Indonesian Text Translator into Database Structured Query Language with Multi Parameters using Natural Language Processing

G Hermawan¹, I Faturohman², N Isharmawan³
Faculty of Engineering and Computer Science, Universitas Komputer Indonesia, Indonesia

Email: galih.hermawan@email.unikom.ac.id

Abstract. In this paper, we present the results of our research on SQL (Structured Query Language) Translator in Indonesian which aims to measure how accurate the designed system in understanding query sentences in Indonesian into SQL query languages. The methods used include pre-processing sentences, stemming, identifying keywords, until the final evaluation of sentences. We prepared 18 sentences divided into several categories as test data. The test results show that the system successfully evaluates all sentences with a total accuracy of 72.22% to 83.87%. Accuracy values were obtained by comparing similarities between systems generated queries and key answers query. In general, the system has been able to translate most of the Indonesian sentences into SQL queries.

1. Introduction
A database is a media used to manage data. One way to access data in a database is to use a command called SQL (Structured Query Language). SQL commands already have the default standard setting, but the syntax generally depends on the DBMS (Database Management Systems) used. DBMS itself is an engine or software used to manage databases. Processing the structure and data in the DBMS is also divided into 2, namely DDL (Data Definition Language) and DML (Data Manipulation Language). DDL is a SQL syntax structure that focuses on building a container for information. The DML is a SQL syntax structure that focuses on data manipulation, such as storing, changing, deleting, and displaying information [1]. The NLP (Natural Language Processing) is one of the fields of artificial intelligence that is more focused on discussing the interaction between human language and computers [2].

Currently, there are several studies related to the use of natural language as SQL Translator media. Most current research uses English as its natural language. Ghosh P K uses certain words to distinguish between keywords and not keywords in SQL [3]. Basic level queries can be done well. The system only handles query data on one table. The model built by Sathick K J has been able to work on complete DML queries, at the basic level [4]. Includes select, update, delete. As for Kaur P, he combined basic level queries with the translated text through voice [5]. Singh G in his research was able to present user-friendly interfaces, dynamic table data, by accommodating basic level queries [6]. Research conducted by Sukthankar N is able to evaluate up to complex query levels, although in some cases with similar sentences there are still failures [7]. Utama P is able to handle basic queries to complex queries, but there are still difficulties when involving data in multiple tables [8]. SQL Translator research in Indonesian has also been conducted by several researchers. Research conducted by Lisangan E A [9] can evaluate queries that are limited to certain databases. The system can only handle basic level queries. The study conducted by Raharjo S only focused on querying specific data in the Qur'an[10]. Karlos J
produces a system that can evaluate basic level queries with a statistical translator engine approach [11]. The results of the research on the instruction sentences were obtained with an accuracy value of 70% - 80%.

In this study, we use MySQL database software as a reference in evaluating the SQL syntax. The focus of the discussion is only on select queries, which aim to extract information in the database. The command sentence to be tested is classified into six types with various parameters, namely queries without conditions, queries with conditions, queries with many conditions, queries involving many tables without conditions, querying many tables with many conditions, and queries for sorting data. The total command sentence has 18 sentences, which have been classified according to the type specified. Next, each sentence will be evaluated by the keys and answers, and to measure the value of total accuracy.

2. Research Methods
In this study, we used descriptive research methods. In this case, we tried to describe the stages of research related to the specific problem [12] of SQL Translator in Indonesian. The stages of system development research generally include database analysis, sentence processing, sentence translation, and output evaluation. The process stages or general description of the system can be seen in Figure 1.

Figure 1. General System Description

Figure 1 contains an overview of the system that contains the stages in translating natural languages to form a SQL query.
3. Results and Discussion

The method of translating commands in natural language into the SQL (Structured Query Language) syntax has several stages including, analyzing the database to get table info, attributes, and relationships between tables, then adding the word also known as in each table and attribute to enrich the keyword of the table name and its attributes. The next stage of the translation process, before translated sentence will be through the process of the pre-processing ie, tokenizing stage to break the sentence, then stemming stage to remove the affixed word, and also known as stage to change the word also known as used in the sentence to be the word in question. After that is done the translation process is divided into 5 processes namely, the process of identification of commands, identification tables relationships, attribute identification, feature identification, and query preparation process.

3.1. Input Command Sentences

In the process of extraction of information, input data in the form of Indonesian command sentences entered by the user. Example of the following command input sentence: "Tampilkan judul film yang termasuk kategori comedy, dimana dalam deskripsi film tersebut mengandung kata 'dentist'" (Show movie titles that are included in the comedy category, where in the film description contains the word 'dentist'). In this case, we use the sakila sampel database [13].

3.2. Database Analysis

There are 3 processes in database analysis block, that is tables and attributes analysis, relation analysis, and also known as assignment. Analysis of tables and attributes is the process of identifying all table names and attributes that exist in the database. Relationship analysis is the process of searching which tables are related. Also known as assignment is the process by which users can register the word also known as of the table name or attribute that has been previously identified.

We can see some examples of tables and attributes analysis in Table 1.

| No | Attributes | Name of Tables/Attributes | Type                |
|----|------------|---------------------------|---------------------|
| 1  | Actor      |                           | Base Table          |
| 2  | 1          | actor_id                 | smallint(5) unsigned|
| 3  | 2          | first_name               | varchar(45)         |
| 4  | 3          | last_name                | varchar(45)         |
| ...| ...        | ...                      | ...                 |
| 23 | 4          | last_update             | timestamp           |

While in Table 2, we can see examples of table relationship analysis.

| No | Relationship Name | Foreign Key            | References          |
|----|-------------------|------------------------|---------------------|
| 1  | fk_address_city   | address.city_id        | city.city_id        |
| ...| ...               | ...                    | ...                 |
| 22 | fk_store_staff    | store.manager_staff_id | store.manager_staff_id|

3.3. Pre-processing Sentences

This pre-processing stage aims to facilitate the application of rule-based methods in solving story problems. In the second block pre-processing sentences there are 3 processes namely, Tokenizing,
Stemming and Identification of the word also known as. Tokenizing is breaking sentences into word tokens. Stemming is the basic word search process of every tokenizing token. The Also known as word identification process is looking for the also known as words listed to be replaced by the table name or attribute in question. Examples of tokenizing results can be seen in Table 3.

**Table 3. Tokenizing Results**

| Name | Object |
|------|--------|
| Tampilkan judul film yang termasuk kategori comedy, dimana dalam deskripsi film tersebut mengandung kata "dentist" | |

**3.4. Translating Sentences**
The result of the preprocessing stage of the sentence is a token list that is ready to be processed at the translation phase of the command line. This stage applies a rule-based system to get information from the preprocessing process and process it in such a way that can translate the sentence of the command. In this process block, there is 5 subprocess, that is. Identify keyword commands, Identify table names, identify attribute names, identify condition conditions, and construct queries. Identification command is the process of searching for command words in tokens. Identify table names in the process of finding table names of tokens. Identify the attribute name is the process of finding what attributes to display. Identification of condition conditions is the process of looking for whether there are any signs about using the facility Where conditions. Preparation of the query is a process that collects all the identification results and compiles them in Syntax SQL.

**3.5. Query Testing Results**
In this last process is to test the results that have been translated into the SQL syntax to be executed syntax and views the data into output. If the translation query is invalid, it will display an error message.

**3.6. Final Test Results**
Testing sentences are divided into six categories, including
1. queries without conditions
2. queries with conditions
3. queries with many conditions
4. queries many tables without conditions
5. queries many tables with conditions
6. queries by sorting data

There are 18 sentences in total. The data material is taken from a combination of Academic and Sakila databases. The test results can be seen in Table 4.
IOP Conf. Series: Materials Science and Engineering 662 (2019) 022095  doi:10.1088/1757-899X/662/2/022095

Table 4. Testing Results.

| No | Sentence | Category | Query Result | Result |
|----|----------|----------|--------------|--------|
| 1  | Tampilkan mahasiswa | 1 | SELECT * FROM 'mahasiswa' ; | Correct |
| 2  | Tampilkan nama mahasiswa | 1 | SELECT 'mahasiswa'.nama FROM 'mahasiswa' ; | Correct |
| 3  | Lihat nilai uas dan uts sebutkan nama, alamat dan kota mahasiswa | 1 | SELECT 'nilai'.uas , 'nama' , 'alamat' , 'kota' FROM 'mahasiswa' ; Correct |
| 4  | Lihat mahasiswa yang nilai uas 80 sampai 90 | 2 | SELECT * FROM 'mahasiswa' WHERE 'mahasiswa'.uas = '80' ; | Correct |
| 5  | Tampilkan mahasiswa yang nilai uas minimal 80 dan uas maksimal 90 | 6 | SELECT 'nama' FROM 'mahasiswa' ORDER BY 'uas' DESC ; | Correct |
| 6  | Lihat mahasiswa yang nilai uas minimal 80 dan uas maksimal 90 | 6 | SELECT 'nama' FROM 'mahasiswa' ORDER BY 'uas' DESC ; | Correct |
| 7  | Tampilkan mahasiswa yang nilai uas minimal 80 dan uas maksimal 90 | 6 | SELECT 'nama' FROM 'mahasiswa' ORDER BY 'uas' DESC ; | Correct |
| 8  | Tampilkan mahasiswa berdasarkan jenis kelamin | 3 | SELECT 'nama' FROM 'mahasiswa' ORDER BY 'jk' DESC ; | Correct |
| 9  | Tampilkan mahasiswa berdasarkan jenis kelamin | 3 | SELECT 'nama' FROM 'mahasiswa' ORDER BY 'jk' DESC ; | Correct |
| 10 | lihat mahasiswa yang tinggal di bandung dan namanya ahmad | 3 | SELECT * FROM 'mahasiswa' WHERE 'mahasiswa'.kota = 'di' ; | Correct |
| 11 | Tampilkan mahasiswa yang berjenis kelamin perempuan dan tinggal di bandung | 3 | SELECT * FROM 'mahasiswa' WHERE 'mahasiswa'.jk = 'perempuan' ; | Correct |
| 12 | lihat mahasiswa yang tinggal di bandung | 4 | SELECT * FROM 'mahasiswa' WHERE 'mahasiswa'.kota = 'di' ; | Correct |
| 13 | sebutkan nama mahasiswa dan nilai uas | 4 | SELECT 'mahasiswa'.nama , 'nilai' FROM 'mahasiswa' INNER JOIN 'nilai' ON 'mahasiswa'.nim = 'nilai' ; | Correct |
| 14 | sebutkan nama mahasiswa dan nilai uas | 4 | SELECT 'mahasiswa'.nama , 'nilai' FROM 'mahasiswa' INNER JOIN 'nilai' ON 'mahasiswa'.nim = 'nilai' ; | Correct |
| 15 | sebutkan nama mahasiswa dan nilai uas | 4 | SELECT 'mahasiswa'.nama , 'nilai' FROM 'mahasiswa' INNER JOIN 'nilai' ON 'mahasiswa'.nim = 'nilai' ; | Correct |
| 16 | Sebutkan mahasiswa yang tinggal di bandung | 5 | SELECT * FROM 'mahasiswa' INNER JOIN 'kelas' ON 'mahasiswa'.kota = 'di' ; | Correct |
| 17 | sebutkan mahasiswa yang tinggal di bandung | 5 | SELECT * FROM 'mahasiswa' WHERE 'mahasiswa'.kota = 'di' ; | Correct |
| 18 | Tampilkan judul film yang Termasuk kategori comedy, dimana dalam deskripsi film tersebut mengandung kata "dentist" | 5 | SELECT 'film'.title AS 'judul' FROM 'film' INNER JOIN 'film_category' ON 'film'.film_id = 'film_category'.film_id INNER JOIN 'category' ON 'film_category'.category_id = 'category'.category_id WHERE 'category'.name = 'comedy' AND 'film'.description LIKE '%dentist%'; | Correct |

The accuracy test used in this research applied the F-Measure method. F-Measure is a representation of the measure of the accuracy of a test by using the precision and recall values of the test. This test looks at whether the system can process sentences entered by the user to be translated so that it can be measured accurately with the method of F-Measure. the sentence entered has been translated manually first so it can be compared with the output of the system. The values calculated in this test are the correct, falsePositive, falseNegative, precision, recall, and F-Measure values of the system dataset iteration. Calculation F-Measure microaverage data by using total correct value, falsePositive, falseNegative overall dataset while calculation F-Measure macroaverage is by calculating average F-Measure each iteration dataset. Here is the result of iterating the entire dataset of system accuracy testing. So the system accuracy based on the test dataset above is: The accuracy value is known to De between 72.22% - 83.87%.
4. Conclusion
After analyzing, designing, and testing using black-box method and accuracy testing, it was obtained that the system can translate the command line in natural language into SQL query syntax format. Test results from several sentences obtained accuracy ranged from 72.22% - 83.87%. The results indicate that when the sentences entered by the user is incomplete it will also get the wrong query or produce the wrong information. This can result in an error in executing the query. Thus the information seeker media can directly access the source information on the database when the data in the database is changed, then the results of the displayed data will also change.

5. Acknowledgements
We would like thanks to faculty of engineering and computer science Universitas Komputer Indonesia for funding our research.

References
[1] Silberschatz, A., Korth, H. F., and Sudarshan, S. (1997). *Database system concepts*, 4. New York: McGraw-Hill.
[2] Neuhauser, L., and Kreps, G. L. (2011, March). Participatory design and artificial intelligence: Strategies to improve health communication for diverse audiences. In *2011 AAAI Spring Symposium Series*.
[3] Ghosh, P. K., Dey, S., and Sengupta, S. (2014). Automatic sql query formation from natural language query. *International Journal of Computer Applications*, 975, pp.8887.
[4] Sathick, K. J., and Jaya, A. (2015). Natural language to SQL generation for semantic knowledge extraction in social web sources. *Indian Journal of Science and Technology*, 8(1), pp.1-10.
[5] Mandal, K. P., Mukherjee, P., Chakraborty, B., and Chattopadhyay, A. (2016). A novel Bengali Language Query Processing System (BLQPS) in medical domain. *Intelligent Decision Technologies*, (Preprint), pp.1-16.
[6] Singh, G., and Solanki, A. (2016). An algorithm to transform natural language into SQL queries for relational databases. *Selforganizology*, 3(3), pp.100-116.
[7] Gao, J., Galley, M., and Li, L. (2019). Neural approaches to conversational AI. *Foundations and Trends® in Information Retrieval*, 13(2-3), pp.127-298.
[8] Utama, P., Weir, N., Basik, F., Binning, C., Cetintemel, U., Hättasch, B., ... and Usta, A. (2018). An End-to-end Neural Natural Language Interface for Databases. *arXiv preprint arXiv:1804.00401*.
[9] Lisangan, E. A. (2013). Natural Language Processing Dalam Memperoleh Informasi Akademik Mahasiswa Universitas Atma Jaya Makassar. *TEMATIKA, Journal of Informatics and Information Systems*, 1(1), pp.39-46.
[10] Raharjo, S., and Hartati, S. (2014). ANTARMUKA BAHASA ALAMI UNTUK MELAKUKAN QUERY TERHADAP TERJEMAHAN AL-QURAN. *Jurnal Teknologi*, 7(1), pp.12-19.
[11] Karlos, J., Sujaini, H., and Anra, H. (2016). Konversi Bahasa Indonesia ke SQL (Structured Query Language) dengan Pendekatan Mesin Penerjemah Statistik. *Jurnal Sistem dan Teknologi Informasi (JUSTIN)*, 4(3), pp.466-470.
[12] Hassani, H. (2017). Research methods in computer science: The challenges and issues. *arXiv preprint arXiv:1703.04080*.
[13] Yue, K. (2013). Using a Semi-Realistic Database to Support a Database Course. *Journal of Information Systems Education*, 24(4), pp327-333.