Responsive Chatbot Using Named Entity Recognition and Cosine Similarity

Entin Martiana Kusumaningtyas, Ekky Regita Laurentino and Ali Ridho Barakbah
Informatics and Computer Scienece Department, Electronics Engineering Institute of Surabaya, Jl Raya ITS, Surabaya, Indonesia

Keywords: Natural Language Processing, Chatbot, Text Mining, Cosine Similarity.

Abstract: Most of the higher education student service systems in Indonesia still use conventional and web-based system to provide academic information to students. This reduces the effectiveness of students in obtaining academic information. In the conventional system, students must visit the academic service center only to obtain academic information. In the web-based system, students can access information anywhere, but students also have to look for the right menu to obtain the academic information needed. This research proposes a chatbot application to help students of Electronic Engineering Institute o Surabaya (EEPIS) obtain information effectively and interactively. This application allows students to ask questions related to academic regulations, grades, class schedules, and attendance recap. This research processes student question using Named Entity Recognition method, this method can recognize the entities contained in the questions asked by the user. To improve the ability of Named Entity Recognition (NER) method to handle words, we also implement synonym checking and keyword matching process. The result entities of previous process will be calculated its proximity to the keywords in the database using cosine similarity to find the right answer. The result of this study showed an accuracy of 84.5% with testing 84 questions obtain 71 accurate answer.

1 INTRODUCTION

Education in Indonesia is a manifestation of one of the country's goals, namely to make the intellectual life of the nation. Therefore, education is included in the rights of every Indonesian citizen. This is in accordance with the constitutional mandate in the 1945 Constitution of the Republic of Indonesia Article 31 paragraphs (1) and (2). It is also in Law no. 39 of 1999 concerning Human Rights Article (12). To realize advanced education, adequate facilities are needed to support the process of running the education itself, one of which is by holding educational units or schools or colleges. Furthermore, in its implementation, educational units, especially tertiary institutions, not only provide services in the form of teaching and learning activities, but also meet the needs of students to obtain student services to realize a systematic education.

At this time, most universities still use conventional and web-based systems to meet the needs of student services such as schedule information, attendance recap, and grades. In the conventional system, students must ask student service staff to obtain this information. This is considered less effective because of time constraints which result in student service employees at higher education being unable to respond to student needs at all times, especially outside working hours. The limited number of employees also makes the service must implement a queue system. These two things certainly make the implementation of student services not optimal and can reduce the level of student satisfaction with student services at the university.

In a web-based student service system commonly called a student portal, students can get these services anywhere and anytime. However, in practice, to access the student portal, many processes are required, such as searching and selecting menus to obtain appropriate information. When compared to administrative services with conventional systems, the conventional system definitely will be easier because users can directly convey the required academic information. Not only that, on a web-based system, there are also some data provided in the form of documents, such as data related to academic regulations. Hence, students have to read the details of the document to obtain the information needed.

Kusumaningtyas, E., Laurentino, E. and Barakbah, A.
Responsive Chatbot Using Named Entity Recognition and Cosine Similarity.
DOI: 10.5220/0011756300003575
In *Proceedings of the 5th International Conference on Applied Science and Technology on Engineering Science (iCAST-ES 2022)*, pages 239-245
ISBN: 978-989-758-619-4; ISSN: 2975-8246

Along with the rapid development of technology, many studies have been carried out to develop student service systems. There are various technologies and methods used to optimize the student service system. Among them is the collaboration between chatbots and digital signage to assist students in obtaining information related to lecture schedules, seminars, and alerts that can be obtained through digital boards and chatbots. While the proposed methods in optimizing student service systems include the K-Nearest Neighbor method with the dataset used is FAQ data which allows the system to find answers to questions asked by users via chatbots. The method that has also been proposed in previous research is context recognition implemented on chatbot with the dataset used is new student admissions data, thus enabling the system to provide answers to questions related to new student admissions submitted via chat.

The objective of this research that will be carried out is to implement named entity recognition method on chatbot system to provide student services related to schedule information, attendance recap, grades, and academic regulations. The major contribution of this research is to model the chatbot system to help student get the academic information based on proposed named entity recognition method.

2 RELATED WORK

Student services is one of the important sectors in the implementation of education. Student can found out the information related to their study through such as schedules, lectures, and grades through student services. Many researchers have conducted research to improve student services system. Rio Junardi et al discussed Chatbot Messenger and digital signage providing academic information services. On digital signage, information will be displayed such as lecture schedules, result seminar schedules, and a comprehension test schedule. In addition, profiles of universities are also displayed as well as several pictures of documentation of activities that have been carried out by these universities. While the chatbot system can be used to request academic information services according to requests by users. Users can type keywords according to the requested data such as location, lecturer, or study program. The chatbot will then provide data according to the keywords provided by the user.

Kristian Adi Nugraha et al from Duta Wacana sChristian University discuss how to build a chatbot to process academic services using the K-Nearest Neighbor method. The chatbot application was built to overcome the problem of decreasing customer service performance due to the limited number of employees or staffs serving. In addition, it also overcame problems related to FAQs that were previously implemented to reduce the customer service workload but made it difficult for users to find the list of FAQs needed. Through this chatbot, users can send questions via chat applications using free language and without a certain format. This chatbot uses the K-Nearest Neighbor method which has been widely implemented to solve problems related to text classification. From this method, answers are taken from the database based on similar questions asked.

Marwan Noor Fauzy et al from Amikom University Yogyakarta propose the academic information service chatbot by using the fuzzy string matching method. The chatbot system in this research is web based and built by using PHP and MySQL database. To ask a question, the user can first access the web then a conversation form and login form will appear. Users are required to login first before asking questions. After the user sends a question, the system will recognize it as input data. Then from the data, the keywords will be searched in it. If the keyword has been found, it will be matched with the data dictionary that has been previously defined using Fuzzy String Matching. Through this method, answers will be obtained based on keywords found from user input.

Rico Arisandy Wijaya build a web service chatbot system by using context recognition and binary cosine similarity methods. The source of data used as a knowledge base in this study is information related to PMB PENS and a list of several questions that may be asked by users related to PMB PENS. In the system built, questions from ussers will be processed by using text mining. Then from the input sentence, only a few keywords will be taken according to what is needed through the context recognition process. This process can speed up the calculation process to find answers using cosine similarity.

In this article there are several informations or uniqueness compared to the related researches mentioned above. This research implement named entity recognition method to provide student services using chatbot technology which allow students of Electronic Engineering Polyechnic Institute of Surabaya to ask several information about academic regulations, recap attendances, schedules, and grades.

3 METHODOLOGY

This research was conducted to develop a chatbot

application with the implementation of named entity recognition method that will be implemented on the student portal of Electronics Engineering Polytechnic Institute o Surabaya. The data used in the knowledge based are academic regulation data and EEPIS student data. The two types of data will give different responses. Academic regulations will answer questions related to academic rules that apply at EEPIS, while student data will answer questions related to grades, class schedules, and attendance recap.

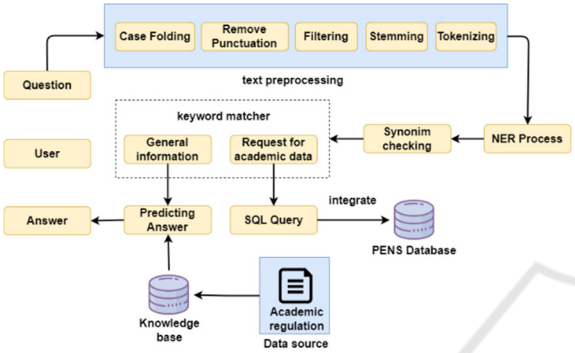

Figure 1: System design.

3.1 Input User

Input is the initial process where the users type in what information they want to find, for example the location of the college, the schedule of courses on certain days in certain classes through a message. Chatbot is designed to simulate a conversation or interactive communication with the user. Therefore, the conversation input is not limited by anything so that the user can freely ask about any academic information. Messages from users can be sent via the chat page which will be provided on the student portal web.

3.2 Text Preprocessing

Text preprocessing is the stage where all user input in the form of text is prepared into data that will be processed further. Text preprocessing aims for uniformity and ease of reading in the next process. This process is the stage of scanning the text of the sentence from the user's question which aims to take important points or keywords. Text preprocessing will perform the process of extracting patterns (useful information and knowledge) from a large number of unstructured data sources, namely questions from the user. In this text preprocessing process, there are several processes to process user input until we get keywords about what information the user wants.

3.2.1 Case Folding

Case folding is the process of converting all letters in the user input sentence into lowercase letters. Only letters 'a' to 'z' are accepted and processed, while numbers and other punctuation marks will be ignored.

3.2.2 Remove Punctuation

Remove punctuation is a process where all punctuation marks contained in the sentence will be removed. This is done because punctuation does not have effect on text preprocessing.

3.2.3 Filtering

Filtering is a process used to retrieve important words in a sentence. Each word in a sentence from the user will be compared with the list of words in the stopword; the stopword contains common words that often appear in everyday language and are considered to have no meaning. Examples of these words are *'yang*/that', *'dan*/and', *'di*/at/in', *'dari*/from', and so on.

3.2.4 Stemming

Stemming is the process of removing affixes to words to produce a root word. One word can often be used in several contexts by giving different suffixes. Therefore, affixed words must be separated from their affixes in order to be processed because the keywords in the database only contain a set of basic words.

3.2.5 Tokenizing

Tokenizing is the process of breaking a sentence into several parts according to the constituent words. Each word generated by the tokenizing process will act as several keywords which will later be compared with several keywords in the database.

3.3 Named Entity Recognition

	Entity 1	Entity 2	...	Entity-n
Sentence	Word 1	Word 2	...	Word-n

	Entity 1	Entity 2	Entity 3
Sentence	Word 1	Word 5	Word 8

Figure 2: NER illustration.

Named Entity Recognition is a process where the system will retrieve several important entities in the sentences entered by the user. The sentence will be a

keyword that will be compared with the keywords in the database and the level of proximity is calculated using cosine similarity to get the answer. An example of the results of the named entity recognition process will be presented in the figure below:

```
Kalimat awal:  Tata tertib mahasiswa itu apa saja

Entity terpilih :  tata tertib mahasiswa apa
```

Figure 3: NER result.

3.4 Synonym Checking

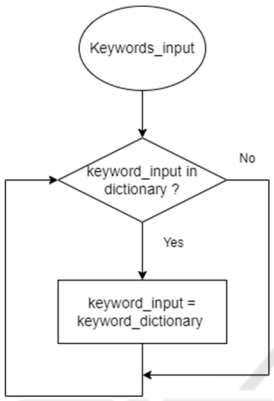

Figure 4: Synonym checking flowchart.

After doing the NER Process on sentences and producing some important words/keywords, the next step is to compare these keywords with words that have the same meaning in the synonym dictionary. Not all words can be entered as keywords in the database. If the keywords contained are not the same as the keywords in the database, then these keywords will not be detected when the search for answers is carried out even though they have the same meaning. Therefore, to expand the knowledge of the system in understanding the intent of the user and providing the right answer, the system will add a synonym dictionary containing keywords and several words that have the same meaning. Then the system will check whether the keyword is in the synonym dictionary and if it is the same, then the word will be replaced with the keyword in the database.

3.5 Keyword Matching

There are two types of questions that are inputted by the user, namely questions about general information, and questions to ask for academic data. Before the process of searching for answers, the types of questions must be known, whether the questions are related to general information or asking for academic data. Rules are needed to check and decide on the type of question. Then after knowing the type of question, the system will be able to determine whether the answer search process will be carried out by involving a database containing general questions, or whether it is necessary to integrate with the PENS database to obtain academic data.

3.6 Predicting Answer

The process of finding answers is done by using the cosine similarity method. This method is used to measure the proximity between the keywords generated after the previous process and the keywords in the database. The cosine similarity method is the right method used in the process of finding answers in this chatbot application because each answer has a different number of keywords, and the advantage of cosine similarity itself is that it is not affected by the short length of a document to be compared so that high accuracy results are obtained from the results of calculations using cosine similarity in comparison of keywords. The following is an example of the cosine similarity calculation process with Table 2 is a sample data of user input and a sample list of questions in the database. While the details of calculating cosine similarity for the sentence "Ada berapa program studi D3 PENS ?" which is calculated for its proximity to the keyword data " program studi d3 pens " is described in Table 2.

Table 1: Cosine similarity result.

Input user	Keyword database	Cosine similarity results
Ada berapa program studi D3 PENS ?	program studi d3 pens	0.75
	program studi d4 pens	0.5
	program studi magister pens	0.4472135
	sosial media pens	0.0
	daftar pens	0.0

Input from the user will go through a text preprocessing process and check synonyms so that several keywords are produced which will later be calculated for cosine similarity with keywords in the database.

Table 2: Cosine similarity calculation.

Keyword	Input (x)	Database (y)	x^2	y^2	(x*y)
berapa	1	0	1	0	0
program	1	1	1	1	1
studi	1	1	1	1	1
d3	1	1	1	1	1
pens	0	1	0	1	0
Jumlah			4	4	3

$$||x|| = \sqrt{1^2 + 1^2 + 1^2 + 1^2} = 2$$
$$||y|| = \sqrt{1^2 + 1^2 + 1^2 + 1^2} = 2 \quad (1)$$

$$\text{Cosine Similarity}(x,y) = \text{Dot product}(x, y) / ||x|| * ||y|| = 3 / 2 * 2 = 0{,}75 \quad (2)$$

3.7 Database System

Database system/Database Management System (DMS) is a system or software designed to manage a database and perform operations on data requested by multiple users. The data referred to in this study are data related to case studies that will be tested later, namely data relating to general university information and academic regulations for students of the Surabaya State Electronics Polytechnic. This data will be used to meet the information needs desired by the user.

4 EXPERIMENT RESULT

Database system/Database Management System (DMS) is a system or software designed to manage a database and perform operations on data requested by multiple users. The data referred to in this study are data related to case studies that will be tested later, namely data relating to general university information and academic regulations for students of the Surabaya State Electronics Polytechnic. This data will be used to meet the information needs desired by the user. The chatbot application testing process is divided into several stages according to the experimental parameters to be tested. Among them are text preprocessing processes including case folding, remove punctuation, filtering, stemming and tokenizing. Then, it is continued with a trial of finding answers using cosine similarity to get answers related to academic regulations and testing to get student data through SQL Query. The trial will be carried out by using sample data from user input. The following will explain the steps in the testing process using sample user input data.

Table 3: Case folding result.

Input Sentence	Case Folding Results
Ada berapa jumlah program studi di PENS ?	ada berapa jumlah program studi di pens ?
PJJ itu apa ya ?	pjj itu apa ya ?
Apa saja program studi di D3 PENS ?	apa saja program studi di d3 pens ?
Saya mau tanya nilai dong	saya mau tanya nilai dong
Saya mau tanya jadwal kuliah	saya mau tanya jadwal kuliah

From Table 3 it can be seen that the case folding process can give the right results. All sentences were successfully uniformed into lowercase. The next stage is remove punctuation, where the results of the folding case will remove all punctuation marks contained in it.

Table 4: Remove punctuation result.

Case Folding Result	Remove Punctuation Result
ada berapa jumlah program studi di pens ?	ada berapa jumlah program studi di pens
pjj itu apa ya ?	pjj itu apa ya
apa saja program studi di d3 pens ?	apa saja program studi di d3 pens
apa saja persyaratan untuk daftar ulang ?	apa saja persyaratan untuk daftar ulang
saya mau tanya nilai dong	saya mau tanya nilai dong
saya mau tanya jadwal kuliah	saya mau tanya jadwal kuliah

From Table 4 it can be seen that the remove punctuation process can give the right results, all punctuation marks contained in the sentence can be erased.

Table 5: Filtering result.

Remove Punctuation Result	Filtering Result
ada berapa jumlah program studi di pens	berapa jumlah program studi pens
pjj itu apa ya	pjj
apa saja program studi di d3 pens	program studi d3 pens
apa saja syarat untuk daftar ulang	syarat daftar ulang
saya mau tanya nilai dong	nilai
saya mau tanya jadwal kuliah	jadwal kuliah

From Table 5 it can be seen that the filtering process can run well. Words marked with yellow are words that are in the stopword list and are not needed in the search for answers. The word was successfully removed through the filtering process. From this process, it can be analyzed that the stopword list needs to be updated periodically according to the data that continues to grow so that the system can run optimally. The next stage is stemming, the results of the filtering will be processed to remove the affixes contained in some words. Sample data will be presented in Table 6.

Table 6: Stemming result.

Filtering Result	Stemming Result
berapa jumlah program studi pens	berapa jumlah program studi pens
pjj	pjj
program studi d3 pens	program studi d3 pens
persyaratan daftar ulang	syarat daftar ulang
nilai	nilai
jadwal kuliah	jadwal kuliah

From Table 6 it can be seen that the stemming process can run well, namely changing the affixed word "requirement" to the basic word "condition". Once done, the next stage of testing is the NER Process. Sample data for questions and entities generated will be presented in Table 7.

Table 7: NER result.

Questions	Entities
Ada berapa jumlah program studi di PENS ?	berapa, jumlah, program, studi, pens
PJJ itu apa ya ?	pjj
Apa saja program studi di D3 PENS ?	program, studi, d3, pens
apa saja persyaratan untuk daftar ulang ?	syarat, daftar, ulang
Saya mau tanya nilai dong	nilai
Saya mau tanya jadwal kuliah	jadwal

From Table VII it can be seen that each question will generate several entities. The resulting entity will be used as a keyword that will be used to find answers.

The nexst testing stage is finding answers using cosine similarity. Details of the results of the calculation of the cosine similarity between the keywords from the user and the keywords in the database will be presented in Table VIII.

Table 8: Cosine similarity experiment.

Input user	Keyword database	Cosine similarity result
Berapa jumlah program studi yang ada di PENS ?	jumlah program studi pens	0.816496580927
	program studi d3 pens	0.612372435695
	program studi paling minat pens	0.547722557505
	lokasi pens	0.288675134594
	psdku pens kota mana	0.204124145231

In Table 8 it can be seen that the keywords with the highest cosine similarity results are found in the keyword "total of pens study programs." These keywords are highly close to the keywords generated by user input. In addition, the answers to these keywords are answers that match the user's questions. It can be stated that the process of finding answers using cosine similarity is running well. The last stage is testing the accuracy of the system in answering questions. The sample data used are 84 questions asked by the user to the system. The chatbot can answer as many as 71 questions correctly. The result of experiment will be shown on figure below.

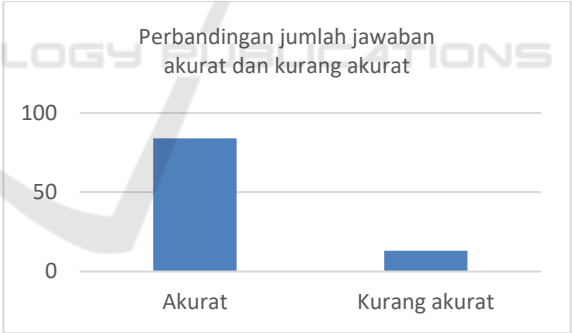

Figure 5: Accuracy experiment result.

5 CONCLUSION

In this study, the author has implemented text processing and named entity recognition method that has been improved by synonym checking and keyword matching process according to the needs of the chatbot to get the appropriate answers. The author also shows how the accuracy of the use of synonym checking based on the dictionary and also the input manually. The accuracy of this study reached 84%. The pre-processing text manually optimize the NER

method to recognize the entities contained in the question. The use of synonym checking can improve system to recognize words that didn't exist in the knowledge base. Dictionary data synonyms also need to be updated along with foreign words that appear so that answers become more accurate. The weakness of using synonym dictionaries is that the admin or the party that makes the decision must have sufficient knowledge, and adding synonyms can sometimes ruin other word or even other synonyms if any duplication synonym word in different keyword.

In the process of finding answers, the authors implement binary systems for cosine similarity that is highly recommended for chatbots because from user questions on chat forums have a straight forward purpose (directly mention the core of the question) because how much the words is not for computing but more focuses on the set of keywords in the question. We test 84 sample data questions and obtain 71 accurate answer.

REFERENCES

Badan Pusat Statistik, Potret Pendidikan Indonesia Statistik. (2019). Pendidikan Indonesia 2019, *Badan Pusat Statistik*, 1st Edition.

Badan Pusat Statistik, Statistik Pendidikan Tinggi Tahun 2019 (2019). *Pusat Data dan Informasi IPTEK DIKTI*, 4th Edition.

Rio Jumardi, Lia Farokhah, Maghfirah. (2020). Kolaborasi Digital Signage dan Chatbot Messenger Sebagai Layanan Penyedia Informasi Akademik, Jurnal Media Informatika Budidarma, Vol. 04, No. 02, Hal. 347-354, STMIK Budi Darma Medan.

Ariyan Zubaidi, Ramdani (2019). Layanan dan Informasi Akademik Berbasis Bot Ttelegram Di Program Studi Teknik Informatika Universitas Mataram, Jurnal Teknologi Informasi Komputer dan Aplikasinya, Vol. 1, No. 1, Program Studi Teknik Informatika Universitas Mataram.

Migunani, Kevin Aditama (2020). Pemanfaatan *Natural Language Processing* dan *Pattern Matching* Dalam Pembelajaran Melalui Guru Virtual, Jurnal Elektronika dan Komputer, Vol. 13, No. 1, Hal. 121-133, STMIK ProVisi Semarang.

Ananda Dwi R, Firdha Imamah, Yusuf Mei Andre S, Ardiansyah. (2018). Aplikasi *Chatbot (MMILKI BOT)* Yang Terintegrasi Dengan *Web CMS* Untuk *Customer Service* Pada UKM MINSU, Jurnal Cendikia, Vol. XVI, Akademi Manajemen dan Informatika DCC.

G. M. D'silva, S. Thakare, S. More, and J. Kuriakose. (2017). Real World Smart Chatbot for Customer Care Using a Software as a Service (SAAS) Architecture, I-SMAC (IoT in Social, Mobile, Analytics, and Cloud (I-SMAC) on 2017 Internatioonal Conference on IEEE, Palladam, India, Hal. 658-664.

Dayinta Warih Wulandari, Putra Pandu Adikara, Sigit Adinugroho. (2018). Named Entity Recognition (NER) Pada Dokumen Biologi Menggunakan Rule Based dan Naïve Bayes Classifier, Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, Vol. 2, No. 11, Hal. 4555-4563, Fakultas Ilmu Komputer Universitas Brawijaya.

Kanya, N., & Ravi, T. (2012). Modelings And Techniques in Named Entity Recognition: an Information Extraction Task. IET Chennai 3rd International Conference on Sustainable Energy and Intelligent Systems (SEISCON 2012), Tiruchengode, Tamilnadu, India.

Kristian Adi Nugraha, Danny Sebastian. (2021). Chatbot Layanan Akademik Menggunakan K-Nearest Neighbor, Jurnal Sains dan Informatika, Vol. 7, No. 1, Program Studi Teknik Informatika Politeknik Negeri Tanah Laut.

Marwan Noor Fauzy. (2019). Kusrini, Chatbot Menggunakan Metode Fuzzy String Matching Sebagai Virtual Assistant Pada Pusat Layanan Informasi Akademik, Jurnal INFORMA Politeknik Indonusa Surakarta, Vol. 5, No. 1, Politeknik Indonusa Surakarta.

Rico Arisandy Wijaya, Entin Martiana Kusmaningtyas, Aliridho Barakbah. (2019). Knowledge Based Chatbot With Context Recognition, 2019 International Electronics Symposium (IES), Surabaya, Indonesia, Hal. 432-438.