INFORMATION EXTRACTION AND SQL PARSER FOR SQL QUERY FROM INDONEISAN LANGUAGE TEXT

Nendi Isharmawan[1], Galih Hermawan, S.Kom., M.T.[2]

[1]Universitas Komputer Indonesia
Jl. Dipatiukur No. 112-116, Bandung 40132
E-mail : nisharmawan@gmail.com[1], galih.hermawan@gmail.com[2]

ABSTRAK

Information is an important part of everyday life, where the information is expected to be useful and can be retrieved easily and faster. One of the information storage media namely databases that exist on the computer, one of which is a Database Management System. Structure and processing data in the DBMS using Structured Query Language (SQL) that divided two DDL and DML. Access to the database and the table data as well as raw formats or have called syntax. But not all the data input obtained to search for information in the form of an SQL syntax. Often the data input in the form of natural human language. Natural human language course has different formats with SQL syntax.

Natural Language Processing is one of the fields of artificial intelligence who study communication between masusia with computer via natural language. Application of Natural Language Processing (NLP) built to translate natural language human language into sql that can be directly executed by the system in order to get information from the source. Based on Blackbox testing, these applications can make a media information seekers in seeking information with the percentage of success sebersar 90.47% of these applications can manage several types of queries, i.e. queries with no conditions, query, query condition with two conditions, query a lot of table without conditions, querying many tables with aggregation, query conditions.

Keywords: information, queries, natural language, NLP.

1. INTRODUCTION

Information is an important part of everyday life, where the information is expected to be useful and can be retrieved easily and faster. One of the information storage media namely databases that exist on the computer DBMS. The DBMS is the engine or the software used to manage the database. Structure and processing data in the DBMS using Structured Query Language (SQL) that divided two DDL (Data Definition Language) and DML (Data Manipulation Language).

DDL is used to manipulate the structure of a database that includes the creation of a database, creating tables, and changing the structure of the table. While DML is used to process the data ditabel include the addition of data, alteration of data, the deletion of data, and the election or preview data. Access to the database and the table data as well as raw formats or have called syntax. But not all the data input obtained to search for information in the form of an SQL syntax. Often the data input in the form of natural human language. Natural human language course has different formats with SQL syntax.

This research focus on natural language processing (natural language processing) as a bridge to access the data in the MySQL DBMS. In order for this research to be more directional some restriction problem set as follows:

1. Structured Query Language are examined on the MySQL DBMS that is specifically the query SELECT
2. The SELECT Query that was performed with 6 types, that is, the query without the query with a condition, condition, condition of two queries, query a lot of table without conditions, querying many tables with aggregation, query conditions.

To achieve the goals of this research are:

1. Natural language Translates into the SQL syntax.
2. Know the accuracy in the presentation of the application of Natural Language Processing in natural language interpreter into the syntax of SQL.

2. CONTENT

At this stage will discuss issues that will be analyzed. Some of the things contained in this stage regarding the review of the literature, discussions and outcomes of the research.

2.1 Review Of The Literature

Some research on tools for media information seekers are making natural language translators, because of differences between human and natural language language used to access databases. Like Sri Hartati research and Eri Zuliarso, conducting research for natural language processing is applied to the XML database, where the research was done at the database information system processing the data book [4]. Setyawan Wibisono, doing similar things by doing

research on natural language for academic database with XML data format by adding the sentence patterns can be a reference to the application users [2]. Sariyun Naja Anwar conducted research in making the query database exam performed at the University of STIKUBANK, like previous research, this research also uses XML in database storage media information [5]. Muslim Aries natural language translators do their research with a virtual machine or which are often known as chatbot where its function is almost the same and that is to get the information provided previously to the virtual machine [1]. Herry Sujani do penetilitan about natural language into a pattern of sql syntax that has only two patterns i.e. pattern syntax with attributes selected calls and no attribute selected both without conditions [3].

Of existing research, the research that will be conducted this focus to the manufacture of a machine translator which can be used on any database on the MySQL DBMS.

2.2 Natural Language Processing

To perform the translation of natural language into a query there needs to be several stages to do can be seen in Figure 2.1.

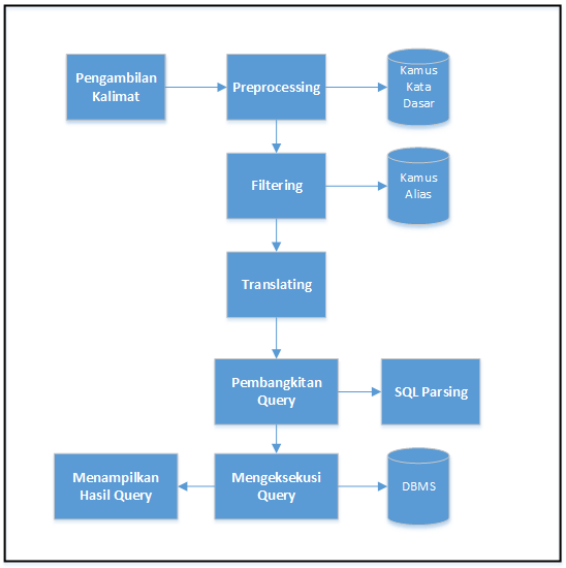


Figure 2. 1 Natural language Processing stages

1. Input Sentences
   Get natural language sentences entered by the user are:
   "Perlihatkan nama mahasiswa"

2. *Scanning*
   At this stage do the scanning stage i.e. natural language sentence scan being entered and then do a contraction of the sentence being said perkata. After being beheaded:
   "Perlihatkan","nama","mahasiswa"

3. *Cleansing*
   The result of the above process obtained a list of tokens in the shape of an array, it will do the process menghilangakan punctuation are not used as ".!,?". The process can be seen in Figure 2.2.

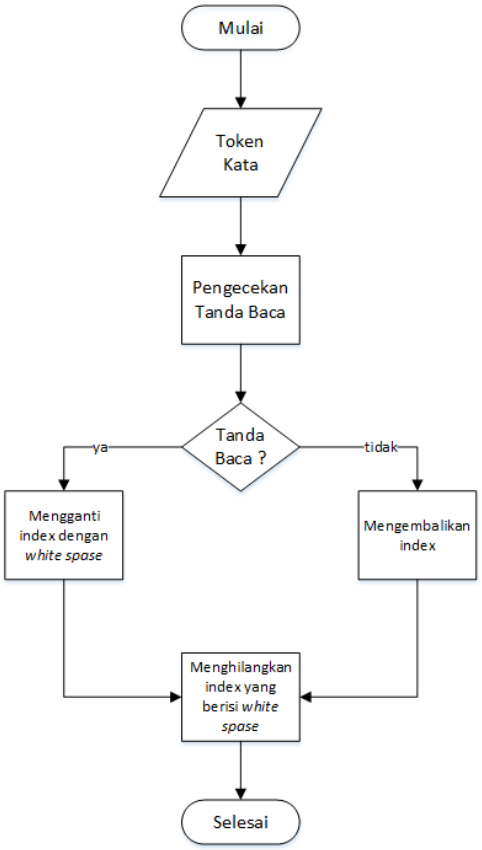


Figure 2. 2 *Cleansing*

4. *Case Folding*
   The result of the previous process will do the process of equating case into lower case all for each in tokennya. Then the result will be. "show me", "name", "student". The process can be seen in Figure 2.3.

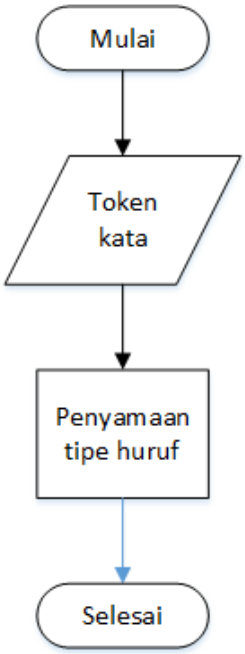


Figure 2. 3 *Casefolding*

5. *Remove Stopword*
The results of the adjustments made to the previous process, will do the process of removing unnecessary words as conjunctions and the substitute person. The process of her can be seen in Figure 2.4.

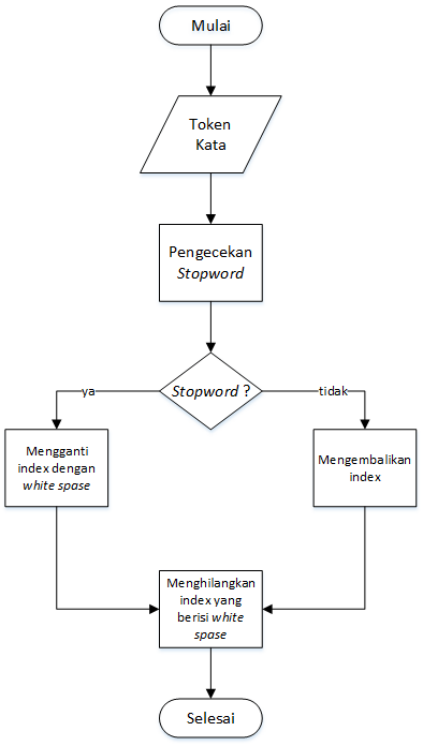


Figure 2. 4 Tahapan *Remove Stopword*

A list of words including the stopword is can be seen in Figure 2.5.

No	Kata *stopword*	No	Kata *stopword*	No	Kata *stopword*
1	Mampu	11	Karena	21	Mungkin
2	Tentang	12	Sudah	22	sebagian
3	Di	13	Oleh	23	Kata
4	Setelah	14	Bisa	24	Sering
5	Semua	15	Dia	25	Saya
6	Hampir	16	Telah	26	Belum
7	Juga	17	Jika	27	Atas
8	Ada	18	Ke	28	Sesuai
9	Seperti	19	Itu	29	Tabel
10	Jadi	20	Dalam	30	pada

Figure 2. 5 *Stopword*

6. *Stemming*
The results of the previous proceedings were taken and would do a basic search of the word stemming process of each token. Then it would be, "see", "name", "student". The process can be seen in Figure 2.6.

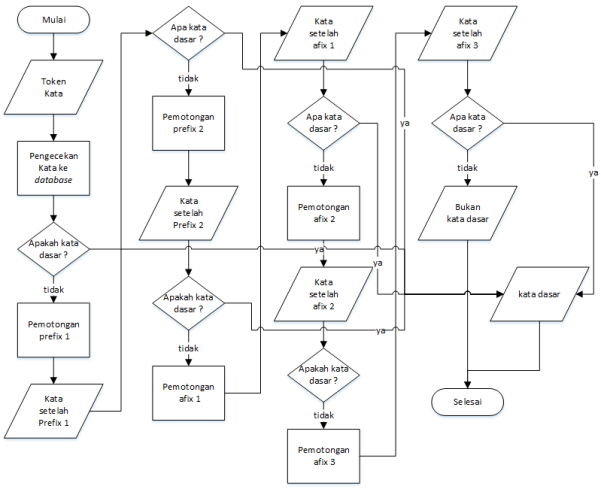


Figure 2. 6 *Stemming*

2.3 *Filtering*

Filtering is a process checking each token and then compared to the dictionary words, i.e. words dictionary commands, parameters, the parameters field of the table, the parameters of the condition, value parameters. The following is a table of Word commands can be seen in Table 2.1:

Table 2. 1 Table Word Order

No	Kata Perintah
1	tampil
2	lihat
3	sebut
4	cari
5	apa

The process of filtering can be seen on the flowchart in Figure 2.7.

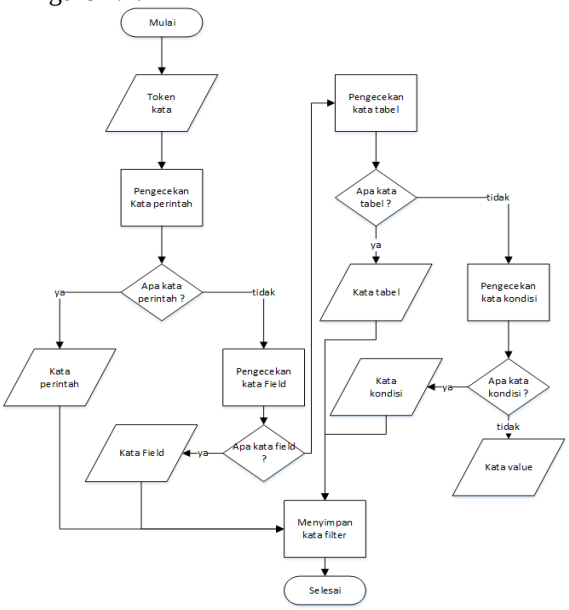


Figure 2. 7 *Flowchart Filtering*

It will be obtained:

a. the Word command: see
b. the Word field : name
c. the Word tables: students

2.4 *Translating*

After the analysis of sentence entered and generates a token already categorized then the token will be transformed into a query understandable words.

The process consists of Translating the process of comparing the word on every token that has gone through the process of checking with the dictionary of words that are already provided, his process can be seen on the flowchart in Figure 2.8.

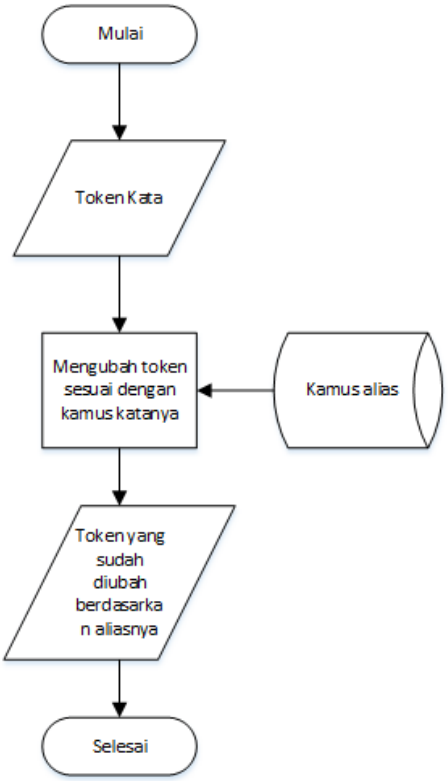


Figure 2. 8 *Flowchart Translating*

It will be didapatan the results of the process of translating is as follows:

a. the Word command: select
b. the Word field : nama
c. the Word tables: mahasiswa

Next will be conducted according to the rules for preparing the query syntax query. Then the result will be:

"select nama from mahasiswa"

2.5 Execute Query

The query has been obtained will be sent to the DBMS to be able to diekseskusi and get the information from the query. Results can be seen in Figure 2.9.

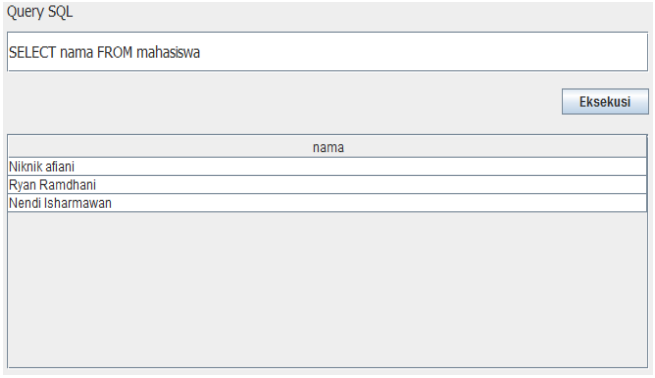


Figure 2. 9 Result of Execute Query

2.6 Testing

Testing done is testing different types of queries are:

a. Query without conditions

This testing is done with the menginputkan sentences in the input form and then press the "process". The sentence cited can be seen in table 2.2.

Table 2. 2 Table First Type Of Sentence

No	Kalimat
1	Tampilkan mahasiswa
2	Tampilkan nama mahasiswa
3	Lihat nilai uas dan uts
4	sebutkan nama, alamat dan kota mahasiswa

The query is expected of sentences in the table 2.2 can be seen in Table 2.3.

Table 2. 3 Table of expected query

No	Kalimat
1	select * from mahasiswa
2	select nama from mahasiswa
3	select uas , uts from nilai
4	select nama, alamat, kota from mahasiswa

The result of the phrase that is entered after pressing the button the process can be seen in table 2.4.

Tabel 2. 4 Test results of a sentence

No Kalimat	Hasil pengujian
1	SELECT * FROM mahasiswa
2	SELECT nama FROM mahasiswa
3	SELECT uas , uts FROM nilai
4	SELECT nama , alamat , kota FROM mahasiswa

b. Query with conditions

This testing is done with the menginputkan sentences in the input form and then press the "process". The sentence cited can be seen in table 2.5.

Tabel 2. 5 Table Second Type Of Sentence

No	Kalimat
1	Lihat mahasiswa yang nimnya 10112322
2	tampilkan nim mahasiswa yang namanya udin
3	tampilkan nama mahasiswa yang jenis kelamin nya pria
4	tampilkan mahasiswa berdasarkan jenis kelamin

The query is expected of sentences in the table 2.5 can be seen in Table 2.6.

Table 2. 5 Table of expected query

No	Kalimat
1	select * from mahasiswa where nim = '10112322'
2	select nim from mahasiswa where nama = 'udin'
3	select nama from mahasiswa where jk = 'pria'
4	select nama from mahasiswa order by jk

The result of the phrase that is entered after pressing the button the process can be seen in table 2.7.

Tabel 2. 6 Test results of a sentence

No	Kalimat
1	SELECT * FROM mahasiswa WHERE nim = '10112322'
2	SELECT nim FROM mahasiswa WHERE nama = 'udin'
3	SELECT nama FROM mahasiswa WHERE jk = 'pria'
4	SELECT * FROM mahasiswa group by jk

c. Query dengan dua kondisi

This testing is done with the menginputkan sentences in the input form and then press the "process". The sentence cited can be seen in table 2.8.

Table 2. 7 Table Third Type Of Sentence

No	Kalimat
1	lihat mahasiswa yang tinggal di bandung dan namanya nendi
2	tampilkan mahasiswa yang berjenis kelamin perempuan dan tinggal di bandung

The query is expected of sentences in the table 2.8 can be seen in Table 2.9.

Table 2. 8 Table of expected query

No	Kalimat
1	select * from mahasiswa where kota = 'bandung' and nama = 'nendi'
2	select * from mahasiswa where jk = 'perempuan' and kota = 'bandung'

The result of the phrase that is entered after pressing the button the process can be seen in table 2.10.

Table 2. 9 Test results of a sentence

No	Kalimat
1	SELECT * FROM mahasiswa WHERE kota = 'bandung' and nama = 'nendi'
2	SELECT * FROM mahasiswa WHERE jk = 'wanita' and kota = 'bandung'

d. Query banyak tabel tanpa kondisi
This testing is done with the menginputkan sentences in the input form and then press the "process" of many of his sentences used was four and the table that is called must have a relationship first. The sentence cited can be seen in table 2.11.

Table 2. 10 Table Fourth Type Of Sentence

No	Kalimat
1	tampilkan nilai dan kelas
2	lihat kelas mahasiswa
3	sebutkan nama mahasiswa dan nilai uas
4	sebutkan nama kelas dan nilai uas

The query is expected of sentences in the table 2.11 can be seen in Table 2.12.

Table 2. 11 Table of expected query

No	Kalimat
1	select * from nilai join kelas on nilai.kode_kelas = kelas.kode_kelas
2	select * from mahasiswa join kelas on mahasiswa.kode_kelas = kelas.kode_kelas
3	select nama , uas from mahasiswa join nilai on mahasiswa.nim = nilai.nim
4	select nama_kelas, uas from nilai join kelas on nilai.kode_kelas = kelas.kode_kelas

The result of the phrase that is entered after pressing the button the process can be seen in table 2.13.

Table 2. 12 Test results of a sentence

No	Kalimat
1	SELECT * FROM nilai join kelas on nilai.kode_kelas = kelas.kode_kelas
2	SELECT * FROM
3	SELECT nama , uas FROM mahasiswa join nilai on mahasiswa.nim = nilai.nim
4	SELECT nama_kelas , uas FROM kelas join nilai on kelas.kode_kelas = nilai.kode_kelas

e. Query banyak tabel dengan kodisi
This testing is done with the menginputkan sentences in the input form and then press the "process". The sentence cited can be seen in table 2.14.

Table 2. 13 Table Fifth Type Of Sentence

No	Kalimat
1	sebutkan mahasiswa yang nilai uasnya 80 sampai 90
2	siapa mahasiswa yang uas nya bernilai 90

The query is expected of sentences in the table 2.14 can be seen in Table 2.15.

Table 2. 14 Table of expected query

No	Kalimat
1	select * from mahasiswa join nilai on mahasiswa.nim = nilai.nim where uas between 80 and 90
2	select * nama from mahasiswa join nilai on mahasiswa.nim = nilai.nim where uas = '90'

The result of the phrase that is entered after pressing the button the process can be seen in table 2.16.

Table 2. 15 Test results of a sentence

No	Kalimat
1	SELECT * FROM mahasiswa join nilai on mahasiswa.nim = nilai.nim WHERE uas between 80 and 90
2	SELECT * FROM mahasiswa WHERE uas = '90'

f. Query agregasi

This testing is done with the menginputkan sentences in the input form and then press the "process". The sentence cited can be seen in table 2.17.

Table 2. 16 Table Sixth Type Of Sentence

No	Kalimat
1	berapa jumlah mahasiswa
2	berapa nilai uas paling tinggi
3	berapa nilai tugas paling rendah
4	berapa jumlah mahasiswa yang nilai uas lebih dari 70

The query is expected of sentences in the table 2.17 can be seen in Table 2.18.

Table 2. 17 Table of expected query

No	Kalimat
1	select count(*) from mahasiswa
2	select max(uas) from nilai
3	select min(tugas) from nilai
4	select count(*) from mahasiswa join nilai on mahasiswa.nim = nilai.nim where uas > '70'

The result of the phrase that is entered after pressing the button the process can be seen in table 2.19.

Table 2. 18 Test results of a sentence

No	Kalimat
1	SELECT count(*) FROM mahasiswa
2	SELECT max(uas) FROM nilai
3	SELECT min(tugas) FROM nilai
4	SELECT count(*) FROM mahasiswa join nilai on mahasiswa.nim = nilai.nim WHERE uas = '70'

2.7 Pembahasan

Applications are made to be able to translate from natural language into a sql query. This can be demonstrated by testing that has been done. The process of translating the following process flow used in natural language processing. Starting with the preprocessing, filtering, translating until finally in compiling a query. The appearance made it easy to be able to see every process so that users can understand any process that is passed by the application.

Of the six existing query types produce a percentage success of 90.47%, because there are some sentences that are to be tested does not comply with expectations or does not match the information it generates.

Words that are not found in the data dictionary then it will not be considered by this application, so the need for large data dictionary so that it can cope with a variety of different databases and a variety of different fields. There are words in the Dictionary data has an adjustment against a dictionary of data obtained from the intended database, so even though the Dictionary data is large but not all of the data dictionary will be used. There needs to be adjustments Dictionary data so the application won't take the Dictionary data will not be used.

3. PENUTUP

3.1 Conclusion

After doing the analysis, design, and testing by the method of blackbox, obtained the information that the system can help the media information seekers because sentences into input could turn into the query. Test results from the few sentences obtained success earning 90.47% showed that when a sentence is entered by users less complete it will be obtained also the wrong query or produce incorrect information. This can lead to errors in executing the query. Thus the media information seekers can directly access the source of the information on the database when data in the database is changed, then the results from data shown will also be changed.

3.2 Advice

Based on the results of the research that has been done, still need to do some further study. As for suggestions for further research are:

1. Use the pengkoreksi sentence when met with a word that is not complete or incorrect writing.
2. Use the data input in the form of voice so that users don't have to do typing in place his sentence.
3. Use the existing database on the network the internet in order for the data which is taken to be more recent..

REFERENCE

[1] Aries Muslim, Kurniawan Robby. Natural Language Processing (NLP) Komunikasi Bahasa Natural Dengan Actor Virtual. Skripsi S1. Teknik Informatika. Depok: Universitas Gunadarma (2008).

[2] Wibisono Setyawan. Aplikasi Pengolahan Bahasa Alami untuk Query Basisdata Akademik dengan Format Data XML. Skripsi S1. Teknik Informatika. Universitas Stikubank (2013).

[3] Kuspriyanto, Sujani Herry, Tjahjono Hary, Kusuma Shofyan. Pernacangan Translator Bahasa Alami Ke Dalam Format SQL. Skripsi S1. Teknik Elektro : Institut Teknologi Bandung (2005).

[4] Hartati Sri, Zuliarso Eri. Aplikasi pengolah Bahasa Alami untuk Query Basisdata XML. Yogyakarta : Universitas Gadjah Mada (2008).

[5] Naja Anwar Sariyun, Suhari Yohanes, Wibisono Setyawan. Query Bahasa Indonesia pada Basisdata Soal Ujian di Fakultas Teknologi Sistem Informasi Universitas STIKUBANK dengan Format XML. Skripsi S1. Teknik Informatika : Universitas STIKUBANK (2013).