AUTOMATIC QUESTION GENERATION FOR TEXT IN INDONESIAN LANGUAGE WITH COMPOUND SENTENCE

Dendy Dwi Ginanjar[1], Ken Kinanti Purnamasari[2]

[1,2] Informatics Engineeringa, Universitas Komputer Indonesia
Jalan Dipatiukur No. 112-116, Coblong, Lebakgede, Bandung, Kota Bandung, Jawa Barat 40132, Indonesia
E-mail : 2003dendy@gmail.com[1], Guinev_gs@yahoo.co.nz[2]

ABSTRAK

Focus of this research is to build the automatic question generator for text in Indonesian language containing compound sentence. Generating research questions using syntactic analysis has been done but only for a simple sentence with a verb clause alone. Whereas in the Indonesian language grammar, the clause is not only shaped by the predicate containing verb alone. Besides the issue of clause above, previous studies did not do phrases identification. Whereas grammatical function in the clause is not occupied by only a single word alone but can be occupied also by a variety of phrases.

The whole process will be done by the system consists of two stages, the preprocess and the main process. Preprocess is the initial stage of the system to get POS tag and NE tag for each word in the sentence, while the main process is a process of syntactic analysis consists of the identification phrases with the pattern matching method, determining grammatical functions with rule-based method and questions generation with template method.

By using 35 text input taken from Buku Sekolah Elektronik Kemendikbud RI, obtained the accuracy amounted 71.06% with details of errors caused by preprocess amounted to 44.55%, 21.78% syntactic analysis and 33.66 % questions generation.

Kata kunci : question generation, identification phrases, compound sentence, rule-based

1. INTRODUCTION

In a study of text-to-question generation, generation systems in question will be given an input of a single sentence, compound sentence, text, text collections even a transcript of a conversation or dialogue, with the aim to produce a set of questions automatically which the answer is contained in the text input [1]. Generating questions useful to assist the evaluation process in teaching and learning activities both for students and for teachers [2].

Benefits of questions generation that can be taken by students is as an evaluation tool, while for teachers, generating questions can be used as tools to prepare for potential questions that might be asked by learners. In addition in the teaching-learning process, questions generation can also be used to build a help system such as Frequently Asked Question (FAQ).

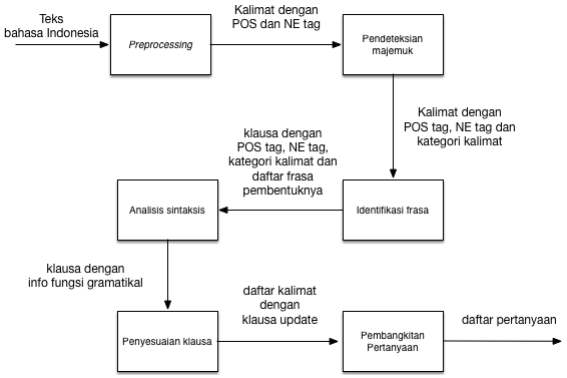
Several studies on question generatin has been carried out. Questions generation in English was conducted using syntactic and statistical methods, while the Indonesian research on Questions generation conducted using templates and semantic methods. The Indonesian questions generation have been made to produce non-factoid and factoid questions [3] [4]. Non-factoid question is a question of 'definition', 'reason', and 'how', while factoid questions is the question of 'what', 'who', 'where', 'where', 'how' and 'when '. However, the factoid questions research has been carried out for simple sentences that are limited to verb clauses. Verb clause is a clause containing predicate verb. In order Indonesian predicate function not only be occupied by the verb, but can be occupied by a phrase or word types such as nouns, adjectives, numbers and others [6] [8].

Questions were raised will need answers about the subject, object information even about the title. Therefore, the determination of grammatical function in a sentence becomes very important. Grammatical function in the sentence can be occupied by a single word or phrase so that the process of identifying the phrase be a major determinant of success determining the grammatical function of each word or phrase that exists.

2. METHODS

To overcome the problem of the generation of the necessary inquiries with compound sentence syntactic analysis from identification phrase to identify grammatical functions. Identify phrases are necessary for the determination of grammatical functions, while the grammatical function is indispensable in the process of generating questions

for factoid questions need answers about the subject, the object, even a description.



Gambar 1 Alur Sistem

In general, this system will have a groove as follows; 1) user to upload a text file containing a sentence or group of sentences raw in Indonesian, 2) the system will perform preproccess which will generate information POS tags and NE tag of each word in the text input, 3) the system will make the detection of complex sentences based on the presence of the word keys that have been defined, 4) system is the process of identifying phrases, 5) system to identify grammatical function to determine the subject, predicate, and object in the sentence, 6) the system performs a process of adjustment clauses for complex sentences similar, and finally, 7) the process of generating questions.

2.1 *Preprocess*

Preprocessing the initial stages performed by the system, which at this stage, the text input will ditokenisasi and labeling do the kind of words (POS tag) and named entity recognition (NER). This process is done entirely by the application of Indonesia NER [9] [13].

The introduction of an entity named or referred to under Named Entity Recognition (NER) is a process used to identify and classify terms and phrases as the entity name of a free text [12]. NER is one of the main components of information extration which aims to detect and classify the named-entity in a text. NER is generally used to detect a person's name, place name and the organization of a document, but can also be extended to the identification of genes, proteins and others as needed.

2.2 Deteksi Majemuk

After the process of labeling the word is completed, the system will make the detection of a compound against all sentences into the input system. There are three categories of sentences that have defined the system, namely; similar compound, compound-rise, and others. Compound detection is based on the existence of conjunctions. Conjunctive meant is conjunctive and compound complex sentences similar tiles.

Table 1 Example of Deteksi Majemuk

Kalimat	Presiden tidak bisa menghadiri peresmian tol barito karena Presiden baru kembali besok.
Kategori	Majemuk bertingkat
Kata kunci	karena

2.2 Identifikasi Frasa

Formation of the phrase refers to a pattern of terms that have been defined by the system. Search phrase patterns performed sequentially from the first word to the last word in the clause using rule-based. The rules for determining phrases in the clause can be seen in Table 2 below.

Table 2 Rule of Phrase Identification

NO	Aturan
1	Pointer is at n-th word, then the system stores the nth word to the variable x and store POS tag of said n-th to the variable y. Further n plus one. If the word nth still there then move the pointer to the nth, if not, copy the variables x and add to variable z ['phrase'] while the variable y is copied and added to the variable z ['category'].
2	Pointer is at n-th word, then the system will combine POS tag nth word with POS tags that exist in the variable y, the combination is stored to the variable k. Then the system will look for patterns of phrases that correspond to the combination of the variable k. If not found patterns similar phrases, the system will continue the process to the point (3). However, if found patterns similar phrases, change the value of the variable y to the value of the variable k and add words to the nth to the variable x. N value plus one and the system will repeat the process point (2). Variable k always emptied at each iteration.
3	The wording in the variable x is determined as the phrase then the variable x is copied and added to the variable z ['phrase'] while the phrase category that corresponds to the pattern of the phrase in the variable y stored to the variable z ['category']. The variables x and y is emptied and the system re-do the process in point (1).

Example : klausa *Presiden Joko Widodo sedang melaksanakan tugas Negara ke Singapura*

Step 1 (aturan 1) :

Pointer is the first word (n = 1), the President of the POS tag NN. The word 'President' is stored to the variable x and Tag 'NN' stored to the variable y. Then

the n plus one (n = 2) and the pointer moves to the nth.
var x = Presiden
var y = NN
var k = ‘’
var z = ‘’

Step 2 (aturan 2) :
The pointer is on the second word (n = 2), ie Joko with POS tag NN. The combination of POS POS tag and said second tag in the variable y stored to the variable k and the system looks for patterns of phrases that correspond to the combinations contained in the variable k. Appropriate pattern so that the word 'Joko' is added to the variable x and the value of the variable y is replaced by the value of the variable k.
Var n ditambah satu (*n=3*) lalu pointer bergerak ke-*n*.

update variabel
var x = ‘Presiden Joko’
var y = ‘NN+NN’
var k = ‘’
var z = ‘’

Step 3 (aturan 2) :
ointer located on the third word that is Widodo with the POS tag NN. The combination of POS tag third word with POS tags in the variable y stored to the variable k and the system looks for patterns of phrases that correspond to the combinations contained in the variable k. Pattern matches found so that the word 'Widodo' is added to the variable x and the value of the variable y is replaced by the value of the variable k.

Var n add by 1 (n = 4), and the pointer moves to the nth.

update variabel
var x = ‘Presiden Joko Widodo’
var y = ‘NN+NN+NN’
var k = ‘’
var z = ‘’

Step 4 (aturan 2,3) :
Pointer located on the fourth word that is moderately POS RB tag. The combination of POS tag third word with POS tags in the variable y stored to the variable k and the system looks for patterns of phrases that correspond to the combinations contained in the variable k. Patterns were not found.

var x = ‘Presiden Joko Widodo’
var y = ‘NN+NN+NN’
var k = ‘NN+NN+NN+RB’
var z = ‘’

Step 5 (aturan 3) :
Pattern phrase not found in Table 2.12 so that a list of words that exist in the variable x is set as a phrase by phrase categories in accordance with the pattern of the variable y. Copy the variable x to variable z [phrase] and save the category phrase to variable z [category] then clear the variables x and y. Then, repeat the process rule points system (1).

update variabel
var x = ’’
var y = ’’
var k = ‘’
var z[frasa] = ‘Presiden Joko Widodo’
var z[kategori] = ‘FN’

Step 6 :
The search process is repeated phrase of the first steps for the next words are still found in the clause.

Final results of the phrase identification for clause *Presiden Joko Widodo sedang melaksanakan tugas Negara ke Singapura* adalah sebagai berikut :
Phrase = 1. *Presiden Joko Widodo (FN)*
2. sedang melaksanakan (FVB)
3. tugas Negara (FN)
4. ke Singapura (FIN)

2.3 Gramatikal Function

The sentence is a unit of language compiled basic constituents are generally in the form of clauses, conjunctions (if any) and the final intonation. In the final written language intonation is expressed by a period (.), The question mark (?) Or exclamation (!) [12]. Structurally, a simple sentence in Indonesian behaves as follows [14] [12].

Tabel 3 Sentence Patern

STRUKTUR
Subjek + Predikat (S-P)
Subjek + Predikat + Objek (S-P-O)
Subjek + Predikat + Pelengkap (S-P-Pel)
Subjek + Predikat + Objek + Pelengkap (S-P-O-Pel)
Subjek + Predikat + Objek + Keterangan (S-P-O-K)

STRUKTUR
Subjek + Predikat + Keterangan (S-P-K)

Identify the grammatical function is done after the process of identifying phrase. This process is performed using the rules of grammar sentence structure as it has been presented in Table 3.2. Determination grammatical functions specifically in multi-storey complex sentences just carried on alone while the parent clauses in complex sentences similar, determining the grammatical function will be performed on two clauses that have been separated at the stage of compound detection process.

The process of determining grammatical functions almost the same as the phrase identification process, which is done checking each word or phrase contained in the clause from beginning to end in sequence. The rules in determining the grammatical function can be seen in Table 3.4 below.

Table 4 Gramatical Function Rule

No	Aturan
1	If the first word in the clause is a conjunction, make those words as a description clause until the comma (,).
2	If there is no subject, and type the word or phrase types now include the type forming the subject, make words or phrases such as subjects.
3.	If the previous word or phrase is the subject, and type the word or phrase types now include the type forming the predicate, make words or phrases such as predicate.
4	If the word or phrase before is a transitive verb predicate manifold (VBT) and type the word or phrase types now includes object-forming type, make words or phrases such as objects.
5	If the word or phrase before an intransitive verb is the predicate manifold (VBI), make word or phrase now as a complement.
6	If the function of the subject, predicate, object already exists. If a word or phrase and now is a conjunction, make words or phrases such as captions clause until the end of the clause.
7	If the function of the subject, predicate, object already exists. If a word or phrase and now contains a preposition, make words or phrases such as captions to the end of the clause.
8	If the function of the subject, predicate, object already exists. And if the word or phrase now NOT contain a preposition, make words or phrases such as complementary to the end of the clause.

2.3 Penyesuaian Klausa (*call reference*)

The process of adjustment clause is only done for similar compound sentences. Its function is to make adjustments to the second clause by clause first. Compound sentences can be broken down into similar clauses are free. Then, the clauses of this free will be processed at the plant module questions one by one alternately so as to enable it generates questions and answers are ambiguous.

Rules for adjustment clause can be seen in Table 5 below.

Tabel 5 Rules for adjustment clause

No	Aturan
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.

Example

Kalimat : *Presiden akan berkunjung ke lokasi kemudian Dia akan kembali ke Jakarta* .

Klausa 1 : *Presiden akan berkunjung ke lokasi*

Klausa 2 : *Dia akan kembali ke Jakarta*

Tabel 6 Contoh Penyesuaian Klausa

No	Proses
1	Subject for second clause is *Dia, Dia* is pronoun (PR). This condition is in accordance with the rules of the 1st. So, *Dia* is the same with subject in clause 1, *Presiden*. Clause 2 (before) : Dia akan kembali ke Jakarta Clause 2 (after) : Presiden akan kembali ke Jakarta

2.4 Pembangkit Pertanyaan

After the preprocessing stage and syntactic analysis is done, then the words that feed the system will have a wide range of important information will be used in the process of generating questions. The question that was raised was the question and the non-factoid factoid.

Generation sentence with a question template assisted by the presence of key words and

grammatical functions obtained from the syntactic analysis. Generation of non-factoid questions will be done by identifying the key words contained in the input sentence. Once the keywords are found, the system will choose the template questions that correspond to the keyword categories and evoke questions and answers using templates that have been mentioned. Here is a list of keywords and query templates that can be used to generate non-factoid question of a sentence.

2.4.1 Deteksi *Non-factoid*

In this process will be the candidate labeling non-factoid. Same way with a detection compound, namely by identifying the presence of non-specific keywords factoid. Table 7 The following is a list of keywords for non-factoid questions [5].

Tabel 7 Keyword for Non-factoid

non-factoid	eyword	
	After target	**Before 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
Method	berfungsi untuk, berguna untuk	dengan cara

2.4.2 Pembangkitan pertanyaan *Non-factoid*

Clauses or sentences that are detected as non-factoid detection process non-factoid will make use template system raises the question of non-factoid. Template non-factoid can be seen in Table 8 below.

Tabel 8 Non-Factoid Template

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

2.4.3 Pembangkitan pertanyaan *factoid*

For the generation factoid questions will be done using the help of templates and syntactic analysis.

Tabel 9 Factoid Template

Target	Template
Subjek	Siapa yang ...? Apa yang ....? Berapa ...?
Objek	Siapa yang ...? Apa yang ....? Berapa ...?
Keterangan	Di mana ...? Ke mana ...? Kapan ...?

3. CONCLUSIONS

The system automatically generating questions from the text speak Indonesia successfully raised questions factoid form of questions: who, what, where, when, how and non-factoid questions such as the definition question, how and why. This system can generate questions with an accuracy of 71.06% with details of errors caused by preprocess amounted to 44.55%, 21.78% syntactic analysis and fault generation question of 33.66%. From these results it can be concluded that the syntactic analysis process which generally consists of the identification process and the determination of grammatical phrase still have error rates high enough to reduce the accuracy of questions generated by the system.

Suggestions for subsequent research to improve the accuracy of the system is as follows. First, the necessary additional rules for processing complex sentences density and mix. Density compound sentence is actually a compound sentence that similar clauses-clauses have the same subject or predicate so the writing in the sentence seemed to consist of one clause when composed of two or more clauses that sealed the subject or predicate. Although similar, the handling of different density compound sentence with a similar compound and is in need of more in-depth analysis so that the process of identifying phrases in complex sentences can be more accurate density.

Second, the rules required the addition of the phrase identification process to be able to process sentences containing extension subject, predicate or object extension. This is necessary because of the limitations of the phrase pattern defined by the system is not optimal recognize grammatical functions expanded.

Third, the system has been built on this research, its preprocess stage relies on supporting applications, so the accuracy of the results will preprocess impact on the accuracy of the syntactic analysis is the identification of phrases and grammatical function determination. As a result of this preprocess is kind word (POS Tag) and named entity (NE tags) for each word that a text input. To improve the accuracy of identification of the phrase, in the preprocess can be added detection process ambiguous words.

Fourth, in order to get the question-answer pair relevant semantic analysis process can be added between sentences, so that if there is a meaning related to one sentence to another sentence that the system can generate more accurate question-answer. And if the addition of template questions on the generation process question, semantic analysis can be used to prevent the emergence of the questions that have the same context as 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] W. Suwarningsih, I. Supriana, A. Purwarianti, and P. P. Informatika-lipi, "Tantangan dan Peluang pada Question Generation 1,3,4," *J. Sist. Inf.*, vol. 6, no. 2, pp. 719–730, 2014.

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

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

[5] W. Suwarningsih, I. Supriana, and A. Purwarianti, "Toward a Framework for Indonesian Medical Question Generator," *TELKOMNIKA (Telecommunication Comput. Electron. Control.*, vol. 13, no. 1, p. 357, 2015.

[6] Efendi, D. Kentjono, and B. Suhardi, *Tata Bahasa Dasar*, Pertama. Bandung: ROSDAKARYA, 2015.

[7] 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.

[8] M. Khairah and S. Ridwan, *Sintaksis : Memahami Satuan Kalimat Persfektif Fungsi*, Pertama. Jakarta: PT Bumi Aksara, 2014.

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

[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] Y. Syaifudin, "Quotations Identification From Indonesian Online News Using Rule-Based Method," Universitas Gadjah Mada, 2016.