# AN CENTERED SENTENCE-BASED CLUSTERING APPROACH FOR QUESTION AND ANSWERING IN FRONT OFFICE

Zamaludin Abdulah[1], Ken Kinanti Purnamasari[2]
[1,2] Informatic Engineering – Indonesia Computer University
Jl. Dipatiukur 114-116 Bandung 40132
E-mail : zamaludin@live.com[1], ken.kinanti@email.unikom.ac.id[2]

## ABSTRACT

Research on a question and answer system has been done previously by using Centered Sentence-Based Clustering method. The study used English data. It also became one of the basic researches undertaken. In addition, the question and answer system should be able to handle the existing questions on the front office environment as well as to handle the case of out of vocabulary (OOV). In this research the built system has two main stages, namely Clustering and searching for answers. Clustering is the initial stage in which the training data will be processed so as to produce a rule that will be the comparison in the search phase of the answer. Clustering will categorize the question, to then look for answers that exist in the database based on the proximity of the question with the sentence in the database. The result of this main process is the pre-filled answer on each training data. The results show the accuracy obtained on the clustering approach for question and answer system in the case of the front office got an accuracy of 81.08%. The accuracy value is due to lack of training data on the clustering process.

**Keywords :** question answering, clustering, centered-sentence based clustering, rule-based, Bahasa Indonesia

## 1. INTRODUCTION

In a previous research, conducted testing on the question and answer system using Rule-Based method. The results obtained that the system can answer 25.88% questions correctly. The study tried to combine word-matching and rule-based methods [2]. Problems arising from the method are the lack of rules and when there is no suitable rule for the question then the answer will not be obtained even though the intent of the question is similar to that of the rule or it can be said that the research has not been able to handle the case of Out of Vocabulary (OOV).

To solve the OOV problem, some studies have used clustering to get answers to the questions given. By taking advantage of the closeness between questions, clustering can minimize the occurrence of OOV [3]. As in other studies using the Centered Sentence-based (CESED) clustering approach method on a question and answer system that uses rule-based method, the result is an accuracy score of 84.62% type "WHERE-type questions" [4].

## 2. CONTENT OF STUDY

This section can be divided into several sub-points of discussion in accordance with the needs of writing.

### 2.1 Q&A

The Q & A system is a computer system that contains an automated way of answering questions using natural language commonly used by humans. This answer contains information coming from a database source [1]. A question and answer system is expected to help users to get information without having to ask other people or related people. Even examples of existing products such as Kata.ai, question and answer systems serve as personal assistants for users who can do things that can generally be done by personal assistants such as bookkeeping transport tickets and hotels [7]. In this research will also build a question and answer system for the case in Front Office Unikom.

### 2.2 Clustering

Basically, clustering of data is a process for classifying a set of data without a pre-defined class attribute, based on the conceptual clustering principle of maximizing and also minimizing intra-class resemblances [10]. For example, a collection of commodity objects can first be clustered into a set of classes and then into a set of rules that can be derived based on a particular classification. In this study, clustering is used as the main process in the formation of question and answer system rules.

### 2.3 Centered Sentence-based Clustering

The learning strategy of CESED is to study the rules in a section of the given training data sample.

**Table 2. 1 Example of Train Data**

| Pair No | Main-Sentence | Co-Sentence | Cosine |
|---|---|---|---|
| | Question No. | Question No. | |
| 1 | 1 | 2 | 0,73 |
| 2 | 1 | 3 | 0,721 |
| 3 | 1 | 4 | 0,757 |
| 4 | 2 | 3 | 0,95 |
| 5 | 2 | 4 | 0,803 |
| 6 | 3 | 4 | 0,816 |

Main-Sentence is a sentence that will be selected center cluster based on the sequence of sentence questions and co-sentence is a candidate sentence which will be calculated whether to enter into the cluster or not.

Clustering starts from counting the question sentences from the training data. Main-sentence will be a center on a cluster. In other words, the purpose of this clustering process is to judge Co-sentence to be similar to Main-sentence compared to all other Co-sentences. At this point, the threshold value has been determined.

**2.4 TFIDF**

The TF-IDF method is a method to calculate the proximity weight of each question sentence. This method is also notoriously efficient, easy and has accurate results [6]. This method will calculate the value of Term Frequency (TF) and Inverse Document Frequency (IDF) on each token (word) in each document in the corpus. This method will calculate the weight of each token t in document d by the formula:

$$idf = \log(\frac{N}{df})$$

D is the number of documents containing terms (t) and idf is the number of occurrences (frequencies) term to D. The algorithm used to calculate the weight (w) of each document against the keyword (query) using the following equation.

$$w_{d.t} = tf_{d.t} * IDF_t$$

Where :
D: d document
t: the th term term of the keyword
tf: term frequency / frequency of words
w: the weight of the dth document against the th term

**2.2 *Cosine Similarity***

Cosine similarity is used to calculate the query relevance approach to the sentence. The greater the value of the query vector similarity with the document vector the query is considered more relevant to the document. Cosine of 00 is 1, and less than 1 for every other angle. Thus, two vectors of the same orientation have cosine similarity of 1, the two vectors in 900 have the similarity of 0. Cosine similarity is mainly used in positive space, where the result is limited (0.1). Cosine similarity then gives a benchmark how similar two documents [7].

Cosine of the two vectors can be derived using the Euclidean dot product of 1.b = || a || || b || cos Θ where the attribute vector, A and B of similarity and magnitude are given by equation (2.3).

$$CS(b_1, b_2) = \frac{\sum_{t=1}^{n} W_{t.b1} W_{t.b2}}{\sqrt{\sum_{t=1}^{n} W_{t.b1}^2} \sqrt{\sum_{t=1}^{n} W_{t.b2}^2}}$$

**2.3 Problem Analysis**

Based on the predecessor of research on the approach centered sentence-based clustering for question and answer system in the case of front office has never been done before. Based on the above, the process analysis is described in Figure 2.1

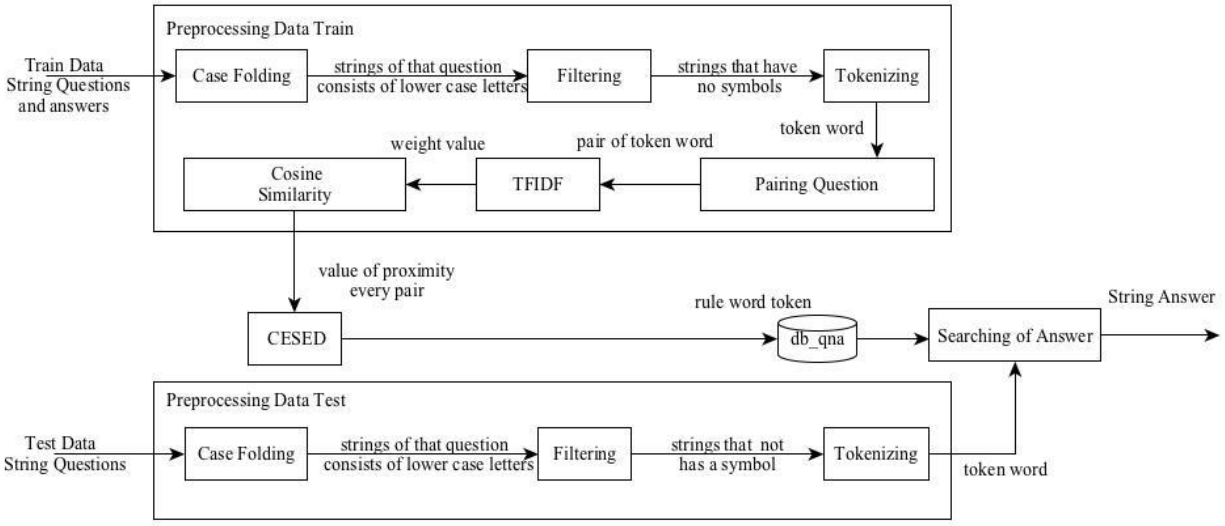



**Figure 2.1 Process Scheme**

### 2.4 Input Data Analysis

Data input to be processed by the system both the training data and test data is taken based on the results and interviews. The file format will be read ie the data in the form of text and the data is inserted into the database.

**Table 2.2 Sample Data Input Questions and Answers.**

| NO | Question | Answer |
|---|---|---|
| 1 | bagaimana cara mendaftar autodebet di unikom? :) | Untuk melakukan autodebet, anda cukup menyetorkan sejumlah uang sesuai dengan jumlah autodebet yang harus dibayarkan ke rekening bank UNIKOM milik anda |
| 2 | Bagaimana jika telat autodebet | Untuk melakukan autodebet, anda cukup menyetorkan sejumlah uang sesuai dengan jumlah autodebet yang harus dibayarkan ke rekening bank UNIKOM milik anda |
| 3 | bagaimana cara mendaftar Kuliah | Pendaftaran dapat dilakukan menggunakan perndaftaran online lalu menyerahkan formulir ke bagian PMB UNIKOM |
| 4 | kenapa Autodebet harus dibayar | Autodebet bertujuan untuk memudahkan pembayaran kuliah dimana mahasiswa cukup mentransfer sejumlah uang sesuai dengan biaya perkuliahan ke rekening mahasiswa tersebut |

### 2.5 *Preprocessing*

In this process, is explained about the data filtering and data extraction so that when the next process the data is clean or ideal for processing and existing features in the data can already be used. An overview of the preprocessing stage can be seen in Figure 2.2.

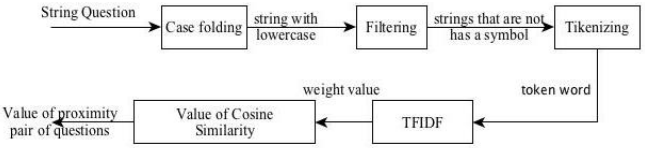



**Figure 2. 2 Preprocessing**

In this study preprocessing has several processes ie:

- Case folding is the process of equalizing the text type, in this study the letter is converted to lowercase so that there is no problem in matching the sentence in the next process.
- Filtering process. In this process, the input data containing symbols and emoticons will be omitted because they are not needed and may hamper the next process, provided that the removal process will remove characters other than the alphabet a-z and 0-9.
- After going through the process of filtering, the next stage of the filtering process will go into the tokenize process. The tokenisasi process in this penlitian is

memisakah each token using a space delimeter. Then the data to be generated from tokenisasi is a set of words or word fractions.

After passing preprocessing then the data generated is the data token on each data train.

**Table 2.3 Sample Data Input Questions and Answers.**

| i | $P_i$ | Before tokenize | After tokenize |
|---|---|---|---|
| 1 | $P_1$ | bagaimana cara mendaftar autodebet di unikom | bagaimana cara mendaftar autodebet di unikom |
| 2 | $P_2$ | ada berapa kah fakultas di unikom | ada berapa kah fakultas di unikom |
| 3 | $P_3$ | berapa biaya kuliah tahun 2016 | berapa biaya kuliah tahun 2016 |
| 4 | $P_4$ | bagaimana cara mendaftar k/uliah | bagaimana cara mendaftar kuliah |

**2.6 *Clustering***

Clustering process that will be done will use CESED method, in this research clustering is used in 2 process that is formation of rule and in search of answer. The process of forming rules using clustering can be seen in Figure 2.4.

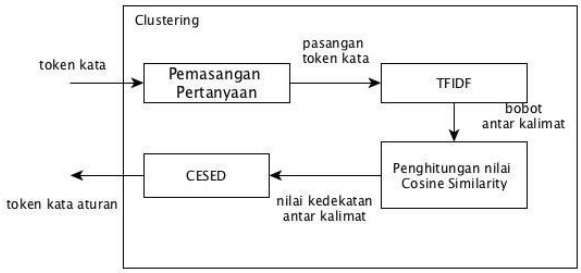



**Figure 2. 4 Clustering Scheme**

To determine the Main Class and Candidate Class, then the process of posting the question dilkaukan. Here are the results of installation of questions. The CESED Main Class is called the Main-sentence and Candidate Class as Co-Sentence.

**Table 2.4 Result of Pairing Question**

| Pasangan No | *Main-Sentence* | *Co-Sentence* |
|---|---|---|
| 1 | $P_1$ | $P_2$ |
| 2 | $P_1$ | $P_3$ |
| 3 | $P_1$ | $P_4$ |
| 4 | $P_2$ | $P_3$ |
| 5 | $P_2$ | $P_4$ |
| 6 | $P_3$ | $P_4$ |

To weight the word, the TF.IDF process is performed. The TF.IDF calculation results can be seen in the following table. In the example TF.IDF used example pair of 3.

Berikut adalah tahapan penghitungan bobot TF.IDF.

1. To fill in a column, count the appearance of a word "autodebet" on every sentence. Word "autodebet" only appears in sentence once, so on $tf - P_1$ value is 1 and tf.S value is 0.
2. Next column df, sum up the word appearance of "autodebet" is once, so $df$ value is 1.
3. Filling column $idf$, ie by the formula (2.1) by entering the value of n (number of words) and $df$ (proccess no.2). Value of $n = 7$, dan $df = 2$, so on idf idf for word "autodebet" $= log(7/2) = 0.35218$
4. Filling column w, ie by the formula (2.2) by entering the value of tf (process no.1) and idf (process no.3). so filling $w - P_1 = tf - P_1.idf = 1.1{,}89209 = 1{,}89209$ and $w - P$ next compute $TF.IDF$ can be seen in Table 2.5.

**Table 2.5 Result of TFIDF**

| Term | W = TF.IDF | | $W_{P1}^2$ | $W_{P4}^2$ | $W_{P1}.W_{P4}$ |
|---|---|---|---|---|---|
| | $P_1$ | $P_4$ | | | |
| bagaimana | 0,544068 | 0,544068 | 0,29601 | 0,29601 | 0,29601 |
| mendaftar | 0,544068 | 0,544068 | 0,29601 | 0,29601 | 0,29601 |
| cara | 0,544068 | 0,544068 | 0,29601 | 0,29601 | 0,29601 |
| autodebet | 0,845098 | 0 | 0,714191 | 0 | 0 |
| di | 0,845098 | 0 | 0,714191 | 0 | 0 |
| kuliah | 0 | 0,544068 | 0 | 0,845098 | 0 |
| unikom | 0,845098 | 0 | 0,714191 | 0 | 0 |
| $\sum$ | | | 3,0306 | 1,60222 | 0,88803 |

After getting the weight value from process TF.IDF the next step is to calculate the value of proximity by using Cosine Similarity method.

**Table 2.5 Result of $W^2$**

| Term | TF | | | IDF | W= TF.IDF | |
|---|---|---|---|---|---|---|
| | P1 | P4 | df | Log(n/df) | P1 | P4 |
| bagaimana | 1 | 1 | 2 | 0,544068 | 0,544068 | 0,544068 |
| mendaftar | 1 | 1 | 2 | 0,544068 | 0,544068 | 0,544068 |
| cara | 1 | 1 | 2 | 0,544068 | 0,544068 | 0,544068 |
| autodebet | 1 | 0 | 1 | 0,845098 | 0,845098 | 0 |
| di | 1 | 0 | 1 | 0,845098 | 0,845098 | 0 |
| kuliah | 0 | 1 | 1 | 0,845098 | 0 | 0,544068 |
| unikom | 1 | 0 | 1 | 0,845098 | 0,845098 | 0 |

Here is the result of cosine similarity calculation on each pair.

**Table 2. 6 Result of Cosine Similarity**

| i | *Main-Sentence* | *Co-Sentence* | *Sim(Main-Sentence, Co-Sentence)* |
|---|---|---|---|
| 1 | $P_1$ | $P_2$ | 0 |
| 2 | $P_1$ | $P_3$ | 0,40299 |
| 3 | $P_1$ | $P_4$ | 0,790021 |
| 4 | $P_2$ | $P_3$ | 0,370568 |
| 5 | $P_2$ | $P_4$ | 0 |
| 6 | $P_3$ | $P_4$ | 0 |

After the proximity value is obtained the next step is to cluster using CESED method. For CESED work steps can be seen in the following flowchart.

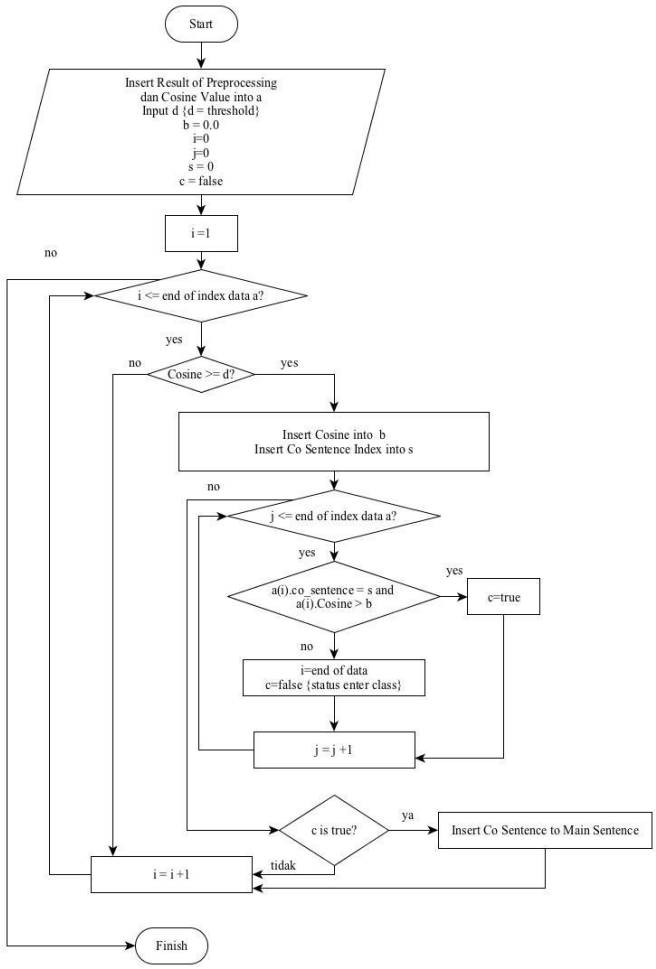



**Figure 2.6 Flowchart CESED**

The first step in the CESED process is to compare each Cosine Similarity value in Table 3.8 and compare it with the predefined threshold value of 0.6. Starting from the first pair whether the value of Cosine Similarity ≥ threshold or not. In the table, can be seen the value of Cosine Similarity on each pair of sentences along with the pair of Main-sentence and Co-sentence. Since main-sentence is measured as a potential center class, the main Sentence will be considered as a cluster and at this stage the class is entered into a variable. Each cluster has a member that is the main-sentence itself.

**Table 2. 8 Dividing Cluster**

| i | *Cluster* | *Co-Sentence* | *Sim(Main-Sentence, Co-Sentence)* |
|---|---|---|---|
| 1 | $C_1$ | $P_2$ | 0 |
| 2 | $C_1$ | $P_3$ | 0,40299 |
| 3 | $C_1$ | $P_4$ | 0,790021 |
| 4 | $C_2$ | $P_3$ | 0,370568 |
| 5 | $C_2$ | $P_4$ | 0 |
| 6 | $C_3$ | $P_4$ | 0 |

Having known the value of Cosine Couple 1, then compared whether the value of Cosine Similarity is ≥ 0.6. Because the value of Sim <0.6 then continue to the next pair Cosine value. Because the value of COSINE <threshold then proceed to the next pair Cosine value. Because the COSINE threshold value is checked whether the COSINE value is the greatest value of all compared to other COSINE values in the same Co-Sentence. Because the COSINE value of pair 5 ≤ the value of COSINE pair 3 then the comparison is done back on the pair that has the value of co-sentence = 4. Because the value of COSINE pair 6 ≤ COSINE value pair 3 then and no more couples that have co-sentence value 4 then clustered with group. After that, the pair that has the value of co-sentence = 4 will not compare again because it has entered into the cluster question 1 is continued by calculating the next COSINE value.

**Table 2. 9 Clustering Process Results**

| Cluster | Co Sentence |
|---|---|
| $C_1$ | $P_1$ |
| | $P_4$ |
| $C_2$ | $P_2$ |
| $C_3$ | $P_3$ |

From the clustering results, into a new rule for the question and answer system is taken from the remaining questions after passing clustering, the remaining question data that is $P_1$, $P_2$ and $P_3$. Then, the rules will appear as in the table 2.10

**Table 2. 10 Rules that Appear After Clustering**

| i | Pertanyaan | Jawaban | $R_i$ |
|---|---|---|---|
| 1 | bagaimana cara mendaftar autodebet di unikom | Untuk melakukan autodebet, anda cukup menyetorkan sejumlah uang sesuai dengan jumlah autodebet yang harus dibayarkan ke rekening bank UNIKOM milik anda | $R_1$ |
| 2 | ada berapa kah fakultas di unikom | Unikom memiliki 6 Fakultas dan 24 Program Studi | $R_2$ |
| 3 | berapa biaya kuliah tahun 2016 | Untuk Semester ganjil biaya kuliah sebesar 5.300.000 untuk fakultas teknik | $R_3$ |

**2.6 Searching of Answers**

In this process, the process of Preprocessing is done again but to the test data. After that, the results of preprocessing will go into the search process answers and will generate answers from the system.

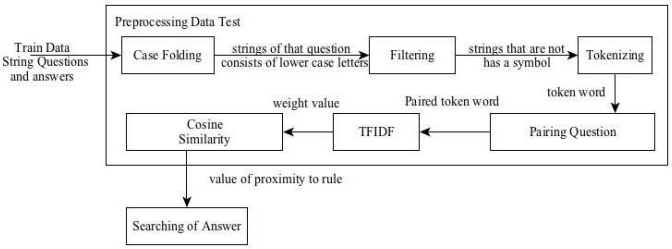



**Figure 2. 7 Searching of Answer**

Here is an example of test data that has passed through preprocessing. Can be seen in table 2.9.

**Table 2. 11 Result of Preprocessing Test Data**

| i | Test Data Before Preprocessing | After Preprocessing | $U_i$ |
|---|---|---|---|
| 1 | Bagaimana cara mendaftar Autodebet? | bagaimana cara mendaftar autodebet | $U_1$ |

After preprocessing into the search process the answer by performing cosine similarity to the test data with the rule on the system of frequently asked questions.

Here is the result of cosine similarity test data calculation of the rule in table 2.8 where the test data is symbolized by $U_1$ and rules symbolized by $R_i$.

**Table 3. 1 Result of TF.IDF**

| Term | TF | | | IDF | W = TF.IDF | |
|---|---|---|---|---|---|---|
| | $U_1$ | $R_4$ | df | Log(n/df) | $U_1$ | $R_1$ |
| bagaiman a | 1 | 1 | 2 | 0,544068 | 0,544068 | 0,544068 |
| mendaftar | 1 | 1 | 2 | 0,544068 | 0,544068 | 0,544068 |
| cara | 1 | 1 | 2 | 0,544068 | 0,544068 | 0,544068 |
| autodebet | 1 | 0 | 1 | 0,845098 | 0,845098 | 0 |
| di | 1 | 0 | 1 | 0,845098 | 0,845098 | 0 |
| kuliah | 0 | 1 | 1 | 0,845098 | 0 | 0,544068 |

**Table 3. 2 Result**

| $U_1^2$ | $R_1^2$ | $U_1 . R_1$ |
|---|---|---|
| 0,29601 | 0,29601 | 0,29601 |
| 0,29601 | 0,29601 | 0,29601 |
| 0,29601 | 0,29601 | 0,29601 |
| 0,714191 | 0 | 0 |
| 0,714191 | 0 | 0 |
| 0 | 0,845098 | 0 |
| 2,316412 | 1,733128 | 0,88803 |

**Table 2. 12 Result Compute Cosine Data Test**

| Data Test | Rule | Cosine Similarity |
|---|---|---|
| $U_1$ | $R_1$ | 0,877971 |
| $U_1$ | $R_2$ | 0,329254 |
| $U_1$ | $R_3$ | 0,246598 |

After that, sought the value of the largest cosine similarity of table IV.8, then the greatest value is jawbannya. Here is the answer generated by the system.

**Table 2. 13 Answer by system**

| No | Question | Answer |
|---|---|---|
| 1 | Bagaimana cara mendaftar Autodebet? | Untuk melakukan autodebet, anda cukup menyetorkan sejumlah uang sesuai dengan jumlah autodebet yang harus dibayarkan ke rekening bank UNIKOM milik anda |

## 3. CLOSING

In this section, will be divided into two parts namely the conclusions and research suggestions.

### 3.1 Conclusion

The conclusion gained during this research about center sentence-based clustering approach, it is found that this CESED method can be implemented in question and answer system with front office case. Where different things with other clustering methods, CESED is used to categorize similar questions and build rules automatically for question and answer systems.

After testing the following test data is the analysis of test results against 37 test data.

1. There is a close answer but the resulting answer is wrong as in the test result No. 32. The question in question is correct the intent of the answer but the detail of the answer is wrong.
2. The answers provided in the training data are incomplete.
3. A similar question can be ambiguous when a word that the system detects is the same but in its description it is different like “kapan pembayaran kuliah” with “kapan pembayaran autodebe” will be ambiguous by the system when it is detected only “kapan” and “pembayaran”

Test results as follows:

Number of correct answers = 30
Number of Question = 37

$$accuracy = \frac{\text{Number of correct answers}}{\text{Number of Question}} x100\%$$
$$accuracy = \frac{30}{37} x100\%$$
$$accuracy = 81.08\%$$

Based on these results, question and answer system was tested using 42 training data obtained accuracy of 81.08%.

### 3.2 Suggestion

Based on the results of the analysis and conclusions, there are several suggestions for improvements to this research, such as the following:

1. Weights do not use the word but use w-shingling so the weighting is done based on the input of many w.

## BIBLIOGRAPHY


[1] Ewerton J. Wantroba and Roseli A.F. Romero, "An Interactive Question-Answer System with Dialogue for a Receptionist Avatar" *SBR Brazilian Robotic Symphony*, 2016.

[2] Romaida Dolarosa Sianturi, "Implementasi Question Answering System dengan Metode Rule-Based pada Banyak Dokumen Berbahasa Indonesia," Institut Pertanian Bogor, 2008.

[3] Thomas Müller and Hinrich Schütze, "Improved Modeling of Out-Of-Vocabulary Words Using Morphological Classes", *Proc. 49th Annu. Meet. Assoc. Computer Linguistic Human Language Technology short Paper*, no. 1, pp. 524–528, 2011.

[4] Shen Song and Yu-N Cheah, "An unsupervised center sentence-based clustering approach for rule-based question answering" *Isc. 2011 - 2011 IEEE Symphony Computer Informatics*, pp. 125–129, 2011.

[5] Zainal A. Hasibuan, "Metodologi Penelitian Pada Bidang Ilmu Komputer Dan Teknologi Informasi", Universitas Indonesia. 2007.

[6] Muhammad Mustaqhfiri, Zainal Abidin, and Ririen Kusumawati, "Peringkasan Teks Otomatis Berita

Berbahasa Indonesia Menggunakan Metode Maximum Marginal Relevance" *Matics*, vol. 4, no. 4, pp. 134–147, 2011.

[7] Abdul Munif, "*Basis Data*" Direktorat Jendral Peningkatan Mutu Pendidik dan Tenaga Kependidikan, Jakarta. 2013.