TEMPLATE BASED QUESTION GENERATOR FROM DOCUMENT SUMMARY USING LEXICAL CHAIN AND CORRELATION BETWEEN SENTENCES

Sani Saefurochman

Informatic Engineering – Indonesian Computer University
Jl. Dipatiukur 112-114 Bandung
E-mail : Saefurochman.sani@gmail.com

ABSTRACT

Technology developments nowadays are rapidly rising that influences all fields, including education. Creating questions from a text to evaluate the reader's understanding of what is presented in the text is something that teachers do to evaluate their students. But humans have limitations in making the question of text that containing many sentences. This is what led researchers to conduct research on how to use the machine in the implementation of intelligent tutoring.

Question Generation (QG) is the process undertaken to create questions by machine that can be accepted by humans from an input. One of a method that is used to generate a question is template-based method. However, this method is considered to have limitations, especially the questions generated are missed from material context. Therefore the research proposed by adding an automatic text summarization to limit the questions to fit the context. Automatic summarization method used is lexical chain and the correlation between sentences.

From the research results, obtained accuracy of questions generated by machine about 70.26% with an accuracy of summarization about 83%.

Keywords: Question Generation, intelligent tutoring, template based, lexical chain, the correlation between sentences.

1. INTRODUCTION

Intelligent Tutoring is one of the topics in the field of artificial intelligence studied by scientists in the field of computer science with the hope that people can learn more quickly gain understanding using the aid of a machine [1]. Generating automatic question included in the study were discussed in intelligent tutoring with the aim to evaluate whether the user is able to understand the contents of a text that has been studied.

Question Generation (QG) is the process undertaken to raise questions that can be accepted by humans of an input with the form of structured (database) or unstructured (text) [2]. Question Generation is a specialization of Natural Language Processing (Natural Language Processing), which also studying how the machine can understand and interact with it using human language.

Based on the results of previous studies of M. Fachrurrozy and Novi Yusliani, the study of the maker of automatic question has long been the subject of Natural Language Processing. The question is raised automatically by the machine Question Generator created, 55.95% of these questions can be accepted [3]. Questions generated by QG contains questions that nonfactoid with constraints that compliance with material context is included as input still has shortcomings thus reducing the accuracy of the right questions. Question to the built generator that is able to make a summary of the material to fit the context of the question and then made peringkasannya results.

In research to be done to increase the percentage of questions that can be accepted to keep it out of the context of the discussion of input to make a summary of the input document first. So expect questions generated by Semantic Question Generator (SQG) produce more accurate in question who related the core of the document inut. With the aim of evaluating whether the user understand the gist of the document that studied and made as input in use SQG. Automatic Text Summarization is a process of making a summary of a document automatically by taking the nucleus from a dokmen. To get a good summary of a text, used methods of lexical chain and correlation of sentences considered better able to make a summary of the level of accuracy (precision) is high [4].

Based on the description of the background described previously, the formulation of the problems found are as follows:

1. How to create a question that can be received in Indonesian automatically from a text document?

2. How Automatic Text Summarization can be implemented in the generation of questions automatically?

The purpose of this study are as follows:

1. Can raise questions that can be received from an Indonesian-language text.
2. Knowing the level of accuracy peringkasan question by implementing automated text to generating automated queries.

2. RESULT AND DISCUSSION

2.1. Question Generator

Question Generation (QG) is the process undertaken to raise questions that can be accepted by humans of an input with the form of structured (database) or unstructured (text). Tasks Generating questions contain multiple subarea [2]. Typically, the approach taken to QG QG depending on application purposes.

Generation question can be done by utilizing a set of documents on the internet (open domain) or a collection of documents at the system itself (closed domain). In general, the question can be classified into five categories, namely the question that is factoid (factoid question), the question is nonfactoid (non-factoid question), question list, yes / no question, and opinion question. This indicates that the questions raised by the automatic query generation systems can be one, several, or all categories of questions [5].

Questions which are non factoid is a question with an answer that is long enough facts which usually consists of one or more clauses. In the manufacture of questions based on predefined templates, there are three classifications for specific words which will refer to the template which will be used to raise questions. Here is a special word that is used to generate questions.

2.2. Automatic Text Summarization

Automatic Text Summarization is the process of summarizing a text document using the help of a computer program by retaining the nucleus of the original document [6]. In general, automatic summarization has three peringkasan techniques, ie extraction-based summarization, abstraction-based summarization, and maximum entropy-based summarization. The technique is often used in cases that summarize this text extraction-based summarization. This technique has a function to give a summary of the results is shorter than the original document. The next technique is a technique of abstraction that is a way of taking the essence of the source text and then make a summary by creating new sentences that represent the essence of the source text in a different form.

2.2.1. Preprocessing

Text mining can be defined as a science in which the user interacts with the document collection from time to time using the tools of analysis. With the analogies into data mining, text mining is used to extract useful information from data sources through identifying and exploring interesting patterns. In the case of text mining, data sources are collections of documents, and an interesting pattern is found not in the database records formal but in a structured textual data in documents in this collection.

a. Text Segmentation

Phase sentence segmentation is the process of separating a string of text documents into a collection of sentences - sentences [7]. The separation process document into sentences using split function (), with a dot ".", "?", "!" As the delimiter to cut the string of documents.

Here is a flowchart of the segmentation stage sentence:

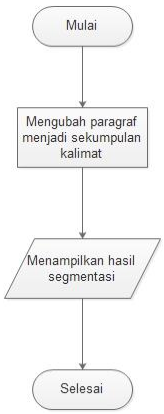


Figure 1. Text segmentation process

b. Text Tokenization

Tokenization stage is the stage of cutting string input into a set of words that compose [7].

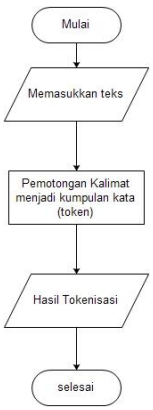
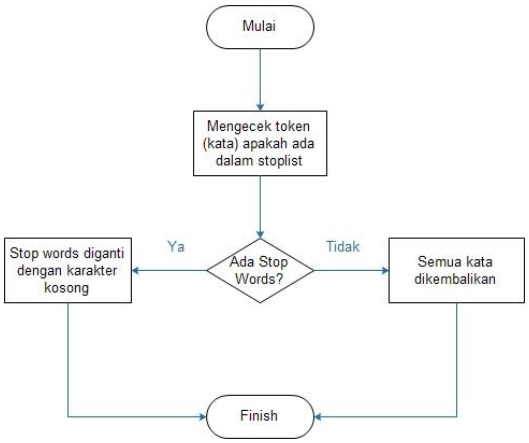


Figure 2. Tokenization Process

c. Stop Words Removal

Stage is the stage of removing stop words removal of words that did not really matter [7]. Every word tokenization result of checking the stop list containing words that are not too important. If it is found in the word tokenization result of the same word in the stop list, then the word will disappear.



Gambar 3. Proses penghapusan *stop words*

d. Stemming

Stemming stage is the stage to change the word of the results of removing stop words into sets of basic words [7]. This stage is not much different from the results stop words because if the base word already exists and there is no change then the result will be the same. This stage will determine the root of each word contained in the phrase so that by the time calculation process which will be calculated is the weight of the weight of the same word. Phase stemming had some sort and as used herein is a process stemming from research Nazief-Adriani. Nazief-Adriani Method are chosen on this research because in the recent research said that nazief-adriani method are better than porter method. The process of stemming can be seen in Figure 4.

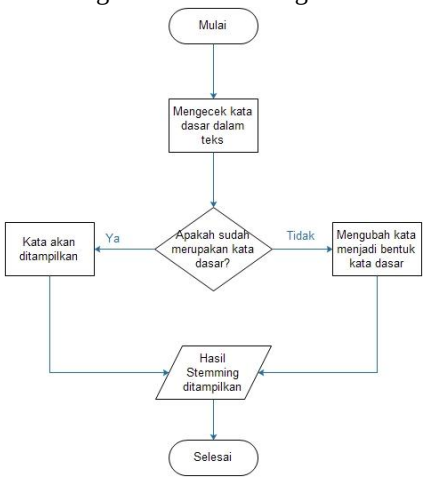


Figure 4. Stemming Process

2.2.2. Extraction Feature

Extraction feature is the processing of sentences using some features so that the values that will be used to summarize any document the extraction features used in this case as follow:

a. Length of Sentences

The length of the sentence is a process for determining the ratio of the number of words that occur in the sentence on the number of words that occur in the document longest sentence. The formula length of the sentence can be seen as follows:

$$S_F1 = \frac{Jumlah\ Kata\ dalam\ Kalimat}{Kalimat\ Terpanjang} \quad (1)$$

b. Weight of Sentences

The weight of the word is the process of counting the frequency of occurrence in the document which has often been used to calculate the words contained in the sentence. Weighting is done by using a calculation tf-idf (term frequency, inverse document frequency). The weight of the sentence can be calculated with the following formula:

$$S_F2 = \mathbf{wi} = \mathbf{TF} \times (\mathbf{log}\frac{\mathbf{N}}{ni}) \quad (2)$$

c. Similarity of Sentences

Similarities between sentences is a word that appears in the same sentence with the word appearing in other words to calculate the similarity, it takes perhitnugan results within a sentence with the following formula:

$$Ki, Kj = \frac{\Sigma Wit \times Wij}{\sqrt{\sum Wit^2} \times \sqrt{\sum Wij^2}} \quad (3)$$

$$S_F4(K) = \frac{\Sigma Sim(Ki,Kj)}{Max(\Sigma Sim(Ki,Kj))} \quad (4)$$

d. Lexical Chain

Lexical chains is an important technique in natural language processing. Lexical chains associated with a sentence that relate to nouns, verbs, adjectives [6]. If the sentence is not related to one another in the text, then the relations between the lexical relate to each other. This is an important concept. Scores from a lexical chain is directly related to the length of the chain whose relationship with the noun. Summarize using lexical chains to choose the most representative sentences.

$$CStrength = Clength \times \{\sum \frac{\Sigma((Tf - Isfwi \times Swi)) + (Tf - Isfwj \times Swj)}{\text{distance between wi and wj}}\} + TotalNumberOfPairs$$

Figure 5. Lexical Chain Formula

2.2.3. Fuzzy Logic

Fuzzy logic was first introduced by Prof. Lotfi A. Zadeh in 1965. Basic fuzzy logic is fuzzy set theory. In the fuzzy set theory, the role of the degree of membership as a determinant of the presence of elements in a set is important. Fuzzy logic can be considered as a black box that connects between the input space to the output space [8]. There are a few things to know in fuzzy systems, including:

a. Fuzzy Variable

Fuzzy variables are variables to be discussed within a system of fuzzy variables. Example: age, temperature, demand, etc.

b. Fuzzy Set

The set of fuzzy variables is a group that represents a particular condition or state of a fuzzy variable. Fuzzy set has two attributes, namely linguistic and numerical.

c. Universal Union

The universe of discourse is the set of real numbers are constantly rising (increases) monotonically from left to right. Value of the universe of discourse can be either positive or negative number.

d. Domain

Domain is a set of real numbers are constantly rising (increases) monotonically from left to right. Domain values can be either positive or negative number.

Fuzzy logic membership functions. The membership function is a curve that shows the mapping of points of input data into membership values (degree of membership) which has the interval between 0 and 1 [8]. One way that can be used to obtain the membership value is through the approach function.

Some of the functions that can be used are input mapping to the degree of membership is described as a straight line. This form is the simplest and be a good choice to approach a concept that is less clear. There are two circumstances that linear fuzzy sets.

First, the linear representation rise is the increase in the value of the set begins domain that has a degree of membership of zero to move to the right toward the domain value having a higher degree of membership. The following is the membership function:

$$\mu(x) = \begin{cases} 0; & x \le a \\ \frac{(x-a)}{(b-a)}; & a \le x \le b \\ 1; & x \ge b \end{cases} \quad (5)$$

Second, the linear representation down is the opposite of the first. A straight line starting from the value of the domain with the highest degree of membership on the left side, then move down to the value of a domain that has a lower degree of membership. The following is the membership function:

$$\mu(x) = \begin{cases} \frac{(b-x)}{(b-a)}; & a \le x \le b \\ 0; & x \ge b \end{cases} \quad (6)$$

2.2.4. Correlation of Sentences

Correlation between sentences is a relationship between one sentence to another where the sentences are related. If a sentence has been linked to any other sentence that the sentence could be a parent or child klaimat sentence for sentence before and after. Clause is useful to explain the sentence stem thus narrowing the meaning of the parent sentence. The linkage between sentences marked by conjunctions in Indoenesian language are as follows;

Agaknya, ... Akan tetapi, ... Akhinya, ... Akiatnya, Artiya, ... Biarpun begitu, ... Biarpun demikian, ... Berkaitan dengan itu, ... Dalam hal ini, ... Dalam hubungan ini, ... Dalam konteks ini, ... Dengan demikian, ... Dengan kata lain, ... Disamping itu, ... Disatu pihak, ... Dipihak lain, ... Jadi,... Jika demikian, ... Kalau begitu, ... Kalau tidak salah, ... Kecuali itu, ... Lagi pula, ... Meskipun begitu, ... Mekipun demikian, ... Namun, ... Oleh karena itu, ... Oleh sebab itu, ... Pada dasarnya, ... Pada hakikatnya, ... Pada prinsipnya, ... Sebagai kesimpulan, ... Sebaiknya, ... Sebaliknya, ... Sebelumnya, ... Sebenarnya, ... Sebetulnya, ...

The Example of system input are as follows.

> Teratai adalah tumbuhan yang hidup di air atau sebagai tanaman air. Oleh Karena itu Apabila kamu perhatikan, bunga teratai nampak indah mengapung di atas permukaan air. Meskipun bunganya mengapung, sebenarnya akar bunga teratai berada di dasar air. Jadi bunga teratai tidak semata mengambang di atas air, melainkan tumbuh dari dasar air. Tangkai daun cenderung tumbuh menjalar, sehingga helai daun selalu terlihat mengapung. Teratai memiliki daun berbentuk bundar dan berukuran sangat lebar.

Figure 6. Text Input

From input text contained in Figure 2, the results based on the method used peringkasan obtained are:

Table 1. Summarization Result

Ki	Sentences
K01	Teratai adalah tumbuhan yang hidup di air atau sebagai tanaman air
K02	Oleh karena itu apabila kamu perhatikan, bunga teratai nampak indah mengapung di atas permukaan air
K05	Tangkai daun cenderung tumbuh menjalar, sehingga helai daun selalu terlihat mengapung

From Table 1 it can be seen that the summary results of the inputs processed there are three sentences the first sentence, second sentence, and the sentence 5. peringkasan done using the 30% rule summary. Where the sentence is taken 2 of 6 core sentence, then having it taken the clause or sentence stem from two basic sentences are based on the method of correlation between sentences. So that the second sentence into a sentence nucleus Because it is a clause of the first sentence. After testing, the material tested from 30 to 10 teachers generated data that can be seen in Table 2.

Table 2. The result of summarization testing

Input	Text Input Title	Sentence result	Sentence approved
1	Bagian Hewan yang Sering Digunakan	12	9
2	Perkembangbiakan Vegetatif Buatan	7	6
3	Manfaat Sumber Daya Alam bagi Manusia	10	8
4	Bunga Pada Tumbuhan	7	6
5	Penebangan Hutan Secara Liar	10	10
6	Hewan yang Terancam Punah di Indonesia	16	13
7	Kemampuan Bertahan Hidup Bunglon	7	7
8	Teratai Sebagai Tanaman Air	8	7
9	Kemampuan Ekolokasi Kelelawar	7	7
10	Kerusakan Lingkungan Hidup	7	6
11	Perkembanganbiakan Vegetatif Alami	13	11
12	Pelapukan pada Batuan	9	9
13	Perubahan Energi Listrik Menjadi Energi Lain	9	9
14	Gaya dan Gerak	7	6
15	Ekosistem	9	7
16	Ciri Khusus Tanaman Kaktus	9	9
17	Pemanfaatan Konduktor dan Isolator	10	7
18	Pelestarian Lingkungan Hidup	9	7
19	Gaya pada Ketapel	6	6
20	Planet	6	6
21	Pengembangbiakan Hewan dan Tumbuhan yang Hampir Punah	8	6
22	Energi Listrik	5	3
23	Penghematan Energi	9	7
24	Sistem Tata Surya	9	7
25	Ciri Khusus Ikan Lele	8	6
26	Jenis Tumbuhan Dikotil	7	5
27	Bagian Tumbuhan yang sering dimanfaatkan	11	8
28	Pembuangan Limbah	6	5
29	Planet Jupiter	7	5
30	Planet Bumi	7	5

Of the total sentence tested, obtained 255 sentences generated by the system and 213 of them declared true sentence is a sentence the essence of the material tested. So it can be said peringkasan conducted has an accuracy rate of 83%.

2.3. Special Words

A special word describes the sentence answer to a question category. Sentences containing the word "karena" it can be used to raise questions categories of reasons. Sentences containing the word "adalah" can be used to raise the question category definitions. While the sentence containing the word "dengan cara" can be used to raise the question category method. The system will classify the first to look for specific words contained in the input to determine which template to be used later. Table 3 shows the specific words each category.

Table 3. Special Words Category

Category	Special words before target	Special words after target
Definition	disebut, dikenal, dinamakan, mendefinisikan	adalah, yaitu, ialah, merupakan, diartikan
Reason	oleh sebab itu, jadi, memungkinkan adanya, , dengan demikian, maka, dikatakan, penyebab terjadinya, sehingga, mengapa, walau demikian, namun demikian	sebab, karena, bertujuan
Methode	dengan cara	berfungsi untuk, berguna untuk

2.4. Question Template

Template question sentence patterns obtained through the analysis of a set of questions. In this study, the questions raised are questions that fall into the category of non-factoid questions Indonesian, namely the definition, rationale, and methods. Therefore, the analysis conducted on a set of non-factoid question in Indonesian. Table 2 shows the template patterns of non-factoid questions sentences in Indonesian.

Table 4. Question template

Category	Question Template
Definition	Apakah yang dimaksud dengan + target kata Apa yang anda ketahui mengenai + target kata
Reason	Mengapa + target kata
Methode	Bagaimana cara + target kata

Categories of questions raised depending on classification results sentence. The information required for the generation question is:

1. Summary of the insert in the form of sentences important to be made of questions.
2. Classification of each word in a sentence containing a particular word. This is done to determine the position of a particular word and the target word in the sentence questions.
3. Knowing the classification of sentences that contain specific words.
4. Knowing the sentence pattern template of questions for each category.

Fourth information is very important for the process pembangkitkan question. If the information obtained by the third sentence pembangkitkan do. Sentence generation process is done by:

1. Determine the template sentence patterns used questions obtained from the determination of specific words and categories specified in the selected sentence.
2. Take the word that comes before or after the specific word to be the target of a question word. This process can be done Because of their classification information of each word in a sentence, said special position, and a special word used.
3. Placing the target word into the sentence pattern template queries. Here is an example of a question that has been Generated:

Table 5. Question Generated by system

Pi	Questions
P1	Apakah yang dimaksud dengan teratai ?
P2	Apa yang anda ketahui mengenai teratai ?
P3	Mengapa oleh ?
P4	Mengapa helai daun selalu terlihat mengapung ?

2.5. Questions Generated Tested by User

Testing the questions generated by the system will be done by giving questionnaires to 10 elementary teachers who will examine several parameters, including:

1. Questions raised system is written correctly.
2. The questions were raised has the answers
3. Questions were raised in accordance with the system context.

These three parameters must be met everything to be able to determine whether the questions generated by the system is substantiated. Just as before, there were 30 materials tested are shown in Table 6.

Table 6. Testing Result

INput	Number Question Generated	Number Question Approved
1	12	9
2	9	7
3	9	8
4	10	10
5	9	10
6	10	9
7	15	10
8	12	5
9	10	10
10	10	7
11	12	10
12	7	9
13	11	7
14	16	9
15	11	9
16	15	4
17	15	12
18	9	9
19	11	5
20	14	4
21	10	8
22	13	6
23	5	8
24	10	10
25	15	7
26	11	5
27	12	7
28	17	9
29	14	9
30	9	9
∑	**343**	**241**

Of the 30 materials tested, the system generates 343 questions and 241 of them are declared righteous. So that the percentage of acceptable accuracy question is really divided Questions The number of questions are raised multiplied by 100 percent. (241/343) * 100% = 70.26%. So from 4:44 table can be concluded that the question generating system is built to produce 343 questions and 241 questions correctly with a percentage of 70.26%. This automatic query generation can still have a high degree of accuracy by selecting the relevant sentences more accurate and with a special word used to enrich and add to the template questions used.

3. CONCLUSION

The conclusions obtained from the research thesis about generating automated queries based on the template of the summary document using lexical chain and Correlation between sentences

1. The accuracy of sentence choosen as summary of the material by using lexical chain and the correlation between sentences that the question will be made is about 83% derived from the selection of the experts in determining the suitability.
2. Automatic Text Summarization applied to the question generation process, generated a number of questions have accuracy of 70.36%.

The advice can be given to improve the accuracy and summarizing questions were performed in this study are:

1. Enriching dictionary for peringkasan circuited, a special word to define the target word, and the template questions.
2. Enrich the base and stop word dictionary list to get a better summary and link chain sentences not only with the same words, but words that have the same meaning.
3. Using stemming techniques that better so text summarization process would have a higher accuracy.

REFERENCES

[1] A. C. Graesser, "Intelligent Tutoring System with COnversational Dialogue," *AAAI,* vol. 22, p. 14, 2001.

[2] G. B. Y. Z. Xuchen Yao, "Semantic-Based Question Generation And Implementation," p. 32, 2012.

[3] M. Fachrurrozy, Novi Yusliani, "SISTEM PEMBANGKIT PERTANYAAN OTOMATIS," *Journal of Research in Computer Science and Applications,* vol. Vol.2, no. No.1, p. 6, 2013.

[4] Kulkarni A.R, S.S.Apte, "An Automatic Text Summarization using Lexical Cohesion and Correlation of Sentences," *International Journal*

of Research in Engineering and Technology, vol. 03, p. 8, 2014.

[5] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval, England: ACM Press, 1999.

[6] M. L. Murphy and M. Ltd, Lexical Meaning, New York: Cambridge, 2010.

[7] F. Damerau, Handbook of Natural Language Processing Second Edition, New Jersey: Taylor and Francis Group, 2010.

[8] H. S. Kusumadewi, Aplikasi Logika Fuzzy Untuk Pendukung Keputusan, Yogyakarta, Yogyakarta: Graha Ilmu, 2010.

[9] S. Upasana Pandey, Chakraverty, B. Juneja, A. Arora and P. Jain, "Semantic Document Classification using Lexical Chaining & Fuzzy Approach," *International Journal of Soft Computing and Engineering (IJSCE)*, vol. 01, p. 5, 2011.