.t@email.kmutnb.ac.th

Phayung Meesad
Faculty of Information Technology
King Mongkut's University of Technology North Bangkok
Bangkok, Thailand
phayung.m@it.kmutnb.ac.th

Choochart Haruechaiyasak
National Electronics and Computer Technology Center
Pathumthani, Thailand
Choochart.Haruechaiyasak@nectec.or.th

***Abstract*— The aim of this research is to present a new model of database creation by natural language. This will allow users, who lack the technical knowledge and skills, to be able to create their own databases without having to practice or learn additional languages. By using a variety of techniques, including the analysis of natural language sentences at the level of words, phrases, and sentences. In addition, semantic patterns and ontology are also used to analyze and specify the structure of the data according to the user requirements. Evaluation of the model is conducted by the 30 samples including students and lecturers by inputting natural language description of data into the model to command the computer for database creation. The results showed that this model can support a variety of sentence syntaxes, and create databases that meet the requirements of users with very high accuracy.**

Keywords*— *Database, Creation, Natural language processing, Semantic patterns, Ontology.

I. Introduction

A database is an important source of information that helps organizations manage their work efficiently. However, the creation and management of the database require technical knowledge and skills. Moreover, the specific language, SQL is needed for database administration [1], therefore unskilled users cannot create or access the database. Though numerous studies related to Natural Language Processing have been conducted to allow users interfaced with computers by human languages in various topics e.g., text summary [2], document analysis [3], [4], language translation [5], and interface with database[6]. However, in the case of interface with database, most of these studies focus on data retrieval as in [7] without any study directly related to database creation. For this reason, the researchers are interested in developing the new model called Database Creation by Natural Language Processing (DCNLP). This will help users who lack technical knowledge and skills to create their own database easier. By using many techniques, including lexical analysis, phrase analysis, semantic pattern parsing and ontology, The DCNLP model is designed to be able to support the natural language sentences in a variety of syntax that corresponds to the actual usage.

II. Research Methodology

As shown in Fig. 1, there are 3 steps in this research: data collection, model development, and model evaluation.

A. Data collection

In the first step, the researcher collected 100 natural language descriptions of data which are required to store in the computer system by 50 experimental samples including students and lecturers.

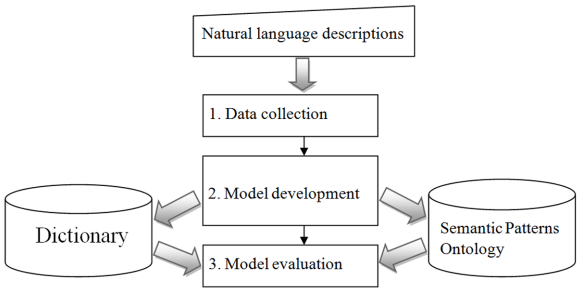

Fig. 1. Example of a figure caption.

B. Model development

The second step, all descriptions were used as a learning dataset which each description was analyzed for keywords, sentence patterns to create a dictionary, Rules and Semantic patterns which are the important parts used by the model to analyzing the sentence meaning and specify the structure of the database according to user specification.

C. Model evaluation

The experimental group consisted of 30 participants input 60 natural language descriptions of data into the model to analyze and create databases. Then the accuracy of the database structure was evaluated according to these descriptions.

III. Processing of the Model

In Fig. 2. There are 6 steps in the DCNLP processing.

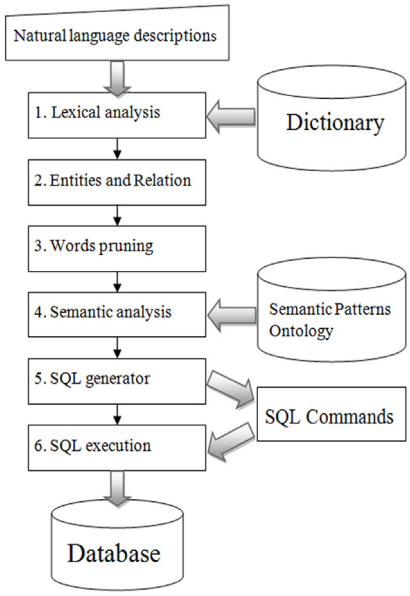

Fig. 2. Processing steps of the DCNLP model

978-616-368-065-5 ©2018 JSCI

A. *Lexical analysis*

In this step, each natural language description that is inputted by the user is analyzed to separate into words and specify the type of words using the TLS-ART[8].

B. *Entities and Relation analysis*

In this step, all words derived from Step 1 were analyzed to identify the key elements which may be used as Entities or Relation i.e., nouns, noun-phrases, verbs, and verb-phrases. noun-phrases and verb-phrases are words that are made up of several types of words, as shown in Fig. 3.

- รหัสนักศึกษา(Noun) + นักศึกษา(Noun) = รหัสนักศึกษา(Noun)
- คะแนน(Noun) + สอบ(Verb) = คะแนนสอบ(Noun)
- วันที่(noun) + ซื้อ(Verb) + สินค้า(Noun) = วันที่ซื้อสินค้า(Noun)
- ลง(Verb) + ทะเบียน(noun) = ลงทะเบียน(Verb)
- การ(Noun) + เรียน(Verb) = การเรียน(Noun)

Fig. 3. Types of noun-phrases and verb-phrases

C. *Words pruning*

In this step, unnecessary, and redundancy words are eliminated.

D. *Semantic analysis*

In this step, all remaining words are parsed to the semantic pattern in the form of Nondeterministic Finite Automaton with output as shown in Fig. 4 to get the results as the entities, relations, and attributes.

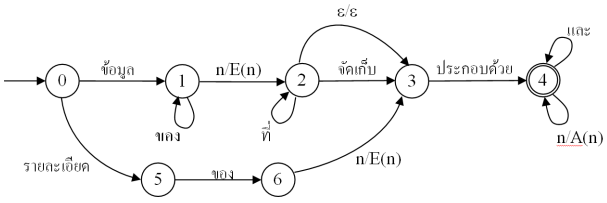

Fig. 4. The semantic pattern in the form of Nondeterministic Finite Automaton with an output

E. *SQL generator*

In this step, the entities, relations, and attributes are mapped to tables and fields and then the SQL commands are created.

F. *SQL execution*

For this step, all SQL commands are executed to create a database.

IV. Experimental result

In the experiment, 60 natural language descriptions are inputted into the model to evaluate the performance. The result showed that the DCNLP model was able to analyze natural language descriptions and create databases with a very high accuracy of 91.67%.

V. Conclusion and discussion

Despite a very high accuracy in the model evaluation, the errors remain a serious problem in database creation. There is 3 main causes of errors: (1) typo error, (2) used of unknown words, and (3) used of unknown sentence syntaxes. Therefore, to improve the efficiency of the model, the number of samples tested should be increased and analyzed for unknown words and more sentence syntaxes to be added into the Model.

References

[1] C. J. Date, An Introduction to Database Systems, 7th ed. Addison-Wesley, Massachusetts: USA, 2000, pp. 83–98.

[2] O. M. Foong, S. P. Yong, and F.A. Jaid, “Text Summarization using Latent Semantic Analysis Model in Mobile Android Platform,” in 9th Modelling symposium, 2015, pp.35-39.

[3] F. Agung, “Software Requirements Specification Analysis Using Natural Language Processing Technique,” IEEE Quality in Research, 2013.

[4] A. S. Hussein, “Visualizing Document Similarity Using N-Grams and Latent Semantic Analysis,” in SAI Computing Conference, London, UK, 2016, pp. 269-279.

[5] D. Moussallem, M. Wauer, A. N. Ngomo, Machine Translation using Semantic Web Technologies: A Survey, Journal of Web Semantics, Volume 51, 2018, pp. 1-19.

[6] M. Llopis, and A. Ferrández, “How to make a natural language interface to query databases accessible to everyone: An example. Computer Standards & Interfaces,”, 2013, pp. 470-481. doi: http://dx.doi.org/10.1016/j.csi.2012.09.005.

[7] A. Shah, J. Pareek, H. Patel, and N. Panchal, “NLKBIDB - Natural language and keyword-based interface to database,” Paper presented at the Advances in Computing, Communications and Informatics (ICACCI), 2013 International Conference on, pp. 1569-1576.

[8] C. Tapsai, P. Meesad, and C. Haruechaiyasak, “Thai Language Segmentation by Automatic Ranking Trie, ” in Proceedings of 9th International Conference Autonomous Systems (AutoSys 2016), Spain.