DETECTION OF RELATIONSHIP AMONG SENTENCES WITH TEMPLATE – BASED METHOD IN AUTOMATIC QUESTION GENERATOR

Purba Saripudin[1], Ken Kinanti Purnamasari[2]

[1,2] Indonesia Computer University
Jl. Dipati Ukur No. 102-116 Telp. (022) 2504119, 2506634, 2533603 Fax. (022) 2533754
E-mail: purbasaripudin@yahoo.com[1], ken.kinanti@email.unikom.ac.id[2]

ABSTRACT

Research that produces questions that use syntactic analysis has been carried out but is limited to only one sentence [4], while the question sentence can come from several sentences. The entire process that will be carried out by the system there are two stages, namely preprocessing that uses previous research and the main process. Preprocessing functions to get the type of word, the name of the entity, the type of phrase, the grammatical function in each sentence and the detection of non-definition sentences while the main process is the process of detecting the interrelationship between sentences using rule-based methods and the question-making process with the template method. The results obtained by the system using 50 text input taken from the Ministry of Education and Culture's Electronic School Book, obtained the question accuracy value of 90.80% with a total of 413 questions asked.

Keyword : question generation, detection of connectivity between sentences, *rule-based, template*.

1. INTRODUCTION

Research on automated question-generating systems has been successfully carried out using English and Indonesian texts with various methods. The study uses English text, the methods used include using syntactic transformation [1] and statistical methods [2]. While the research that uses Indonesian-language texts includes using template methods [3][4] and semantic methods [5]. Generating questions for Indonesian texts produces factoid and non-factoid questions.

Based on several studies that have been mentioned, there are Indonesian language studies using template-based methods [3] [4]. In this study, research on generating questions for simple sentences [3] [4] and compound sentences [4]. This study uses template-based methods with syntactic functions to find verbs [3] and grammatical rules [4] that have been determined to generate questions.

From the above research[3] [4], it is known that the question sentences raised are limited to only processing questions from each sentence. In fact, there are questions that can be processed in several sentences, because there are interrelated sentences. These sentences should be processed and used as sentence questions. Processing many sentences can be done by taking the syntax of each sentence, then adjusting to the conditions in each question template rule. This syntactic information is taken by using special rules to detect which sentences are interrelated.

1.1 Goal and Purpose

Based on the existing problems, the purpose of this study is to build a question-generating system that can detect the interrelationship between sentences and generate questions based on several sentences. The purpose of this study is to determine the accuracy of system-generated questions that can generate information questions from several sentences.

1.2 A Method of Software Development

In the construction of the question generating system using the prototype method, because the development of research will continue to be refined with the evaluation of the results of the successful test [6]. The Prototype method is very suitable for implementing a particular method or algorithm in a case [10].

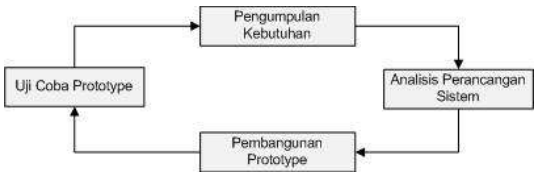


Figure 1. Prototype Method

There are several stages in the Prototype method :

1. Gathering needs
   This stage is the process of identifying system requirements, gathering the required data, collecting dictionary words from various sources and analyzing Indonesian grammar rules.
2. System design analysis
   This stage is a process of method analysis, system design, to determine functional and functional needs, as well as the design of the interface database from the development of this application.
3. Prototype development
   Implementing the system design that has been made. In this study, the software will be built using the PHP programming language.
4. Prototype trial
   This stage is testing the software that was built. If the results of the trial are still found to be errors, the software development process will be corrected.

2. CONTENT OF RESEARCH

The solution to the problem regarding generating questions is limited to just one sentence, this research will carry out the detection process between sentences after the syntactic process. The process of detection between sentences is done by searching for subjects, predicates, and objects in all sentences.

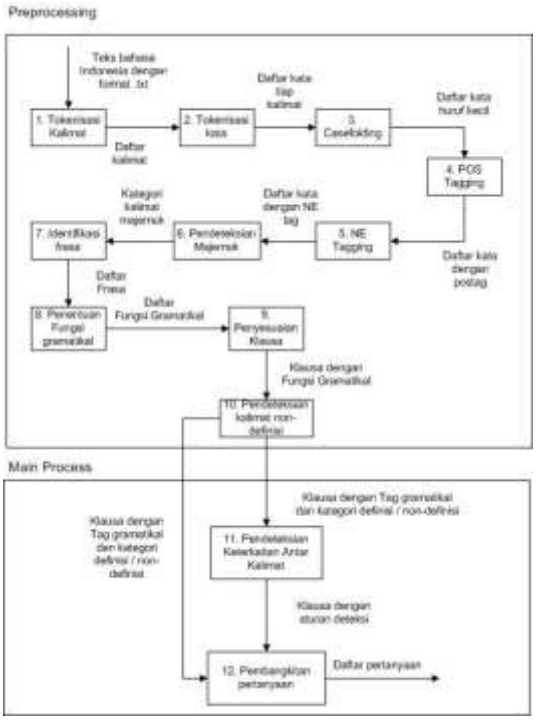


Figure 2. System Flow Chart Block

In this system to be built, the user selects text containing sentences or a set of standard sentences in Indonesian, then this system will process the input text which has the following paths: 1) the system will process the sentence tokenization, 2) the system will process the word tokenisation, 3) The system performs Case folding (homogenizes all into lowercase letters), 4) The system performs word type detection (POS Tagging), 5) the system performs the process of recognizing named entities (NER), 6) the system performs compound sentence detection processes, 7) The system performs the phrase identification process, 8) the system performs the process of determining the grammatical function to determine the subject, predicate, object, description and complement, 9) the system performs a custom clause adjustment process for equivalent compound sentences, 10) the system performs non-definition sentence detection, 11) the system selects many sentences, 12) the system performs the process of generating questions.

2.1 *Preprocessing*

Preprocessing is the initial process carried out by the system, where at this stage the input text will be separated into single words and carried out word type labeling, the introduction of named entities (NER), compound sentence detection, phrase identification, grammatical function determination, clause adjustment and

non-definitional sentence detection. This process is carried out based on previous research [3] [4].

2.1.1 Sentence Tokenization

The sentence characterization is used to break the input string into a collection of sentences based on a period (.), Question mark (?) And an exclamation point (!). The sentence separator mark will not be used in the next process.

2.1.2 Word Tokenization

After the sentences are separated in sentence tokenization, word tokenization is used to separate words in sentences using a space separator, so that a single word will be formed in each sentence.

2.1.3 Casefolding

Casefolding is used to convert all capital letters to lowercase characters. The results of this process are only used in the post tagging process.

2.1.4 POS *Tagging*

After having a lower-case word, the next process that the system will do is detect word type, this process uses the Indonesian NER application [7][8]. This process will automatically generate a word class label on a word, the results of this tag are very influential on subsequent processes, especially the identification of phrases.

2.1.5 NE *Tagging*

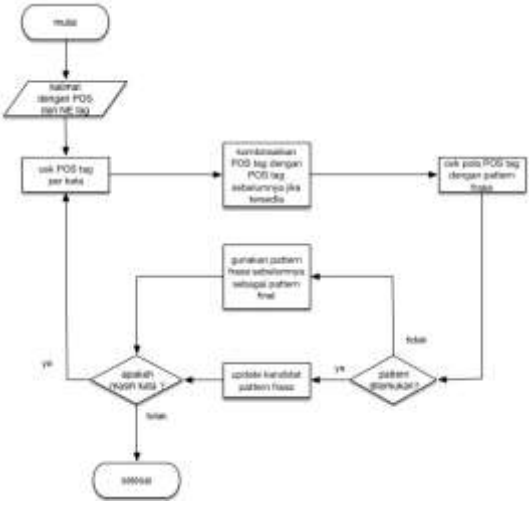
The next process that the system will do is the introduction of named entities, this process uses the Indonesian NER application [7][8]. This process is used to detect people, places, and organizations from a text, the results of this tag are very influential in determining the question template to be used.

2.1.6 Detection of Compound Sentences

The next process that will be carried out by the system is compound sentence detection. This process will detect categories of all sentences, there are three categories that will be defined, namely: equivalent compounds, multilevel compounds, and others. This process is carried out by detecting conjunctions in equivalent compound sentences and multilevel compounds.

2.1.7 Phrases Identify

The next process that will be carried out by the system is the identification of phrases, this process depends on the results of POS tags and the pattern of phrases that have been defined. This process is done sequentially by adjusting the combination of word types (POS tags) with defined phrase patterns. Phase identification phase can be seen in Figure 3 below.



Gambar 3. Flowchart Identify Phrases [4]

2.1.8 Grammatical Function Determination

After getting the phrase in all sentences, the next process is determining the grammatical function. This process is almost the same as the phrase identification process, which is checking each word or phrase from the beginning of the sentence to the end in sequence. The rules for determining grammatical functions can be seen in Figure 4 below.

Figure 4. *Flowchart* Grammatical Function Determination [4]

2.1.9 Penyesuaian Klausa (call reference)

Adjustment of clauses is only done on sentences detected equivalent compound sentences. The function of this process is to satisfy the subject of the second clause with the first clause so that the questions and answers raised in the second clause are not ambiguous. Clause adjustment rules can be seen in Table 1 below.

Table 1. Rules for adjustment clause [4]

No	Rules
1	If the subject of the second clause is a pronoun (PR), the pronoun refers to the subject in the first clause.
2	If the second clause does not have a subject, the subject of the second clause similar to the subject of the first clause.
3	If the object in the second clause is a word reserved (NNG), then the object is equal to the object in the first clause.

2.1.10 Detection of Non-Definitional Sentences

The process of detecting non-definition sentences is used to determine that the sentence that has the words 'is' and 'is' is not always a sentence of definition. This process is carried out sequentially from the first word to the last word by searching for words before and after the keyword. Non-definition sentence detection rules can be seen in Figure 5 below.

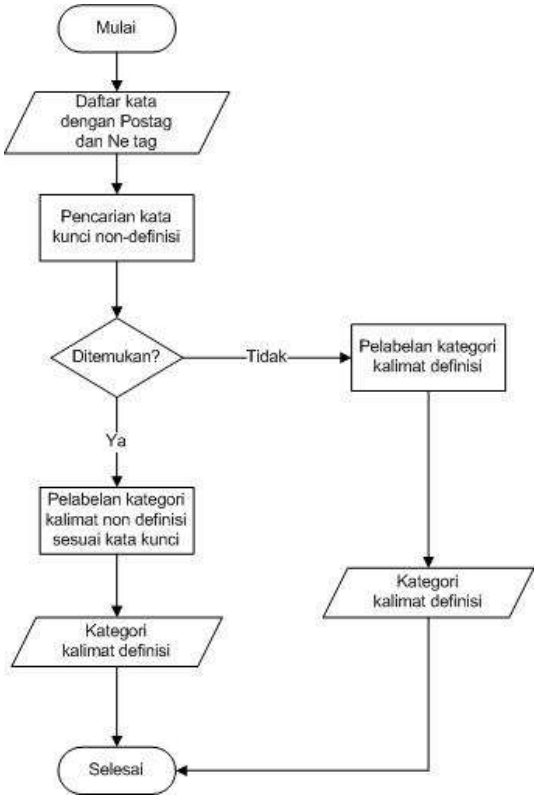


Figure 5. Flowchart Detection of Non-Definitioanal

2.2 Main Process

The main process in this research is the process of detecting the linkages between sentences and the process of generating questions that are carried out by the system after the preprocessing process. The process of detecting the interrelationship between sentences is done to examine which sentences are interrelated, in order to process many sentences.

2.2.1 Detection of Connectivity Between Sentences

This process is done by finding the same subject, predicate, and object in all sentences.

This stage is carried out sequentially, by looking for the same subject in all sentences then looking for predicates and objects. This process will get what sentence conditions will be raised in the question generation process several sentences. The complete condition of detection of the linkages between sentences can be seen in Table 2 below.

Table 2. Detection of Connectivity Between Sentences

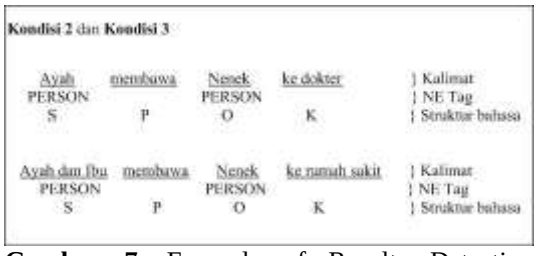
No	Detection
1	1. If the subject is the same / found 2. If the predicate is different
2	1. If the subject is the same / found 2. If the predicate is the same / found
3	1. If the predicate is the same / found 2. If the object is the same / found
4	1. If the subject is the same / found 2. If the predicate is the same / found 3. If the object is the same / found
5	1. If the subject is the same / found 2. If the predicate is the same / found 3. If the object is different

Based on the detection of connectivity between sentences in Table 2, the following Figure 6 will explain the flow of the detection conditions of the interrelations between sentences.

Gambar 6. Flowchart Detection of Connectivity Between Sentence

2.2.2 Question Generator a Few Sentences

The question generation process of several sentences will be carried out using the help of templates based on the detection rules of interrelated sentences. Figure 7 is an example of the detection of the relationship between sentences.



Gambar 7. Example of Results Detection Connectivity Between Sentences

The above clauses are the result of the detection of the interrelationship between sentences, which are included in Condition 2 and Condition 3. The above clause in the NE tagging process can recognize the words Father and Grandma as PERSON, the results of this tag will help in the process of determining the template. The following table 3 is an example of a multiple sentence question generation process based on a question template.

Table 3. Template Question Generation a Few Sentences

Question Mark	Template Question
Siapa	**NE Tag :** PERSON, ORGANIZATION **Condition :** Kondisi 2 ***Template* :** Siapa yang + …? **Question Result :** Siapa yang membawa nenek ke dokter P O K1 dan ke rumah sakit? K2 **Answer :** Ayah
Siapa	**NE Tag :** PERSON, ORGANIZATION **Template :** Siapa yang + …?

	Condition : Kondisi 3 **Question Result:** Siapa yang dibawa ayah dan Ibu P S ke dokter dan ke rumah sakit? K1 K2 **Answer :** Nenek

2.2.3 Question Generation every sentence

Question generation every sentence there are several processes, namely non-factoid detection, non-factoid question generation and factoid question generation.

2.2.3.1 Non-Factoid Detection

In the process of generating questions, sentences or clauses that become input systems will be detected first in the non-factoid detection process. The non-factoid detection process is carried out based on the presence of non-factoid specific keywords. A list of non-factoid keywords can be seen in Table 4 below.

Tabel 4. Keyword Non-Factoid [4]

Category Non-Factoid	Keyword	
	Before Target	**After Target**
Definisi	Disebut, dikenal, dinamakan, mendefinisikan	Adalah, yaitu, ialah, merupakan
Alasan	Oleh sebab itu, jadi, memungkinkan adanya, dengan demikian, maka, dikatakan, penyebab terjadinya, sehingga, walau demikian, namun demikian	Sebab, karena, bertujuan
Cara	Berfungsi untuk, berguna untuk	Dengan cara

2.2.3.2 Non-Factoid Question Generator

The clause that is detected as a non-factoid question candidate based on the keyword in the non-factoid detection process will make the system generate questions using a non-factoid template. Non-factoid question templates can be seen in Table 5 below.

Table 5. Template *Non-Factoid* Question Generator[4]

Category	Template
Definisi	Apa yang dimaksud …?
Alasan	Mengapa ..?
Cara	Bagaimana cara ..?

2.2.3.3 Factoid Question Generator

The process of generating factoid questions will be carried out if the clause is not detected as a non-factoid question. The factoid question generator process is carried out based on customized templates based on NE tag results and syntactic analysis. The factoid question template can be seen in Table 6 below.

Table 6. Template Factoid Question Generator[4]

Kata tanya	Template Question
Apa	**The function asked :** Subjek
	NE Tag : OTHER **Template :** Apa yang + ….. ?
	The function asked : Objek
	NE Tag : OTHER **Template :** Apa yang + …. ?
Siapa	**The function asked :** Subjek
	NE Tag : PERSON, ORGANIZATION **Template :** Siapa yang + ….?
	The function asked : Objek
	NE Tag : PERSON, ORGANIZATION **Template :**

	Siapa yang + …… ?
Mana	**The function asked :** Place information
	NE Tag : LOCATION **Template :** (di / ke / dari) + mana + ….. ?
Kapan	**The function asked :** Time information
	NE Tag : TIME **Template :** Kapan + …… ?

3. CONCLUSIONS

Detection of connectivity between sentences in the automatic question-generating system for Indonesian-language texts has successfully detected which sentences can be raised into questions. The question template developed, to handle many sentences can also be applied, so the system does not only generate perverse questions. After testing using 50 input texts, this system produced 413 questions with the accuracy of all questions at 90.80%.

Suggestions for subsequent research to improve system accuracy are as follows. First, can do more in-depth research on the process of identifying entities named (NER) and types of words (POS Tags) because the process in this system depends on supporting applications. NER errors will have an impact on determining the inappropriate question template and POS Tag errors will have an impact on the identification process so that the grammatically formed becomes wrong. Second, the next research can add semantic analysis between sentences to detect the interrelationship between sentences. Because the detection is done is only looking for the same grammatical in each sentence. This needs to be done because there are the same grammatical errors but have different meanings, so the system will raise the wrong questions. Third, further research can be added to the coreference process before the questions are raised. The core is needed when there are two or more words that lead to the same reference, the system can consider which words or phrases will be entered into questions and answers. Fourth, additional rules are needed to process compound and mixed sentence sentences. Solid compound sentences are actually equivalent compound sentences whose clauses have in common the subject or predicate so that the writing in the sentence seems to consist of one clause whereas there is more than one clause. The more in-depth analysis is needed to handle compound and mixed sentence because different handling is needed with equivalent compound sentences.

REFERENCES

[1] M. Heilman, "Automatic Factual Question Generation from Text," Carnegie Mellon University, 2011.

[2] K. Beulen and H. Ney, "Automatic Question generation for Decision Tree Based State Tying," University of Technology, 1998.

[3] M. R. Iqbal. 2016. "Pembangkit Pertanyaan Otomatis Untuk Teks Berbahasa Indonesia Berdasarkan Template Sintaksis". Skripsi. UNIKOM.

[4] D. D. Ginanjar. 2017 "Pembangkit Pertanyaan otomatis untuk Teks Berbahasa Indonesia yang Mengandung Kalimat Majemuk". Skripsi. UNIKOM.

[5] F. Ferdian, "Implementation of Semantic Analyzer in Indonesian Text-Understanding Evaluation System," pp. 6–10, 2012.

[6] F. Ferdian, "Implementation of Semantic Analyzer in Indonesian Text-Understanding Evaluation System," Bandung Institute of Technology, pp. 6–10, 2012.

[7] M. Fachri, "Named Entity Recognition For Indonesian Text Using Hidden Markov Model," Universitas Gadjah Mada, 2014.

[8] Y. Syaifudin, "Quotations Identification From Indonesian Text Using Hidden Markov Model," Universitas Gadjah Mada, 2014.

[9] W. Hidayat. 2017 "Pendeteksian Kalimat Non-Definisi Pada Pembangkit Pertanyaan Otomatis Untuk Teks Berbahasa Indonesia". Skripsi. UNIKOM.

[10] R. Susanto, "Perbandingan Model Waterfall dan Prototypiing untuk pengembangan sistem informasi," Universitas Komputer indonesia, 2016.