NL2SQL for Chatbot with Semantic Parsing Using Rule-Based Methods

Adi Kurniawan[a,1,*], Abdiansah Abdiansah[b,2], Alvi Syahrini Utami[b,3]

[ab]Informatics, Faculty of Computer Science, Sriwijaya University, Palembang, Indonesia
[1] hello.adikurniawan@gmail.com*; [2] abdiansah@unsri.ac.id; [3] alvisyahrini@ilkom.unsri.ac.id
* corresponding author

ARTICLE INFO

Article history
Received 12 Feb 2024
Revised 24 Feb 2024
Accepted 24 Feb 2024 `

Keywords
Chatbot
Natural Language to Structured Query Language
NL2SQL
Rule Base

ABSTRACT

Structured Query Language (SQL) is a command language that allows users to access database information. Ordinary people generally do not know how to make queries with SQL to a database. The chatbot is a computer program developed to interact with its users via text or voice. In this study, chatbots were developed to help and facilitate users in the Natural Language to Structured Query Language (NL2SQL) process to search for information in an Academic Information System database with semantic parsing using a rule-based method that accepts input in the form of interrogative sentences or order. In the Natural Language toStructured Query Language (NL2SQL) process several problems arise, namely input problems with unique parameters for the knowledge base, and slow searching or translation processes, which make Natural Language to Structured Query Language (NL2SQL) inefficient, problems This problem will be solved using a semantic parsing approach using a rule-based method that is proven to be efficient in solving issues such as the Natural Language to Structured Query Language (NL2SQL) process. The results showed that the semanticparsing approach using the rule-based method succeeded in obtaining an accuracy rate of 96.72% using 122 test data in the form of question sentences or command data about the Academic Information System of the Department of Informatics Engineering, Sriwijaya University in Indonesian, and an average execution time of 50.68 milliseconds. seconds or 0.05 seconds.

1. Introduction

The chatbot is a conversational interface software program that allows users to communicate in the same way they would interact with humans [1]. One of the technologies implemented together with chatbots is Natural Language to Structured Query Language (NL2SQL).

Natural Language to Structured Query Language (NL2SQL) or previously known as Text to SQL can be described as a system that accepts input in the form of a Natural Language Query (NLQ) against a Relational Database (RDB), and produces a Structured Query Language (SQL) which is equivalent in meaning to NLQ, and is valid for the RDB [2]. NL2SQL is a natural language interface to the database that is able to assist users in accessing the database without the need for knowledge of Structured Query Language (SQL) and technical databases. Query results from NL2SQL need to be formed with precision, because the results of forming a precise query greatly affect the information or data accessed from the database.

Several methods have been used to solve NL2SQL problems, including rule-based [3], Cosine Similarity [4], and Top-Down Parsing [5]. Cosine Similarity has the advantage of accuracy in calculating the level of similarity between two objects expressed in two vectors using keywords but has a weakness when the natural language input contains parameters that are unique to the knowledge base that has been built. Top-Down Parsing has the advantage of parsing a sentence fromthe largest constituent to the smallest constituent, but Top-Down Parsing has the disadvantage of recursive searching and backtracking being slow when handling complex natural language input.

The rule-based method can solve problems in the Cosine Similarity and Top-Down Parsing algorithms, namely the natural language input contains parameters that are unique to the knowledge base that has been built, and the searching process is carried out recursively and slow backtracking. But the rule-based method has a weakness in the difficulty of determining the meaning of words contained in a natural language. Semantic Parsing is an approach that can solve problems in determining the meaning of words in a form of meaning representation (logical form) that can be understood by machines to get the exact meaning of the natural language.

This study aims to develop an NL2SQL system with Semantic Parsing using rule-based methods in forming SQL queries in NL2SQL so that more optimal and accurate results are expected.

2. Literature Study

a. Chatbot

Chatbots are supported by a set of rules and sometimes artificial intelligence that is built to help humans in terms of providing information on specified topics[6]. Chatbot consists of two words namely "chat" and "bot". Chat which means conversation and bot is taken from the word "robot". In other words, Chatbot is a robot or virtual machine that can simulate conversations with humans through a conversational interface via text or speech.

b. Natural Language to Structured Query Language (NL2SQL)

Natural Language to Structured Query Language (NL2SQL) is a system that converts statements in natural language into queries in the form of Structured Query Language (SQL) to access information stored in database management systems [7]. NL2SQL or previously known as Text to SQL can be described as a system that accepts input in the form of a Natural Language Query (NLQ) against a Relational Database (RDB) and produces a Structured Query Language (SQL) which is equivalent in meaning to NLQ and is valid for the RDB [2].

c. Semantic Parsing

Semantic Parsing is the process of translating natural language into a form of representation of its meaning (logical form). Mapping from Natural Language to SQL (NL2SQL) is one of the important tasks of semantic parsing in a question-answering system [8]. The result of the semantic parsing process is the result of natural language translation in a logical form that can be understood by machines so that the exact meaning of the natural language can be obtained.

d. Natural Language Processing (NLP)

Natural Language Processing (NLP) is a programming technique in which computers can understand and provide output in the form of human language or simply facilitate communication between humans and machines [9]. The goal of NLP is to provide appropriate answers or responses based on machine understanding of the meaning of human language [10].

e. Text Processing

Preprocessing is one of the stages of eliminating problems that can interfere with the results of data processing. In many NLP studies, the preprocessing process is very important to produce better accuracy [11]. In this research, there are several kinds of processes carried out such as case folding, punctuation removal, stemming, stop word removal, word conversion, and tokenizing.

f. Rule Based System

A rule based system is a computer program that processes information contained in working memory with a set of rules contained in the knowledge base to produce new information [12]. A rule-based system has several essential elements, namely:

1. A collection of facts can be in the form of statements, data, or conditions.
2. Set of rules, these rules define all the steps that must be taken when given a set of facts.
3. Standard termination, which is a condition that determines whether a solution has been found ornot to avoid an infinite loop.

g. Database

The database is a collection of data that is organized in such a way that data is easily stored and manipulated, such as updating, searching, processing with certain calculations, and deleting [13]. The database consists of 2 words, namely data, and base. The base can be interpreted as a gathering place, while data is a representation of real-world facts that represent an object recorded in the form of numbers, letters, symbols, text, images, sounds, or a combination thereof.

h. Structured Query Language (SQL)

Structured Query Language (SQL) is a command language that allows database users to create, manipulate data, and extract information [14]. SQL commands are also known as queries. There are three subcommands in the SQL command, namely Data Definition Language (DDL), Data Manipulation Language (DML), and Data Control Language (DCL).

i. Data Manipulation Language (DML)

Data Manipulation Language (DML) is a SQL command method that is used to process data in tables such as displaying, entering, changing, deleting data contents, and is not related to changes in the structure and definition of data types from database objects [15]. In this study, the DML command used on this system is only the SELECT command to retrieve or display data from a tableview in the database system.

j. Agile Software Development Methodology

Agile Software Development Methodology is a modern software development methodology based on the principles of short-term system development that requires developers to adapt quickly to changes of any kind [16]. Agile Software Development Methodology allows developers to develop software that has to change requirements quickly [17]. The Agile Software Development Methodology method consists of several stages, namely requirements, design, development, test, deployment, review, and launch [18].

k. Relevant References

Research on NL2SQL has been carried out in previous years. Fu'adi has conducted research on "Implementation of Commands to Display Data Using Indonesian with Natural Language Processing". In this study, the method used is a rule-based method with a Top-Down Parsing algorithm, the input that can be used to access the database is in the form of Indonesian sentences with production rules made in such a way that commands in Indonesian can conform to the format of SQL commands.

Stratica, et al [19] researched "NLIDB Templates for Semantic Parsing". This research developeda template for Natural Language Interface to Database (NLIDB) with a Semantic Parsing approach,the system can translate commands in English into SQL form, but the system has not been able to form SELECT SQL queries in the form of relations with JOIN between tables.

Another study was conducted by Jemi, et al [20]. In the Research "Conversion of Indonesian to SQL (Structured Query Language) with a Statistical Translation Machine Approach", the system is implemented using a statistical machine translation approach to convert Indonesian into SQL with a BLEU accuracy value of 64.89% and a database application of 70%.

Fauzan and Alif[3] examined a system capable of translating compound sentences containing the word referring to them using a rule-based method. This study entitled "Coreference Implementation on the Indonesian Language to Structured Query Language (SQL) Conversion on Changing The Structure of The Table" applies a rule-based method without looking at the table structure and produces SQL language with the Data Definition Language (DDL) command. The final value obtained from the test is 88.54%.

3. Methodology

a. Types of Data, Data Sources, and Data Collection Methods

The type of data used in this study is primary data in the form of a set of command sentences or questions in Indonesian with the scope of questions regarding student, lecturer, and course data in the Sistem Informasi Akademik (SIMAK) Department of Informatics Engineering, Sriwijaya University.

Data on a set of questions or command sentences were obtained directly by distributing questionnaires to informatics engineering student respondents at Sriwijaya University. The questionnaire method is a data collection technique that is carried out by giving a set of written questions to the respondent to answer, which can be given in person or via post or the internet.

The data obtained were 148 data, but not all data were used in this study because there were data with sentences that did not fit the scope of the interrogative sentences or commands used in this study, namely regarding student, lecturer, and course data. So, there are 122 data that are used, and 26 data that are not used, and then the data obtained will be determined by the Expected Structured Query Language (SQL) based on the production rules for constructing sentences that have been adapted to the syntax format in Structured Query Language (SQL). Table 1 shows an example of the data collected.

Table 1. Example of Data Used

No.	Questions or Commands	Structured Query Language Hope

1.	Apa nama dan sks mata kuliah dengan kode FIK20022?	SELECT mata_kuliah.nama, mata_kuliah.sks
		FROM mata_kuliah
		WHERE mata_kuliah.kode = “FIK20022”

2.	berapa nilai suliet mahasiswa dengan nim 09021181823003	SELECT mahasiswa.suliet
		FROM mahasiswa
		WHERE mahasiswa.nim = “09021181823168”

3.	Berikan data dosen pembimbing dari mahasiswa yang bernama Nadia	SELECT mahasiswa.
		dosen_pembimbing_akademik
		FROM mahasiswa
		WHERE mahasiswa.nama = "nadia"

4.	Apa nama dan sks mata kuliah dengan kode FIK20022?	SELECT mata_kuliah.nama, mata_kuliah.sks
		FROM mahasiswa
		WHERE mata_kuliah.kode = “FIK20022”

5.	Temukan nama dosen yang memiliki NIP 90923xxxx	SELECT dosen.nama
		FROM dosen
		WHERE dosen.nip = “90923xxxx”

6.	Temukan nama mahasiswa yang memiliki NIM 0902118182xxxx	SELECT mahasiswa.nama
		FROM mahasiswa
		WHERE mahasiswa.nim = “0902118182xxxx”

7.	Tolong carikan nama mahasiswa yang angkatan 2018 urutkan berdasarkan NIM secara menurun	SELECT mahasiswa.nama
		FROM mahasiswa
		WHERE mahasiswa.angkatan = “2018”
		ORDER BY mahasiswa.nim DESC

8.	Tampilkan nama, dan NIP dosen yang memiliki nama berawalan huruf a urutkan berdasarkan NIP secara meningkat	SELECT dosen.nama, dosen.nip
		FROM dosen
		WHERE dosen.nama LIKE “a%”
		ORDER BY NIP ASC

9.	Tampilkan nama mahasiswa dengan IPK lebih besar dari 3.5 batasi 5 data	SELECT mahasiswa.nama
		FROM mahasiswa WHERE mahasiswa.ipk > “3.5”
		LIMIT 5

10.	Tolong temukan data mahasiswa dengan angkatan dibawah 2018 batasi 10 data	SELECT * FROM mahasiswa
		WHERE angkatan < “2018”
		LIMIT 10

After determining the Structured Query Language (SQL) from the data used in this study, the distribution of query types based on optional clauses from the data used is obtained which is presented through the Venn Edwards diagram in Fig. 1.

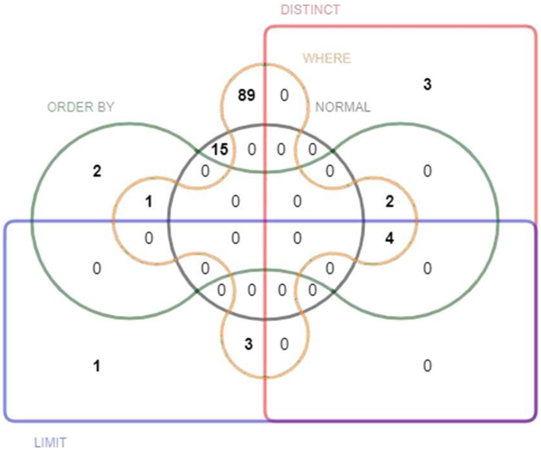


Fig. 1. Distribution of Data Query Types used

b. Software Architecture

Architecture outlines what components will be developed to form the system as a whole. The process of this system starts from user input in the form of natural language, namely Indonesian language questions or commands, then received by message recipients on the chatbot server, and processed at the stages in the NL2SQL REST API in the form of Pre-Processing, Parsing, Translating, and query execution , the expected output is SQL query translation result, and execution result data from SIMAK SIMULATION database when input can be translated into Structured Query Language (SQL) language format, as well as error validation messages when input cannot be translated into Structured Query Language format (SQL), the output will be sent by the message sender on the chatbot server to the user as a reply message. The software architecture in this study can be seen in Fig 2.

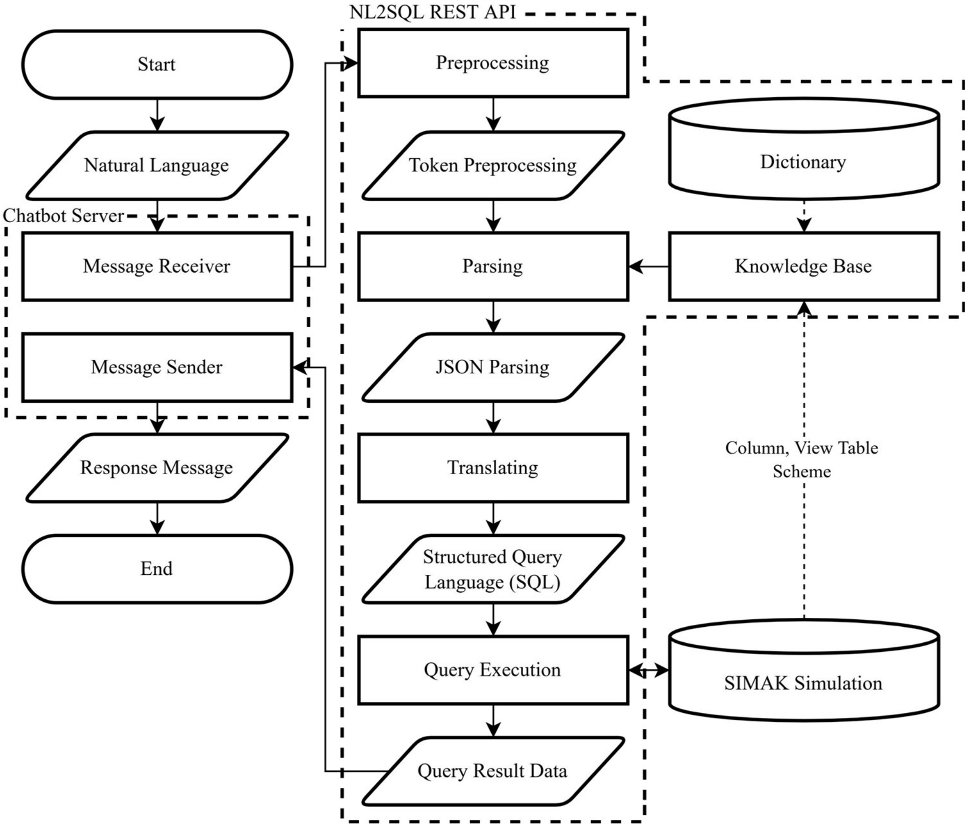


Fig. 2. Software Architecture

i. Preprocessing

This stage is the initial process of preparing input data in the form of natural language, namely Indonesian question sentences or commands before proceeding to the next stage. This research has several preprocessing stages, namely case folding, punctuation removal, stemming, stop word removal, word conversion, and tokenizing. The flowchart of preprocessing can be seen in Fig 3.

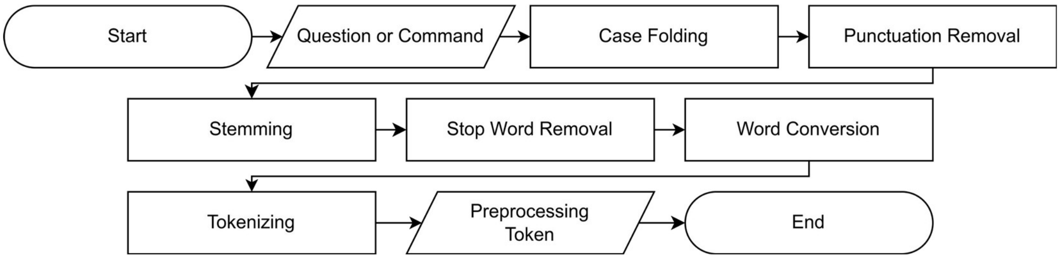


Fig. 3. The flowchart of preprocessing

ii. Parsing

At this stage, the input parsing process is carried out in the form of tokens resulting from preprocessing based on the production rules of the Structured Query Language (SQL) compiler and produces output in the form of an object that contains a list of words identified according to the production rules of the Structured Query Language (SQL) compiler. The steps at this stage include identifying keywords, identifying table views, identifying column view tables, and identifying conditions that are carried out asynchronously with the knowledge base of word dictionaries, schema view tables, and column view tables in the SIMAK SIMULATION database. The flowchart of Parsing can be seen in Fig 4.

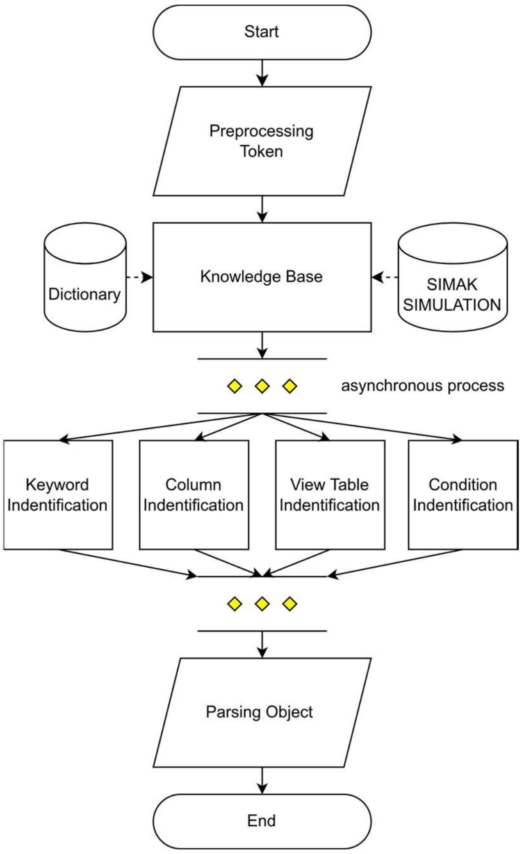


Fig. 4. The flowchart of parsing

iii. Translating

At this stage, the process of mapping the results of the parser according to the production rules is carried out into the resulting language, namely the Structured Query Language (SQL) language. The specified production rules can produce complex sentences because some identified words such as condition, keyword order, column, and condition column can have zero to more than one identified word. However, paying attention to questions that are commonly used to access data from the Academic System database of the Sriwijaya University Informatics Engineering Department, the translator is only designed to change the parser results whose arrangement matches the pattern of the questions. The flowchart of Translating can be seen in Fig 5.

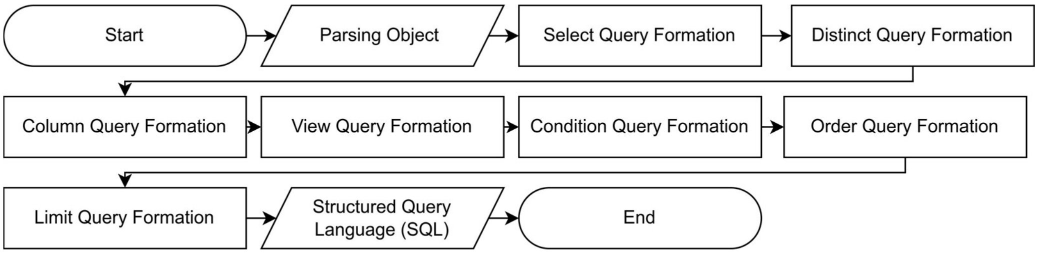


Fig. 5. The flowchart of translating

iv. Query Execution

The Structured Query Language (SQL) generated in the previous stage is executed to the SIMAK SIMULATION database and the results of the data view table output are seen, if the query results from the previous stages are invalid it will produce an error message with status code 500, if the query does not have query result data on the database will generate a message in the form of Structured Query Language (SQL) the results of the previous stage along with information about there is no data in the databasewith status code 404, and if the query is valid it will produce a message in the form of Structured Query Language (SQL) the results of the previous stage along with query result data in the database with status code 200. The flowchart at this stage can be seen inFig 6.

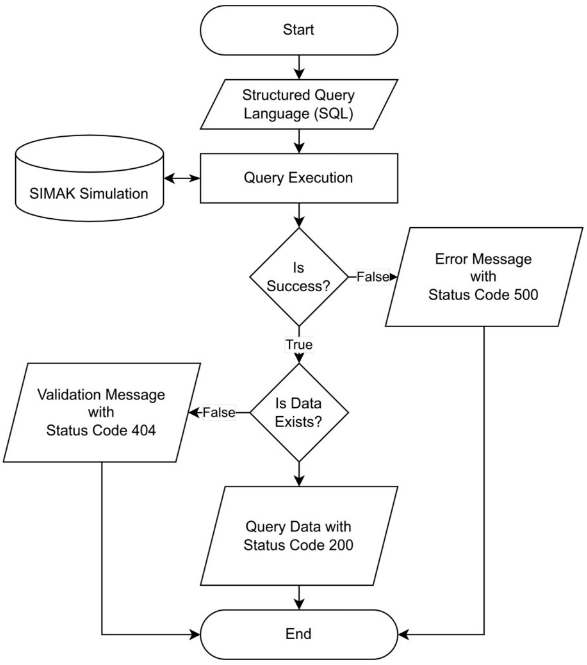

Start
Structured Query Language (SQL)
SIMAK Simulation
Query Execution
Is Success?
False
Error Message with Status Code 500
True
Validation Message with Status Code 404
False
Is Data Exists?
Query Data with Status Code 200
End

Fig. 6. The flowchart of query execution

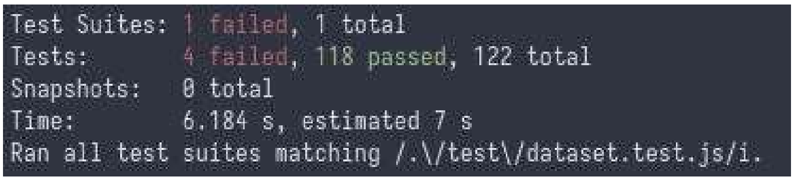
4. Result

a. Test Configuration

Tests were carried out on test data in the form of command sentences or questions in Indonesian with the scope of questions regarding Academic Information Systems (SIMAK) of the Department of Informatics, Sriwijaya University, totaling 122 data. The test was carried out by using an integration test script to automate the testing of all 122 test data in JavaScript Object Notation (JSON) format with the key "nl" for commands or questions, and "sql" for SQL expectations from NL2SQL. The integration test script checks the SQL NL2SQL results from the input sentence "nl" with the expected SQL from "sql", if the check is correct then the result of the integration test is worth PASS, and if the check is not correct then the result of the integration test is worth FAIL so that from the testing process the results of the number of data that are suitable and not suitable are based on the suitability of the expected SQL with the NL2SQL SQL results.

b. Test Result

The result of the test is the conformity of the answers given by the chatbot in the form of SQL NL2SQL results to the sentences entered by the user with the expected SQL. Fig 7 shows all the data tested along with the execution time and the final test results.



```
Test Suites: 1 failed, 1 total
Tests:       4 failed, 118 passed, 122 total
Snapshots:   0 total
Time:        6.184 s, estimated 7 s
Ran all test suites matching /.\/test\/dataset.test.js/i.
```

Fig. 7. The flowchart of query execution

Based on the results of the tests that have been carried out, then the accuracy value will be calculated. To find out the accuracy value of the chatbot, the test will be based on the number of SQL NL2SQL results that match the expected SQL in the script integration test. It can be seen that from 122 test data in the form of interrogative sentences or commands, 118 test data were suitable and 4 test data were not appropriate.

The accuracy of the NL2SQL software for Chatbot with Semantic Parsing Using Rule-Based Methods will be shown in table 2 of software accuracy.

Table 2. Software Accuracy

Final Test Results	Amount	Accuracy
Appropriate	118	96.72 %
Inappropriate	4	3.28 %
Total	122	100%

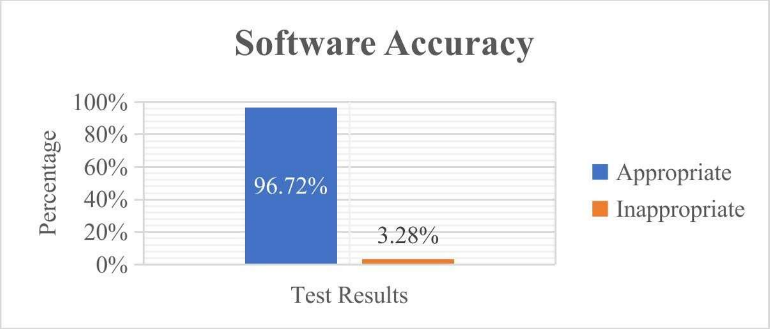


Fig. 8. Software Accuracy

It can be seen from Table 2 and Fig 8 that the accuracy percentage value of the NL2SQL software for Chatbot with Semantic Parsing Using the Rule-Based Method reaches 96.72% for appropriate answers, and 3.28% for inappropriate answers.

5. Conclusions and Recommendations

a. Conclusions

Based on the research that has been described from the description of the previous chapter, conclusions can be drawn from this research including the following:

1. NL2SQL software for Chatbot with Semantic Parsing Using a Rule-Based Method has been successfully developed.
2. From the results of the research that has been done, it is found that the accuracy level of the NL2SQL software for Chatbot with Semantic Parsing Using the Rule-Based Method is 96.72%,and the average execution time is 50.68 milliseconds or 0.05 seconds.

b. Recommendations

In future research, it is hoped to make the following suggestions:

1. Added optional clauses such as HAVING, GROUP BY, and DATE TIME in DML SELECT.
2. Development on a wider scope of DDL, DCL, and DML.
3. This research can be used as a reference for research on Natural Language to Structured Query Language with further rule-based methods.

References

[1] G. K. Vamsi, A. Rasool, and G. Hajela, "Chatbot A Deep Neural Network Based Human to Machine Conversation Model," *IEEE*, 2020.

[2] G. Katsogiannis-meimarakis and G. Koutrika, "Deep Learning Approaches for Text-to-SQL Systems," *open Proc.*, pp. 710–713, 2021.

[3] F. Abdulwahid and A. Finandhita, "Coreference Implementation On Indonesian Language to Structured Query Language (SQL) Convertion on Changing The Structure of The Table," *Tek. Inform. –Universitas Komput. Indones.*, 2019.

[4] R. Agustina, "Akses Basis Data Melalui Perintah Suara Berbahasa Indonesia Menggunakan Algoritma Cosine Similarity," *Univ. Sumatera Utara*, 2018.

[5] F. Fu'adi, "Implementasi Perintah Menampilkan Data Menggunakan Bahasa Indonesia Dengan Natural Language Processing," *Proceeding of KMICE'08*, no. 021, pp. 1–8, 2015.

[6] Eka Yuniar and Heri Purnomo, "Implementasi Chatbot 'Alitta' Asisten Virtual Dari Balittas Sebagai Pusat Informasi Di Balittas," *Antivirus J. Ilm. Tek. Inform.*, vol. 13, no. 1, pp. 24–35, 2019, doi: 10.35457/antivirus.v13i1.714.

[7] J. Huang, Y. Wang, Y. Wang, Y. Dong, and Y. Xiao, "Relation Aware Semi-autoregressive Semantic Parsing for NL2SQL," *arXiv*, 2021, [Online]. Available: http://arxiv.org/abs/2108.00804.

[8] T. Guo and H. Gao, "Content Enhanced BERT-based Text-to-SQL Generation," *arXiv*, pp. 2–7, 2019, [Online]. Available: http://arxiv.org/abs/1910.07179.

[9] I. Iswandi, I. S. Suwardi, and N. U. Maulidevi, "Penelitian Awal : Otomatisasi Interpretasi Data Akuntansi Berbasis Natural Language Processing," *J. Sist. Inf.*, vol. 5, no. 2, pp. 622–628, 2013.

[10] R. Alamanda, C. Suhery, Y. Brianorman, and J. S. Komputer, "Aplikasi Pendeteksi Plagiat Terhadap Karya Tulis Berbasis Web Menggunakan Natural Language Processing Dan Algoritma Knuth- Morris-Pratt," *J. Coding, Sist. Komput. Untan*, vol. 04, no. 1, 2016.

[11] K. K. Purnamasari and I. S. Suwardi, "Rule-based Part of Speech Tagger for Indonesian Language," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 407, no. 1, 2018, doi: 10.1088/1757-899X/407/1/012151.

[12] L. Valatehan, "Identifikasi kalimat pemborosan menggunakan rule base reasoning," *Annu. Res.Semin.*, vol. 2, no. 1, pp. 205–208, 2017.

[13] S. Saiful and B. Rahman, "Sistem Informasi Hasil Produksi Dan Penjualan Roti Pada Usaha Dagang Lala Roti Kendari," *Simtek J. Sist. Inf. dan Tek. Komput.*, vol. 1, no. 1, pp. 53–61, 2016, doi: 10.51876/simtek.v1i1.8.

[14] S. Zygiaris, *Structured Query Language (SQL): Data Management*. 2018.

[15] D. M. P. P, K. K. Purnamasari, and J. D. Bandung, "INDONESIAN TEXT TRANSLATOR INTO DML (DATA MANIPULATION LANGUAGE) WITH SUB-QUERY."

[16] K. S. Haryana, "Penerapan Agile Development Methods Dengan Framework Scrum Pada Perancangan Perangkat Lunak Kehadiran Rapat Umum Berbasis Qr-Code," *J. Comput. Bisnis*, vol.

13, no. 2, pp. 70–79, 2019.

[17] M. A. Muslim and N. A. Retno, “Implementasi Cloud Computing Menggunakan Metode Pengembangan Sistem Agile,” *Sci. J. Informatics*, vol. 1, no. 1, pp. 29–37, 2015, doi: 10.15294/sji.v1i1.3639.

[18] I. Rabbani and E. Krisnanik, “E – Commerce Perlengkapan Haji Dan Umroh Berbasis Web Menggunakan Metode Agile Software Development,” *Semin. Nas. Mhs. Ilmu Komput. dan Apl.*, vol. 1, no. 2, pp. 432–443, 2020.

[19] N. Stratica, L. Kosseim, and B. C. Desai, “NLIDB Templates for Semantic Parsing,” *Proc. 8th Int. Conf. Appl. Nat. Lang. to Inf. Syst.*, pp. 235–241, 2015.

[20] J. Karlos, H. Sujaini, and H. Anra, “Konversi Bahasa Indonesia ke SQL (Structured Query Language) dengan Pendekatan Mesin Penerjemah Statistik,” *J. Sist. dan Teknol. Inf.*, vol. 3, no. 1, p. 1, 2016.