INDONESIAN TEXT TRANSLATOR INTO DML (DATA MANIPULATION LANGUAGE) WITH SUB-QUERY

Debby Meidi P. P.[1], Ken Kinanti Purnamasari[2]

[1,2] Teknik Informatika – Universitas Komputer Indonesia
Jl. Dipatiukur 112-114 Bandung 40132
Email : debbymeidi13@gmail.com[1], ken.kinanti@email.unikom.ac.id[2]

ABSTRACT

Some research that has been done on translating natural languages into query languages, is the SQL parser carried out by Nendi Isharmawan, where the research focuses on natural language processing which serves as a bridge to access data that is in the database. Another study conducted by Ihsan Faturohman, this study focuses on the select function, in the two studies this sub-query command has still not been detected. The method used in this study is the rule-based method. The main process is divided into two stages, namely preprocessing and translation. Preprocessing stage consists of folding cases, filtering, word tokensizing, stemming, and stopword removal. Translation stage consists of keyword detection, table and column detection, tokenizing commands, identification of DML commands, identification of content, and preparation of queries. Based on the results of testing of 28 command sentences consisting of 6 query combinations, an accuracy of around 82,35% was produced. Errors that occur are not handled by date search sentences and also sub-queries with more than two tables and sub-queries with nested query types. This can be handled by adding detection rules for nested sub-queries and adding date detection for searches.

Keywords: Translation, Query, Data Manipulation Language, Database, Natural Language Processing

1. INTRODUCTION

Databases or more often called databases are groups of interrelated tables, these relations can be indicated by keys from each existing table [1], to access the database bridged by a system called the DBMS (Database Management System). SQL is the most widely used standard language in the DBMS, in SQL itself is divided back into two, namely DDL (Data Definition Language) to define tables in the database and DML (Data Manipulation Language) to manipulate data or information that is in the database so that the information useful, Sub-queries are part of DML which if interpreted is a query in a query, so to get information on a table or in a very complex relational table, we can use conditions with certain sub-requests. Sub-queries themselves can make SQL commands relatively simpler when compared to join functions so that they shorten commands and can use logic functions that are easier to understand than ordinary queries[1].

Research that has been done before about the translation of natural languages into query languages is research conducted on how to translate natural languages into SQL language by paying attention to the words in the sentences that are entered without looking at the structure of the word, which later the words will be compared with the list of keywords which exists[2], but in this study only can detect ordinary queries without containing sub-queries. Another research that has been done is the SQL parser carried out by Nendi Isharmawan [3], where the research focuses on natural language processing which serves as a bridge to access data that is in the database, in this study the results of the translation of the query have not produced a sub-query. Other research that has been done is the translation of natural language into SQL language conducted by Ihsan Faturohman, this study focuses on the select function which is divided into six, namely select with conditions, without conditions, many conditions, many tables without conditions, many tables with many conditions and table order [4], in this study the sub-query command is still not detected.

Based on the explanation above, it can be seen that there is no research that can translate from Indonesian into SQL language that contains the use of sub-queries in it. In the process of translating into SQL language, it will be carried out using rule-based methods following the previous research.

2. RESEARCH CONTENT

2.1 *NLP(Natural Language Proessing)*

NLP is part of AI science which has a focus on natural language processing, where natural language is a language commonly used by humans to communicate with each other, but the language received by computers must be processed first and understood so that the intent of the user can be understood[5].

2.2 Database

Database is a group of interrelated tables, these relationships can be shown by the key of each table that exists. One database shows one data set that is used in one company or agency scope[6].

Databases are very important in information systems because the database is a data storage warehouse that can be processed further. The role of the database becomes very important because it can organize data and also avoid duplication of data or more commonly referred to as redundancies. In the database there is a database management system process or abbreviated as DBMS which is a system that allows the database admin to access data, control and maintain data efficiently. In this study the database is used to store system supporting data and is also used as a testing medium from the command of the results of translating Indonesian into the SQL language.

2.3 *Sub-query*

If the sub-query itself is interpreted as a query in the query. So to get information on a table or in a very complex relational table, we can use conditions with certain sub-requests[1].

2.4 Research Method

The research method that will be used is an experimental method wherein this method is observed under artificial conditions that are made and arranged, where manipulation of the object under study is carried out so that there is control in the study[5]. The description of the research flow can be seen in Figure 1.1 below.

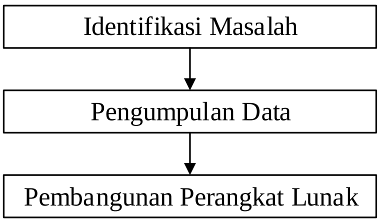


Figure 1. 1 Metode Penelitian

The following are the steps.

1. Identify Problems
   In this step an observation of the problems that existed in the previous research and also looked at examples of sub-query command cases that exist so that it can help to define the needs that will help in achieving the research conducted.

2. Data Collection
   At this stage data collection is carried out that can support research such as literature studies of electronic books, books, journals and existing case examples to determine what variables are needed to achieve the research conducted.

3. Software Development
   The method used in this stage is the prototype method, this method is used because in the process prototypes are first made and then tested, when the test results obtained do not meet the desired criteria then re-analysis can be done.

2.5 Problem Analysis

Data Manipulation Language (DML) is a SQL command method that is used to process data contents in tables such as displaying, entering, changing, deleting data contents and not related to changes in structure and definition of data types of database objects. SQL commands that are included in DML include select, update, and delete. In this study, the DML command used is only the select command.

2.6 System Overview

In this study an Indonesian language translator system will be built into DML queries. The system stage starts from the reception of natural language sentence input in Indonesian then processed at the preprocessing and translation stages. At the preprocessing stage there are several steps including folding case, filtering, word tokenizing, stemming, and removing stopword, while at the translation stage there are stages of keyword detection, table and column detection, DML command identification, content identification, and query compilation.

After the preprocessing stage generates basic word tokens that will be processed at the translation stage by detecting each word token with the initial keyword in the dictionary, after detecting the initial keyword, it is continued by comparing each word token with the predefined keyword dictionary..

If the results of comparisons between word tokens and keywords are then compared with table and column data, then the word token is identified as a DML query, then the content identification stage is a search of the remaining word tokens to fill the content in the query template that has been identified previously, and in the final stage is the preparation of the query, then the query results from the SQL token mapping process will be displayed.

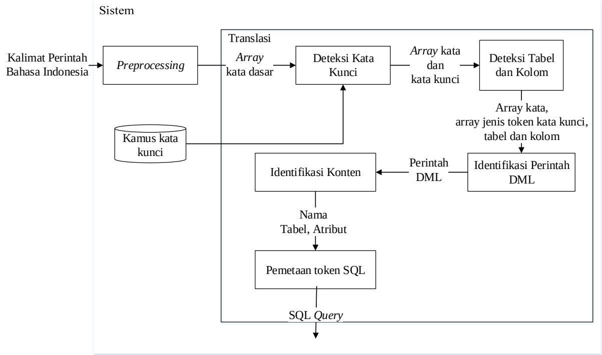


Figure 1. 2 Gambaran Sistem

2.7 Input Data Analysis

This study has input data in the form of command sentences in Indonesian that are entered by the user and output data in the form of SQL queries. The examples of input sentences that will be handled in this study can be seen in the following table.

Table 1. 1 Example Input Command

Example of DML Order in Bahasa Indonesia
Tampilkan data transaksi dengan kode_mitra yang mitranya berasal dari kota Bandung!

2.8 Preprocessing

This stage is the initial process for preparing input data in the form of Indonesian text sentences before the translation process. In many NLP studies the preprocessing process is very important to produce better accuracy[7]. This process has several stages, namely case folding, filtering, tokenizing words, stemming, and removing stopword.

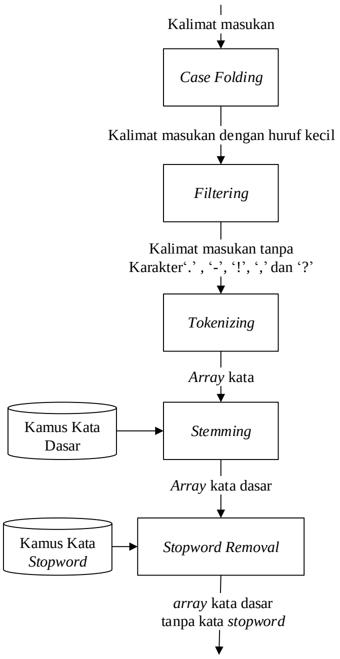


Figure 1. 3 *Proses Preprocessing*

1. Case Folding

   The case folding process is the uniformity of each letter to captal or lowercase letters [8]. In this study each letter is uniformed into lowercase letters.

Table 1. 2 Case Folding Example

Before	After
Tampilkan data transaksi dengan kode_mitra yang mitranya berasal dari kota **B**andung!	tampilkan data transaksi dengan kode_mitra yang mitranya berasal dari kota bandung!

2. *Filtering*

   Filtering is the process of sorting out each character in the input sentence which aims to reduce the character that is not considered necessary by the next process, so it can reduce noise. In this study, characters that are allowed to be input are only 'a' to 'z' and '0' to '9'.

Table 1. 3 Filtering Example

Before	After
tampilkan data transaksi dengan kode_mitra yang mitranya berasal dari kota bandung**!**	tampilkan data transaksi dengan kode_mitra yang mitranya berasal dari kota bandung

3. *Tokenizing*

   Tokenizing is the process of separating sentences into words (tokens) with a dividing parameter in the form of spaces (''). This process is done because the next process will be processing each word. The results of wordkeeping are in the form of word tokens. In this study the tokenizing process is applied so that the system can detect every word that has been separated into a set of tokens.

Table 1. 4 Tokenizing Example

Before	After	
	Indeks Token	Token Kata
tampilkan data transaksi dengan kode_mitra yang mitranya berasal dari kota bandung	1	tampilkan
	2	data
	3	transaksi
	4	dengan
	5	kode_mitra
	6	yang

Table 1.4 Tokenizing Example (extension)

	7	mitranya
	8	berasal
	9	dari
	10	kota
	11	bandung

4. Stemming

Stemming is the process of changing a word that has an additive into a basic word. In this study, the stemming process uses the library stemmer for Indonesian from literature which has adopted the Nazief & Andriani Algorithm [9]. In the Indonesian language the prefix and suffix are referred to as affixes, for example, is the suffix "-nya" as in the word "partner" which has the basic word "partner". The stemming process is carried out because the word to be detected is a basic word so it needs to be done by the process of eliminating the affix..

Table 1. 5 Stemming Example

Indeks Token	Before	After
1	tampil**kan**	tampil
2	data	data
3	transaksi	transaksi
4	dengan	dengan
5	kode_mitra	kode_mitra
6	yang	yang
7	mitra**nya**	mitra
8	**ber**asal	asal
9	dari	dari
10	kota	kota
11	bandung	bandung

5. *Stopword Removal*

Stopword removal is a process of word refinement that often appears and is considered insignificant, in a stopword language it helps to structure the sentence but does not represent any content from the contents of the sentence [10], some examples are the words' which ',' and ',' with ',' or 'and others. The stopword removal process is done to erase sentences that are considered insignificant which can result in the translation process being unfavorable or wrong.

Table 1. 6 Stopword Removal Example

Indeks Token	Before	After
1	tampil	tampil
2	**data**	
3	transaksi	transaksi
4	**dengan**	
5	kode_mitra	kode_mitra
6	**Yang**	
7	mitra	mitra
8	asal	asal
9	**Dari**	
10	Kota	kota
11	bandung	bandung

2.9 Translasi

in the translation process translation is from word tokens to the results of preprocessing into the clash of DML query commands. input sentences that have gone through the preprocessing stage in the form of DML queries. In this study, the translation process is divided into five stages, namely keyword detection, table and column detection, identification of DML commands, content identification and SQL token mapping.

1. Keyword Detection

in the process of keyword detection, each basic word token that has been obtained from the results of the preprocessing stage is done by the keyword detection process to determine the type of each token that will be used in the next process.

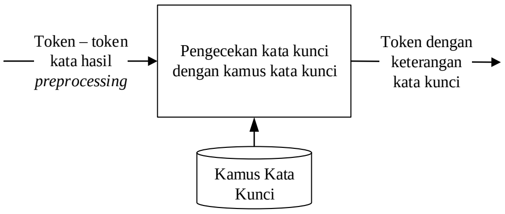

Token – token kata hasil *preprocessing*
Pengecekan kata kunci dengan kamus kata kunci
Token dengan keterangan kata kunci
Kamus Kata Kunci

Figure 1. 4 Proses keyword detection

From the above process keywords can be detected as follows.

Table 1. 7 Result keyword detection

Posisi Token	Token Kata	Jenis Token
1	tampil	Perintah
2	asal	Kondisi

2. Table and Column Detection

   In this process, each basic word token that is not included in the keyword will go through the process of detecting tables and columns to determine the type of each token that will be used in the next process.

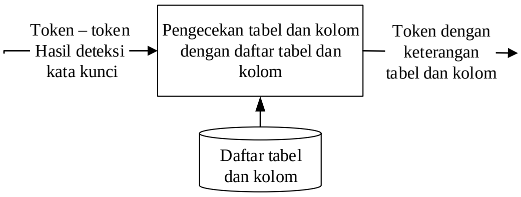


Figure 1. 5 Table and Colum Detection Process

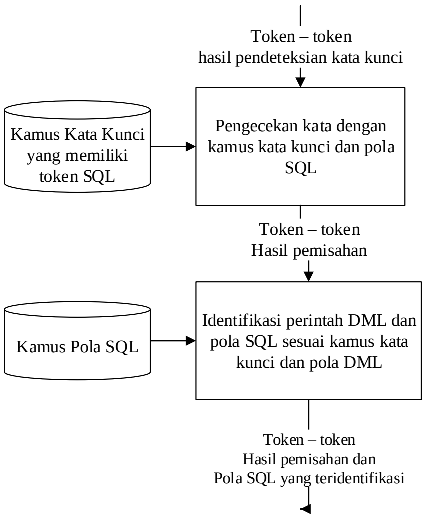
From the process of detecting tables and columns above, an array is generated which already has the type of token as follows.

Table 1. 8 Result of Table and Column Detection

Posisi Token	Token Kata	Jenis Token
1	Tampil	Perintah
2	transaksi	Tabel
3	kode_mitra	Kolom
4	mitra	Tabel
5	Asal	Kondisi
6	kota	Kolom
7	Bandung	-

3. DML Command Identification

   This process identifies all tokens that have a type of token that includes keywords such as the type of command token and the type of content conditions that are compared with the data in the database so that it can determine the appropriate DML command.


Token – token
hasil pendeteksian kata kunci
Kamus Kata Kunci yang memiliki token SQL
Pengecekan kata dengan kamus kata kunci dan pola SQL
Token – token
Hasil pemisahan
Kamus Pola SQL
Identifikasi perintah DML dan pola SQL sesuai kamus kata kunci dan pola DML
Token – token
Hasil pemisahan dan
Pola SQL yang teridentifikasi

Gambar 1. 6 DML Command Identification Proccess

From the process the comparison above obtained results as follows.

Table 1. 9 Result of DML Command Identification

Posisi Token	Token	Jenis Token	Token Dalam DML
1	tampil	**Perintah**	**SELECT**
2	transaksi	Tabel	transaksi
3	kode_mitra	Kolom	kode_mitra
4	mitra	Tabel	mitra
5	Asal	**Kondisi**	**IN**
6	kota	Kolom	kota
7	Bandung	-	bandung

After obtaining the appropriate DML command, the SQL pattern search process is then performed by comparing the arrangement of types of tokens with the data dictionary in the database so that the corresponding SQL pattern is obtained, in this process the SQL pattern found is' Commands + Tables + Columns + Tables + Conditions + Column '.

Table 1. 10 Result of Pola SQL

Pola SQL Teridentifikasi
SELECT * FROM [TABEL1] WHERE [KOLOM1] [KONDISI1] (SELECT [KOLOM1] FROM [TABEL2] WHERE [KOLOM2] = [KATA_KUNCI_PENCARIAN])

4. Konten Identification

In this stage a process is carried out to determine the content of the pattern in the previous process. Content checking is done by checking each token on the array of tokens that have been marked with the type of token, table, column and condition. In the previous process a pattern containing [table1], [table2], [column1], [column2], [condition] and [word_key_page] was found, so this process is to determine table1, table2, column1, column2, search_data_keyword and condition, which will be made.

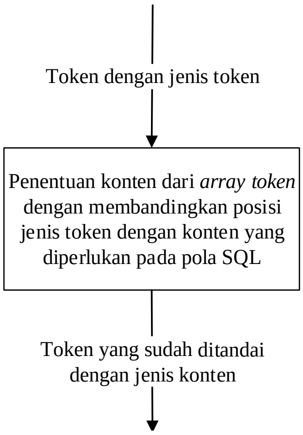

Token dengan jenis token
Penentuan konten dari *array token* dengan membandingkan posisi jenis token dengan konten yang diperlukan pada pola SQL
Token yang sudah ditandai dengan jenis konten

Figure 1. 7 Identification Content Process

Explanation of the above process can be seen as follows.

a. Determine table content for [table1] and [table2], columns for [column1] and [column2]

Table 1. 11 Identification table and column

Posisi Token	Token	Jenis Token	Jenis Konten
1	tampil	Perintah	-
2	transaksi	**Tabel**	**Tabel1**
3	kode_mitra	**Kolom**	**Kolom1**
4	mitra	**Tabel**	**Tabel2**
5	IN	Kondisi	-
6	kota	**Kolom**	**Kolom2**
7	bandung	-	-

b. Determine condition content for [kondisi]

Table 1. 12 identification content for condition

Posisi Token	Token	Jenis Token	Jenis Konten
1	tampil	Perintah	-
2	transaksi	Tabel	Tabel1
3	kode_mitra	Kolom	Kolom1
4	mitra	Tabel	Tabel2
5	IN	**Kondisi**	**Kondisi**
6	kota	Kolom	Kolom2
7	bandung	-	-

c. Determine content [kata_kunci_pencarian]

Tabel 1. 13 Identification Content for [kata_kunci]

Posisi Token	Token	Jenis Token	Jenis Konten
1	tampil	Perintah	-
2	transaksi	Tabel	Tabel1
3	kode_mitra	Kolom	Kolom1
4	mitra	Tabel	Tabel2
5	IN	Kondisi	Kondisi
6	kota	Kolom	Kolom2
7	bandung	-	**kata_kunci_pencarian**

5. SQL Token Maping

At this stage the token mapping process is carried out in accordance with the results of the process of identifying DML commands and content identification.

Table 1. 14 SQL Token Maping

Jenis Konten	Konten	Sebelum	Sesudah
		Pola SQL	**Pola SQL**
Tabel1	transaksi	SELECT * FROM [TABEL1] WHERE [KOLOM1] [KONDISI1] (SELECT [KOLOM1] FROM [TABEL2] WHERE [KOLOM2] = [kata_kunci_pencarian])	SELECT * FROM **transaksi** WHERE **kode_mitra IN** (SELECT **kode_mitra** FROM **mitra** WHERE **kota** = '**bandung**')
Kolom1	kode_mitra		
Tabel2	mitra		
Kolom2	kota		
Kondisi1	IN		
kata_kunci_pencarian	bandung		

2.10 Test Result

Accuracy testing conducted in this study aims to determine the success of the system of translating Indonesian sentences into DML with sub-queries. The DML function that is tested is the function select with the condition "=", IN and NOT IN. Input sentence data in the form of raw Indonesian sentences containing DML commands with sub-queries. Testing is done by comparing the translation results of the system with a predetermined query query.

Table 1. 15 Test Result

No.	Jenis Perintah	Jumlah kalimat	Hasil Pengujian	
			Benar	Salah
1	SELECT '=' MAX	4	3	1
2	SELECT '=' MIN	4	4	0
3	SELECT '=' AVG	3	2	1
5	SELECT '>' MIN	3	3	0
6	SELECT '>' AVG	3	3	0
7	SELECT '<' AVG	5	4	1
8	SELECT IN	6	4	2
9	SELECT NOT IN	6	5	1
Total		34	28	6

The calculation method refers to the credit measure written by Abidin [11] for the overall accuracy of the input data in the Indonesian sentence translation system into the SQL language that contains sub-queries.

Based on the testing that has been done using black box testing and accuracy testing is concluded as follows.

1. System functionality is running properly.
2. The accuracy obtained is quite good, but the translation results are still not there due to some things, namely:
   a. The alias column whose contents are the same in more than one table cannot be detected by the alias intended by the user for which table.
   b. Search by using the date has not been formatted so that in testing the query on the database is not accurate

3. CLOSING

3.1 Conclusions

Based on the results of the implementation and testing of the research that has been carried out it can be concluded that the system can translate Indonesian DML command sentences containing sub-queries and generate DML queries with sub-queries. This study can translate Indonesian sentences into DML commands with sub-queries with an accuracy of 82,35%.

3.2 Suggest

Based on the results of the analysis and testing that has been carried out, suggestions are obtained so that this research can produce better translations, including the following.

1. Add rules that can be nested sub-queries or nested sub-queries and can handle more than two tables.
2. Add checks and handling for spelling the wrong words.
3. Add rules for detecting full date writing, months, and months.

BIBLIOGRAPHY

[1] B. Nugroho, *Database Relasional Dengan Mysql*. Yogyakarta: Andi, 2005.

[2] Kuspriyanto, H. Sujani, H. Tjahjono, and S. Kusuma, "Perancangan Translator Bahasa Alami Ke Dalam Format SQL (Structured Query Language)," vol. 10, no. 2005, p. 12, 2005.

[3] N. Isharmawan, “Ektraksi Informasi Dan SQL Parser Untuk Query Sql Dari Teks Berbahasa Indonesia,” Universitas Kompuuter Indonesia, 2017.

[4] I. Faturohman, “Perancangan Translator Bahasa Alami Ke Dalam Format SQL (Structured Query Language) Dengan Fokus Pada Kasus Retrival Informasi,” Universitas Kompuuter Indonesia, 2018.

[5] W. Budiharto and D. Suhartono, *Artificial Intelegece Konsep dan Penerapannya*. 2014.

[6] A. Fadilsyah, *Pemrograman Database (Konsep dan Implementasi)*. Yogyakarta: Graha Ilmu, 2008.

[7] K. K. Purnamasari and I. S. Suwardi, “Rule-based Part of Speech Tagger for Indonesian Language,” *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 407, Sep. 2018.

[8] I. H. Witten, “Text mining,” 2002.

[9] M. Adriani, J. Asian, B. Nazief, and E. Williams, “Stemming Indonesian : A Confi x-Stripping Approach,” *ACM Trans. Asian Lang. Inf. Process*, vol. 6, no. 4, pp. 1–33, 2007.

[10] B. Liu, *Web Data Mining Second Edition*, 2nd ed. Springer-Verlag Berlin Heidelberg, 2011.

[11] Abidin, “Accuracy Measure,” *Bahan kuliah data mining, Progr. Stud. Tek. Inform. FMIPA Univ. Syiah Kuala*, 2012.