DETECTION OF NON-DEFINITION SENTENCES IN AUTOMATIC QUESTION GENERATOR FOR TEXT IN INDONESIAN LANGUAGE

Wahyu Hidayat[1], Ken Kinanti Purnamasari[2]

[1,2] Informatics Engineering, Indonesia Computer University
Jalan Dipatiukur No. 112-116, Coblong, Lebakgede, Bandung, Kota Bandung, Jawa Barat 40132, Indonesia
E-mail : wahjoe16@gmail.com[1], kinantiken@gmail.com[2]

ABSTRACT

This research is the development of previous research on constructing automatic question generator for Indonesian text containing compound sentences. Question generator research on compound sentences is done to generate non-factoid questions that are questions about definitions, reasons, and method and to generate factoid questions that are questions about what, who, where, how, and when. However, in the study still found a question that is not in accordance with the context of the sentence used as input. This happens to a sentence with a non-factoid keyword like ‘is’ always detected as a non-factoid question with the definition category but not all sentences containing the word ‘is’ are sentence definitions.

The overall process that will be done by the system consists of two stages: preprocessing that is adopted from previous research and main process. Preprocessing is the initial stage of a system for obtaining a type of word, named entity, phrase type, and grammatical functions for every word in the sentence, while the main process is a non-definitional sentence detection process that uses rule-based methods and question generating processes with template methods.

Using 48 text inputs taken from the Indonesian language textbook of the Indonesian Elementary School of Kemendikbud RI, the accuracy of the question was obtained by 68.84% with total questions raised as many as 398 questions.

Keywords : question generation, sentence definition, non-definition sentence detection, compound sentence, rule-based

1. INTRODUCTION

Research on question generating systems has been done before with syntactic methods and statistical methods [1] [2] [3]. For Indonesian the research on question generator is done by template method [4] [5] and semantic method [6]. The question generator in Indonesian is done to generate non-factoid questions that are questions about definitions, reasons, and ways [4] [5] as well as to generate factoid questions which are questions about what, who, where / where, how, and when [7]. Generating questions is helpful to assist the evaluation process in teaching and learning activities for both learners and teachers. Question generators can also be used to build help systems such as Frequently Ask And Question (FAQ) [1] [5].

In a previous study, questionnaire research was conducted for simple sentences [7] [8] and compound sentences [8]. The test results from the research have resulted in accuracy up to 64.52%. The test used input data of 35 input texts which produced 327 questions. The results of the 327 questions 211 questions of which become accurate questions and 116 questions that are not accurate. Among the questions that are not accurate in the research is still found questions that are not in accordance with the context of the sentence that is used as input system. This happens in sentences containing non-factoid question keywords with definition categories like 'is' and 'is' always detected as non-factoid candidate candidates by definition category, but not all sentences containing non-factoid question keywords with categories Definition is the definition sentence.

To solve the problem of not being able to distinguish definitional and non-definitional sentences is required a process of detecting non-definition sentences based on the meaning contained in the clause before and after the non-factoid question keywords by definition category. The process of detecting non-definition sentences is done by searching words that contain the meaning of the person (person), location, quantity, and tool before and after the non-factoid question keywords by definition category.

2. METHOD

To overcome the problem of generating inaccurate questions due to the question that should be a factoid question into a non-factoid question, in this research will be a non-definitional sentence detection process after the process of determining

grammatical functions and before generating question process. The process of detecting non-definition sentences is done by searching from the first word to the last word of the word that forms the meaning of the offender, the quantity, and the location before and after the words 'is' and 'is'.

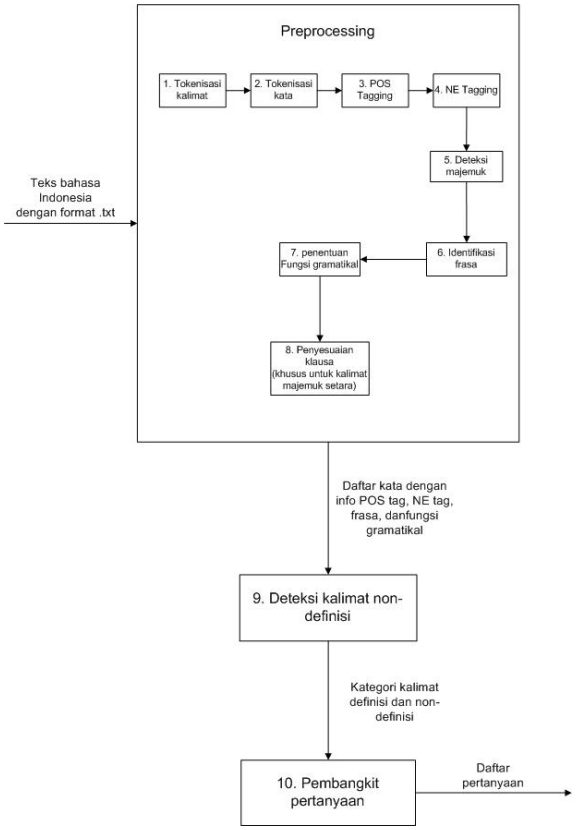


Figure 1. System Flow Chart Block

On the system to be built, the user uploads a text containing a sentence or set of standard sentences in the Indonesian language, then this system will process input text that has the following flow; 1) system to process the tokenisasi sentence, 2) system to process word tokenisasi, 3) system to detect the type of word (POS Tagging), 4) system performs entity recognition process (NER), 5) system conducts detection process compound sentence based on existence (6) the system performs the process of identifying phrases based on phrase patterns that have been defined, 7) the system performs the process of determining grammatical functions to determine the subject, predicate, object, and description based on grammatical function filler rules, 8) the system performs a process of adjustment A special clause for equivalent compound sentences, 9) the system performs non-definitional sentence detection, 10) the system does the question generating process.

2.1 Preprocessing

Preprocessing is the initial process performed by the system after obtaining a text file with .txt format. At this stage the input text will be tokenisasi sentence, tokenisasi kata, labeling of word type (POS tagging), introduction of named entity (NER), the detection of compound sentences, phrase identification, grammatical function determination, and clause adjustment for equivalent compound sentences. This process was done in previous studies [8].

2.1.1 Sentence Tokenisasi

Text uploaded into the system in the form of paragraphs taken from the Indonesian language textbook. The paragraph is certainly composed of a collection of sentences so it takes tokenisasi sentence to separate the sentence. This is necessary because of the questions the system will generate on a per-sentence basis. The phrase callenization is done by recognizing punctuation (.), Exclamation point (!), Or question mark (?).

2.1.2 Word Tokenisasi

After the sentence is separated in the process of tokenisasi sentence, the next process is the tokenisasi word that is done to separate the words in every sentence into a unit of words. The word's name is cutting the input string based on each word that compiles it. Solving sentences into single words is done by scanning sentences and separating word for word with space as a delimiter.

2.1.3 POS *Tagging*

The next process that will be done by the system is the detection of word type (POS tagging). Part of Speech Tagging (POS-Tagging) is a process that automatically labeled word classes in a word in sentence [9]. The result of this POS is very influential on the output of the Parsing process [10]. POS tagger used in this research is POS tagger with HMM method [11].

2.1.4 NE *Tagging*

The next process the system will perform is the introduction of a named entity (NER). Named Entity Recognition (NER) is a process used to recognize and classify terms and phrases as the name entities of a free text [12]. NER is one of the main components of information extration that aims to detect and classify named-entities in a text. NER is commonly used to detect people's names, place and organization names from a document, but can also be extended to the identification of genes, proteins and others as needed. This study uses Indonesian NER application, and uses the HMM method [13] [14].

2.1.5 Detection of Compound Sentences

The next process that will be done system is the detection of compound sentences to all sentences

that are in the input text. Sentences processed will be grouped into three, ie compound sentences equivalent, multi-layered sentences and other sentences. The process of clashing a sentence is based on identifying the existence of a conjunctive word in the sentence.

2.1.6 Phrases Identify

The next process that the system will perform is phrase identification. The process of identifying phrases refers to the pattern of phrases that have been described and performed on previous studies [8]. The identification of a phrase in a clause is done by matching a combination of word types within a clause with a defined phrase pattern. The phrase detection is done sequentially from the first word to the last word in the clause.

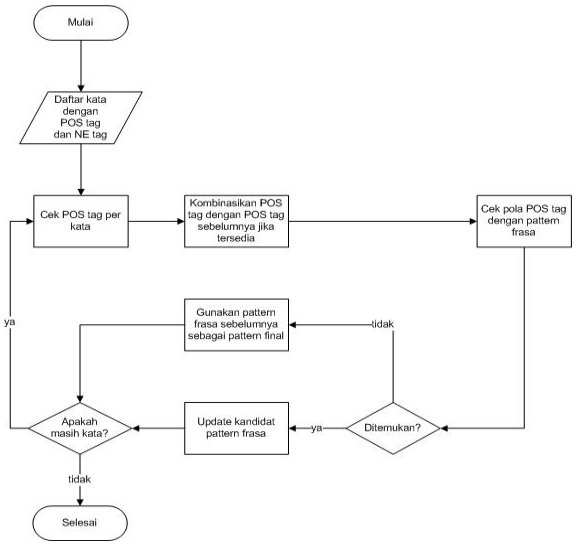


Figure 2. Flowchart Identify Phrases

An example of the result of the phrase identification for clause "*Ciri utama dari sistem tanam paksa adalah mewajibkan rakyat di Jawa untuk membayar pajak dalam bentuk barang dengan hasil-hasil pertanian yang mereka tanam"*

Final Result:

1. **Ciri utama (FN)**
2. **dari sistem tanam (FIN)**
3. **paksa adalah (FVB)**
4. mewajibkan (VB)
5. rakyat (NN)
6. **di jawa (FIN)**
7. **untuk membayar (FIN)**
8. pajak (NN)
9. **dalam bentuk barang (FIN)**
10. **dengan hasil-hasil pertanian (FIN)**
11. yang (SC)
12. mereka (PR)
13. tanam (VB)

2.1.7 Grammatical Function Determination

The next process that will be done system is the process of determining grammatical functions. The process of determining grammatical functions is done based on the grammatical rules of the formation of sentence patterns that have been done in previous studies [8]. Determination of special grammatical functions in multilevel compounds is performed only on parent clauses only whereas in equivalent compound sentences, the determination of grammatical functions will be performed on both clauses that have been separated at the stage of the compound detection process. The process of determining grammatical functions is almost identical to the phrase identification process, ie checking every word or phrase contained in the clause from beginning to end in sequence.

Table 1. The Rules of Grammatical Function Determination

No	Rule
1	If the first word in the clause is a conjunction, make it as a description clause until the comma (,).
2	If no subject has been present, and the type of word or type of phrase now includes the subject-forming type, make that word or phrase as the subject.
3	If the previous word or phrase is the subject, and the type of word or type of phrase now includes the predicate-type form, make that word or phrase as a predicate.
4	If the previous word or phrase is a transitive verb type predicate (VBT) and the type of word or type of phrase now includes the object-forming type, make that word or phrase as the object.
5	If the previous word or phrase is a type of intransitive verb (VBI), make the word or phrase now complementary.
6	If the subject function, predicate, object already exists. And If the word or phrase is now a conjunction, make that word or phrase as a description clause to the end of the clause.
7	If the subject function, predicate, object already exists. And If a word or phrase now contains a preposition word, make that word or phrase as a description until the end of the clause.
8	If the subject function, predicate, object already exists. And If the present word or phrase does NOT contain a preposition, make that word or phrase as a complement to the end of the clause.

Table 2 below is an example of grammatical determination in the case of a sentence *Sistem tanam*

paksa adalah kebijakan Gubernur Jendral Van den Bosh yang mewajibkan para petani Jawa untuk menanam tanaman-tanaman yang dapat diekspor ke pasaran dunia.

Table 2. Example of Grammatical Function Determination

No	Proses
1	Frasa / jenis kata pertama *Sistem* merupakan jenis kata NN. Kondisi ini memenuhi aturan ke-2. Maka jenis kata *Sistem* ditentukan sebagai subjek (S)
2	Frasa / jenis kata berikutnya *tanam paksa* merupakan FVB dengan jenis verba transitif (VBT). Kondisi ini memenuhi aturan ke-3. Maka frasa *tanam paksa* ditentukan sebagai predikat (P)
3	Kata atau frasa berikutnya *adalah kebijakan Gubernur Jendral Van den Bosh yang mewajibkan para petani Jawa untuk menanam tanaman-tanaman yang dapat diekspor ke pasaran dunia* merupakan klausa keterangan. Kondisi ini memenuhi aturan ke-6. Maka klausa tersebut ditentukan sebagai keterangan (K)

2.1.8 Clause Adjustment

The next process the system will perform is the clause adjustment process. The clause adjustment process is only done for equivalent compound sentences. Its function is to make adjustments to the second clause with the first clause. Compound equivalent sentences can be broken down into free clauses. Then, these free clauses will be processed on the question generating module one by one in turns to allow for the generation of questions and answers that are ambiguous.

Table 3. The Rule Of Clause Adjustment

No	Rule
1	If the subject of the second clause is a pronoun (PR), then the pronoun refers to the subject in the first clause.
2	If the second clause does not have a subject, then the subject of the second clause equals the subject of the first clause.
3	If the object in the second clause is a proprietary word (NNG), then the object is the same as the object in the first clause.

2.2 Detection of Non-Definitional Sentences

The process of detecting non-definition sentences is done by means of a word search that can determine that the sentence is a non-definition sentence. The word can be a single word or a combination of two or more words that form a certain meaning. The meaning can form meaning about location, time, actor, etc.

The word search process is done sequentially from the first word to the last word. The determination of a word that is the determinant of a non-definitional sentence can be from words before the word "adalah" or from words after the word "adalah". The complete non-definition sentence detection rule can be seen in table 4.

Table 4. Non-Definition Sentence Detection Rules

No	Rule
1	If there is a combination of words between nouns and their noun companions (several, all, all, numerous, many) before the word 'is', 'is' and constitutes the meaning of quantity or number, then the sentence is a non-definitional sentence.
2	If there is a combination of words between nouns and their noun companions (this, that) before the word 'is', 'is' and forms the nominator's meaning, then the sentence is a non-definitional sentence.
3	If there is a combination of words between nouns and their noun companions (in, on, from, to, to) before the word 'is', 'is' and forms the location meaning, then the sentence is a non-definition sentence.
4	If there is a NER in the PERSON category before the word 'is', 'is' then the sentence is a non-definition sentence.
5	If there is a word combination between nouns and its noun companions before the word 'is', 'is' and is an extension of the meaning of the noun, then the sentence is a non-definitional sentence.
6	If the first word is a description equipped with a comma punctuation (,) and before the connecting word 'is', 'is', then the sentence is a non-definition sentence.
7	If there is a word combination between the verb and its verb companion (not yet, will, should, may, must, must, not, not) before the word 'is', 'is' and form the meaning of various modalities, then the sentence is a sentence Non-definitions.

Under the rule of non-definition sentence detection in table 4, the following figure 3 will explain the flow of the non-definitional sentence definition rule.

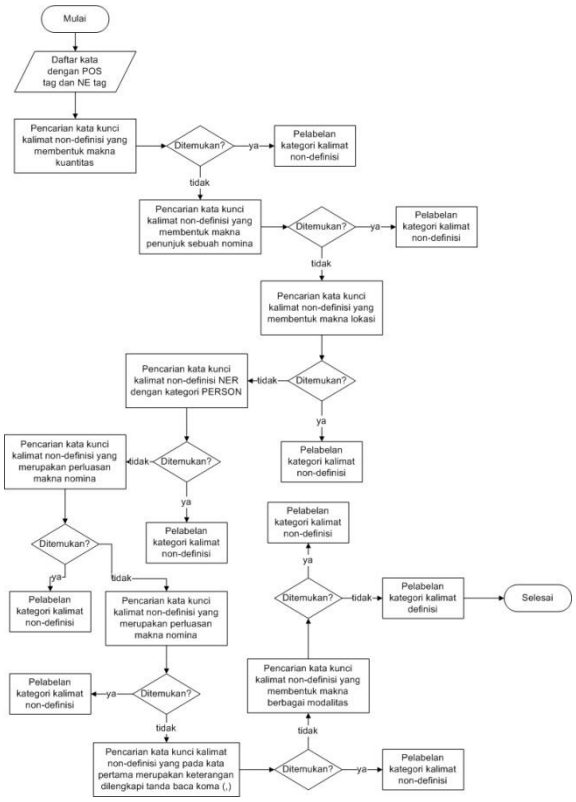


Figure 3. Flowchart Detection of Non-Definitional Sentences

Here are the results of non-definition sentence detection.

Table 5. Example of Non-Definition Sentence Detection Process

No	Process
1	*Sistem tanam paksa* ***adalah*** *kebijakan Gubernur Jendral Van den Bosh yang mewajibkan para petani Jawa untuk menanam tanaman-tanaman yang dapat diekspor ke pasaran dunia.* The words before the word "adalah" are Sistem, tanam and paksa. This condition does not meet the rules in table 7, then this sentence is a definition sentence.
2	*Ciri utama dari sistem tanam paksa* ***adalah*** *mewajibkan rakyat di Jawa untuk membayar pajak dalam bentuk barang dengan hasil-hasil pertanian yang mereka tanam.* Words before the word "adalah" are the ciri, utama, dari, sistem, tanam, and paksa. This condition meets the 1st rule in table 7 that is the word from which is the noun companion accompanying the word system. Merging these two words will shape the meaning of the location. So this sentence is a non-definition sentence.

2.3 Question Generator

After the preprocessing step and the non-definitional sentence detection is done, the sentence that the system input will have various important information that will be used in the question generation process. The questions raised are factoid and non-factoid questions.

Question generators are assisted with templates based on the existence of keywords and grammatical functions derived from syntactic analysis results. Generation of non-factoid questions will be done by identifying the keywords contained in the input sentence. Once a keyword is found, the system will select a template of questions that match the keyword category then generate questions and answers using the selected template.

2.3.1 Non-factoid Detection

The process of labeling non-factoid questions is done by detecting the presence of non-factoid specific keywords. Table 6 below is a list of keywords for non-factoid questions [15].

Table 6. Non-Factoid Keywords

Non-factoid Category	Keywords	
	Before Target	**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, walau demikian, namun demikian	sebab, karena, bertujuan
Methode	berfungsi untuk, berguna untuk	dengan cara

2.3.2 Non-Factoid Question Generator

Clauses or phrases detected as non-factoid through a non-factoid detection process will make the system generate questions using non-factoid templates. Non-factoid templates can be seen in Table 7 below.

Table 7. Template Non-Factoid

Jenis	Template
Definition	Apa yang dimaksud...?
Reason	Mengapa ...?
Method	Bagaimana cara. ..?

2.3.3 Factoid Question Generator

For generating factoid questions will be done using template help and syntactic analysis that produces grammatical functions.

Table 8. Factoid Template

Target	Template
Subject	Siapa yang ...? Apa yang ....? Berapa ...?
Object	Siapa yang ...? Apa yang ....? Berapa ...?
Information	Di mana ...? Ke mana ...? Kapan ...?

3. CONCLUSIONS

The development of automatic question generators for Indonesian subtitles has generated factorial and non-factoid questions. The development of this system can generate questions with an accuracy of 68.84%. From these results it can be concluded that the process of detection of non-definition sentences can overcome errors caused by the generating process of celebration. However, this system still has a high error rate that reduces the accuracy of the question generated system.

Suggestions for subsequent research to improve the accuracy of the system is First, more in-depth research is required for the introduction of entity named (NER), because if there is a recognition error named entity it will impact on the process of detection of non-definition sentence because no detection NER with category PERSON, LOCATION , etc. Such naming entity errors often occur on NER labeling with PERSON and non-PERSON categories causing generating error questions that should result in 'who' to 'what' questions or vice versa.

Second, to overcome the question-answer pairs are not logical and error context, in the next research can be added semantic analysis process between sentences, so that if there is a related meaning between one sentence with the other sentence system can generate more accurate questions. And if the questionnaire template is added to the question generation process, semantic analysis can be used to prevent the occurrence of questions that have the same context as the other questions.

REFERENCES

[1] H. D. A. D. Ali, "Thesis: Automatic question generation : a syntactical approach to the sentence-to-question generation case," 2012.

[2] M. Heilman, "Automatic Factual Question Generation from Text," *Dissertation*, no. June, p. 203, 2011.

[3] K. Beulen and H. Ney, "Automatic question generation for decision tree based state tying," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2, no. November, pp. 805–808, 1998.

[4] M. Fachrurrozi, "Sistem Pembangkit Pertanyaan Otomatis Dengan Metode Template-Based," 2013.

[5] W. Suwarningsih, I. Supriana, and A. Purwarianti, "Indonesian Medical Sentence Transformation for Question Generation," pp. 68–71, 2015.

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

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

[8] D. D. Ginanjar *et al.*, "Pembangkit Pertanyaan Otomatis Untuk Teks Berbahasa Indonesia Yang Mengandung Kalimat Majemuk."

[9] D. Jurafsky and J. H. Martin, "Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition," *Speech Lang. Process. An Introd. to Nat. Lang. Process. Comput. Linguist. Speech Recognit.*, vol. 21, pp. 0–934, 2009.

[10] R. A. SUKAMTO, "Penguraian Bahasa Indonesia Dengan Rosa Ariani Sukamto Nim :," vol. 23507024, 2009.

[11] A. F. Wicaksono and A. Purwarianti, "HMM Based Part-of-Speech Tagger for Bahasa Indonesia," *4th Int. MALINDO (Malay Indones. Lang. Work.*, no. October 2016, 2010.

[12] I. Budi, S. Bressan, G. Wahyudi, Z. A. Hasibuan, and B. A. A. Nazief, "Named Entity Recognition for the Indonesian language: Combining contextual, morphological and part-of-speech features into a knowledge engineering approach," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 3735 LNAI, pp. 57–69, 2005.

[13] M. Fachri, "Named Entity Recognition For Indonesian Text Using Hidden Markov

Model," Universitas Gadjah Mada, 2014.
[14] Y. Syaifudin, "Quotations Identification From Indonesian Online News Using Rule-Based Method," Universitas Gadjah Mada, 2016.
[15] a a Zulen and a Purwarianti, "Study and implementation of monolingual approach on indonesian question answering for factoid and non-factoid question," *PACLIC 25 - Proc. 25th Pacific Asia Conf. Lang. Inf. Comput.*, pp. 622–631, 2011.