AUTOMATIC QUESTION GENERATION FOR 5W-1H OPEN DOMAIN OF INDONESIAN QUESTIONS BY USING SYNTACTICAL TEMPLATE-BASED FEATURES FROM ACADEMIC TEXTBOOKS

[1,2]SETIO BASUKI, [3]SELVIA FERDIANA KUSUMA
[1,2]Informatics Engineering, Universitas Muhammadiyah Malang, Indonesia
[3]Informatics Engineering, Politeknik Kediri, Indonesia
E-mail: [1,2]setio_basuki@umm.ac.id, [3]selvia@poltek-kediri.ac.id

ABSTRACT

The measuring of education quality in school can be conducted by delivering the examination to the students. Composing questions in the examination process to measure students' achievement in the school teaching and learning process can be difficult and time consuming. To solve this problem, this research proposes Automatic Question Generation (AQG) method to generate Open Domain Indonesian Question by using syntactical approach. Open Domain questions are questions covering many domains of knowledge. The challenge of generating the questions is how to identify the types of declarative sentences that are potential to be transformed into questions and how to develop the method for generating question automatically. In realizing the method, this research incorporates four stages, namely: the identification of declarative sentence for 8 coarse-class and 19 fine-class sentences, the classification of features for coarse-class sentence and the classification rules for fine-class sentence, the identification of question patterns, and the extraction of sentence's components as well as the rule generation of questions. The coarse-class classification was carried out based on a machine learning with syntactical features of the sentence, namely: Part of Speech (POS) Tag, the presence of punctuation, the availability of specific verbs, sequence of words, etc. The fine-class classification was carried out based on a set of rules. According to the implementation and experiment, the findings show that the accuracy of coarse-class classification reaches 83.26% by using the SMO classifier and the accuracy of proposed fine-class classification reaches 92%. The generated questions are categorized into three types, namely: TRUE, UNDERSTANDABLE, and FALSE. The accuracy of generated TRUE and UNDERSTANDABLE questions reaches 88.66%. Thus, the obtained results show that the proposed method is prospective to implement in the real situation.

Keywords: *Automatic Question Generation (AQG), Coarse-class Classification, Fine-class Classification, Open Domain Question, Syntactical Approach*

1. INTRODUCTION

One of the main components in school teaching and learning process is the availability of questions for the achievement evaluation. The quality and type of questions should get more attention because learning outcomes should be assessed by the capability of students to pass the examination. To actualize this, it is necessary to provide various types of questions. The provision of examination questions becomes an important issue. Composing questions can be difficult and time consuming. Moreover, the challenge faced is how to provide the various types of questions to comprehensively measure students' understanding. Therefore, this research aims to develop a software that is able to generate the questions in Indonesian language automatically from the text documents, school textbooks, and internet articles. The main challenge in this research is how to identify the type of declarative sentence that potential to be transformed into questions. In this case, the variety of declarative sentences that might appears in the data sources becomes the biggest issue. The next challenge is the way how to develop the technique that is able to generate the questions automatically. The proposed technique should be able to accommodate the variation of sentences in the data source and the needs of required evaluation questions in school. In this research, we follow the 5W+1H patterns (what, when, where, who, why and how). These needs could be accommodated by

Natural Language Processing (NLP) technique, especially Automatic Question Generation (AQG).

There are AQG-related researches on both English and Indonesian Language (Bahasa Indonesia). In the English AQG, there are two research categories according to the type of questions and their domain. In regards to the type of questions, there are two types of questions: multiple-choice questions [1][2][3][4] and W-H questions (what, why, when, who, where, or how) [5][6][7][8][9][10][11][12][13][14][15][16]. The number of references related to the W-H questions is higher than multiple-choice question since this research aims to create the Indonesian version of W-H questions. However, multiple choices AQG research is needed to be referred in order to know the position of this research among the existing AQG researches. The first [1] multiple choice AQG research aims to encompass the biological domain by using Knowledge Descriptor, Question Generator, and E-Learning Executer techniques. The research uses pre-defined template for the type of questions such as: The Concept definition, Correct or Wrong options, and Problems. There are 12 chapters with 239 sub topics and 46.345 questions used as the dataset. The multiple-choice AQG researches in [2][3][4] are using semantic-based approach. Nevertheless, there are differences among them. Researches [2][4] are focused on biological domain whereas research [3] is focused on open domain. Research [2] uses GENIA EVENT corpus with three stages of generating questions: (1) the extraction of information to extract semantic relations; (2) automatically-generated questions by using semantic relations; and (3) the distractor of components by using distributional similarity. Research [4] uses GENIA as corpus and British National Corpus (BNC) as a general corpus. This research uses unsupervised Relation Extraction to identify the most important terminology and name entity in the documents and then take its relations. Research [3] uses TREC 2007 dataset. This research uses Semantic Role Labels (SRL) and Named Entities in generating questions.

In W-H AQG research, there are two domain categories, which are open question and specific question. Open question means a question is not intended to sole one subject or purpose as in [6][7][10][12][15]. Research [6] employs techniques consisting of Syntactic Parsing, Part of Speech (POS) tagger, and Named Entity. The data sources used in this research are Wikipedia, Yahoo! Answers, and Open Learn (QGSTEC). Research [7] also uses QGSTEC 2010 as the data sources. This research employs techniques consisting of Sentence Simplification, Name Entity, and Predicate Argument Structure. To measure the importance of generated questions, this research uses Latent Dirichlet Allocation (LDA). Research [10] uses templates and rules consisting of three stages; they are training module, question generator engine that generates the questions, rank, review, and user interface. The research aims to assist the academic evaluation process. Furthermore, the generated questions in [12] are used for educational purposes such as: self-study, quiz, etc. by using a template approach of more than 50. Finally, the research [15] uses datasets from the 2010 QGSTEC and the 2007-TREC. This research uses two techniques: Part of Speech (POS) tagger and Named Entity (NER) in order to generate questions and to get words or phrases needed. Furthermore, the questions are formed based on the rule. The specific domains are the questions set for more specific purposes in a particular field or purpose as in [5][8][9][11][13][14][16]. Research [5] aims to build AQG in the form of Trigger Question for the purpose of assisting academic writing. The research uses 504 citations taken from 45 academic papers as a training data and consists of three stages: Citation Extraction, Citation Classification, and Rule-based Question Generation. Research [8] uses two techniques: learning lexico-syntactic patterns and test item generation that is able to generate 24 question patterns in order to get AQG in biographical texts. Research [9] focuses on AQG by using extraction technique of simplified sentences from appositives, subordinate clauses, and other constructions. The research uses 25 Encyclopedia of Britannica Articles as the data sources that are related to cities in the world. Research [11] builds AQG to generate feedback questions for critical review support. The approaches used in this research are syntax-based and template-based approaches; Further, this also uses Wikipedia as the data sources. Research [13] aims to generate a question pattern for ontology-based question by using textual entailment in Movie and Cinemas domain. The techniques used are predictive questions asked by users in the ontology domain. Research [14] conducts the question generation based on question templates formed from the training process in many medical articles. Finally, research [16] builds AQG based on the semantic pattern to generate various kinds of depth questions for self-study or tutoring. The data used in this research come from science textbooks (www.ck12.org).

In regards to Bahasa Indonesia, there are several current AQG researches as in [17][18][19][20]. Although they are developed for Bahasa Indonesia, those four researches have the differences in their field, techniques, and purposes. The researches [17][18][19] employ various health articles as the data sources. All those researches generate W-H question; however, there are still differences among them. Research [17] uses manually written rules to perform the syntactic transformations (e.g. the identification of keywords or key phrases to Medical Name Entity Recognition based on PICO Frame, namely: Patient/Problem, Intervention, Comparison, and Outcome to change the declarative sentences into questions. These generated questions use question marks such as who, what, and how. Meanwhile, research [18] aims to form not only a question but also a couple of question-answer. The technique used in this research is semantic-based template by using a combination of semantic role labeling (SRL) with PPPICCOODTQ components (Problem, Patient, Intervention, Compare, Control, Outcome, Organs, Drug, Time, Quantity). All the generated questions use question marks, such as: what, how, why, how many, and when. Research [19] uses the formation questions of what, why, and how by using P-A templates based on a statistical measure. As in [17], this research is also carried out based on P-A extraction on the PICO frame. Apart from those three researches, research [20] uses 30 articles as the data sources from various fields, such as: the solar system, evolution, light, electrical energy, etc. The technique used in this research is a syntactic template-based method comprising sentence extraction, sentence classification, and question generation. The kinds of question that are possible to generate are related to Definition, Reason, and Method with what, why, and how.

Related to the main contribution of developing open domain Indonesian AQG, this research differs from researches [17][18][19] that focus on the medical field; yet, it is similar to research [20] that concentrates on an open domain and uses template-based techniques. However, there are several differences from previous researches, such as: (1) research [20] uses 3 types of sentences, namely: Definition, Reason, and Method while this research identifies 8 (coarse-class) templates and is subsequently followed by depth elaboration from 8 into 19 (fine-class) templates; (2) The data sources in this research are varied from different school e-books and internet articles. It also has a variety of declarative sentences. The variation of data sources inevitably triggers the necessity of identifying keywords and syntactic patterns of each sentence. In addition, this research also requires rule-based filtering techniques since the data sources used are not only in form of documents or articles, but also of e-books; (3) The classification of research [20] merely takes one stage with certain word features to represent each class while the classification of this research takes two stages. The first classification stage uses 35 features (syntactic patterns) of each sentence type and the second phase is done based on collection rules. The first classification is performed based on a machine learning; and eventually, (4) this also contributes in building a set of rules to generate questions. One declarative sentence can be transformed into one or more questions, depending on the sentence type and the components contained therein.

2. DECLARATIVE SENTENCES

2.1 The Pattern of Sentences

Selected E-books were converted into plain text by removing the title, subtitle, references, and other contents that do not have a contribution. This process is called filtering. The next process is called segmentation in which the whole plain texts were split into sentences. The pattern identification of sentences was carried out manually to the sentences. The formed question comes from the extraction of several components in a sentence. Several identified sentences are Definition, List, Cause-Effect, Reason, Method, Location, Entity, and Date-Time.

A definition sentence is a type of sentence aiming to explain an entity. There is a variation to explain the entity, particularly if it is seen in syntactical point of view. On this research, four variations of definition sentence are identified to explain an entity.

DefinitionType-1

Sentence Pattern: Entity + definition-keyword + explanation

Example: Lokasi relatif (entity) adalah (definition-keyword) lokasi sesuatu objek yang nilainya ditentukan oleh objek-objek lain di luarnya (explanation) *(Relative location (entity) is (definition keyword) the location of an object in which the value is determined by other objects outside (explanation))*

DefinitionType-2

Sentence Pattern: Entity + comma + definition-keyword + explanation

Example: Litosfer (entity) , (comma) yaitu (definition-keyword) lapisan yang terletak di atas lapisan pengantara, dengan ketebalan 1.200 km (explanation) (*Litosfer (entity) , (comma) is (definition-keyword) a layer located above the mediator layer, with a thickness of 1.200 km (explanation)*)

DefinitionType-3

Sentence Pattern: Entity + (colon/semi-colon) + explanation

Example: Skala peta (entity) : (colon) dapat dihitung jaraknya dari rumah Anda atau kota Anda ke kota tempat kejadian (explanation) (*Map scale (entity) : (colon) the distance can be calculated from your home or your city to the city scene (explanation))*

DefinitionType-4

Sentence Pattern: Explanation + definition-keyword + entity

Example: Ilmu yang mempelajari gempa bumi, gelombang-gelombang seismik serta perambatannya (explanation) disebut (definition-keyword) Seismologi (entity) (*The study of earthquakes, seismic waves and the spreads (explanation) is called (definition-keyword) Seismology (entity))*

Definition-keywords: is (*merupakan, adalah, ialah*), known as (*dikenal, dikenal dengan*), called as (*disebut juga, disebut sebagai*), is called (*disebut*), etc.

List sentence is a sentence that gives details to an entity, either example pronunciation or an explanatory. There is only one type of list sentence pattern in this research.

List

Sentence Pattern: List-statement + comma + list-keyword + entity-1 + entity-2 + ... + entity-n

Example: Berdasarkan kenyataan tersebut terdapat dua jenis lapisan batuan utama (list-statement) , (comma) yaitu (list-keyword) lapisan kedap (impermeable) (entity-1) dan lapisan tak kedap air (permeable) (entity-2) (Based on the fact there are two kinds of main rock (list-statement) , (comma) namely (list-keyword) impermeable layer (entity-1) and permeable layer (entity-2))

List-keywords: Namely, specifically, i.e (yaitu), as follows (sebagai berikut), covers (meliputi) including (diantaranya), etc.

Cause-effect sentence is used to show the causal connection and there are three ways to write this sentence.

CauseEffectType-1

Sentence Pattern: Effect + CauseEffect-1-keyword + Cause

Example: Perbedaan warna air laut (Effect) disebabkan oleh (CauseEffect-1-Keyword) perbedaan kandungan zat larutan atau organisme yang ada di dalam laut tersebut (Cause) (*The difference of sea color (effect) is caused by the difference of solution substances or the organisms living in that sea (cause))*

CauseEffectType-2

Sentence Pattern: Cause + CauseEffect-2-keyword + Effect

Example: Tinggi rendahnya permukaan bumi (relief) (Cause) mempengaruhi (CauseEffect-2-Keyword) pola penyinaran matahari (disebut juga faktor fisiografi) (Effect) (*The high and low of Earth surface (relief) (cause) influence (causeEffect-2-keyword) the sunlight cycle (is also knows as physiographic factors) (effect))*

CauseEffectType-3

Sentence Pattern: "Jika" + Cause + "maka" + Effect

Example: Jika (CauseEffect-3-Keyword) penyebab primernya adalah infeksi (Sebab) maka (CauseEffect-3-Keyword) ditangani dengan pemberian antibiotika (Effect) (*If (causeEffect-3-keyword) the primer cause is an infection (cause) therefore (causeEffect-3-keyword) use antibiotics (effect))*

CauseEffect-1-keyword: Caused by (disebabkan, disebabkan oleh, diakibatkan, or diakibatkan oleh), affected by (dipengarhi or dipengaruhi oleh) etc.

CauseEffect-2-keyword: Influence (mempengaruhi), cause (menyebabkan, mengakibatkan), therefore (sehingga), inflict (menimbulkan), resulting in (berakibat pada), etc.

CauseEffect-3-keyword: If-then (jika-maka, apabila-maka), because-then (karena-maka), because-so that (karena-sehingga), to-then (untuk-maka), because of-then (olehkarena-maka), etc.

Reason sentence aims to present a reason and justification towards an event, process, or activity. There are three types of reason sentences identified in this research.

ReasonType-1

Sentence Pattern: Fact + reason-keyword-1 + reason

Example: Mesir Pertengahan mengalami kemunduran (fact) karena (reason-keyword-1) serangan Hykos yang gemar berperang (reason) *(Middle Egypt had lost ground (fact) because (reason-keyword-1) Hykos's attack that tends to be warlike (reason))*

ReasonType-2

Sentence Pattern: Reason-keyword-2 + reason + (comma or not) + Fact

Example: Karena (reason-keyword-2) Indonesia terletak di daerah sekitar khatulistiwa (reason), Indonesia sering mengalami hujan (fact) (*Because (reason-keyword-2) Indonesia lies along the equator (reason), Indonesia often encounters seasonal rain (fact))*

ReasonType-3

Sentence Pattern: reason-keyword-3 + (comma or not) + Fact

Example: Oleh karena itu (reason-keyword-3), (comma) Homo sapiens dianggap sebagai jenis yang paling sempurna yang menjadi nenek moyang manusia (fact) (*Therefore (reason-keyword-3) , (comma) Homo sapiens is regarded as the perfect species that becomes human ancestors (fact))*

Reason-keywords-1: Karena (Because), oleh karena (therefore), apabila (if), bila (when), terjadi karena (occurs because), terjadi akibat (caused by), agar (in order), akibat (result), semakin (more), dengan tujuan (with the purpose of), bertujuan untuk (aims to), etc

Reason-keywords-2: Karena (Because), oleh karena (therefore), apabila (if), bila (when), terjadi karena (occurs because), terjadi akibat (caused by), agar (in order), akibat (result), semakin (more), dengan tujuan (with the purpose of), bertujuan untuk (aims to), etc

Reason-keywords-3: Oleh karena itu (therefore), karena itu (because), dengan demikian (thus), oleh sebab itu (therefore), hal ini (in this matter), hal ini terjadi (it happens), ini terjadi karena (it happens because), akibatnya (as a result), alasannya (the reason), etc

Method sentence aims to give explanation on how the activity is done. Method sentence can be done in three ways.

MethodType-1

Sentence Pattern: Target-method + passive verb/passive VBT or intransitive verb/VBI + method-keyword-1 + detail-method

Example: Reproduksi aseksual (target-method) dilakukan (passive verb) dengan (method-keyword-1) membentuk tunas eksternal atau tunas internal (detail-method) *(Asexual reproduction (target-method) is produced (passive verb) by (method-keyword-1) forming the external or internal buds (detail-method))*

MethodType-2

Sentence Pattern: detail-method + method-keyword-2 (optional) + active verb + Target-method

Example: Satu bagian bumi didorong masuk ke selubung untuk meleleh kembali, bagian lainnya didorong ke atas (detail-method) membentuk (active verb) pematang (target-method) *(One part of the Earth pushes into the sheath to melt back, the other part pushes upwards(detail-method) forming (active verb) embankment (target-method))*

MethodType-3

Sentence Pattern: Method-keyword-3 + (comma or not) + target-method

Example: Dengan cara ini (method-keyword-3), (comma) dapat memaksimalkan homogenisasi mikroorganisme pada pelarut (target-method) *(This way (method-keyword-3), (comma) could maximize the homogeneity of microorganism toward the solvent (target-method))*

method-keywords-1: dengan (with), secara (by), dengan cara (in such a way), melalui proses (through the process), etc

method-keywords-2: cara (way), proses (process), mekanisme (mechanism), etc

method-keyword-3: Dengan cara ini (this way), dengan cara itu(lah) (that way), melalui proses ini (through this process), dengan cara demikian (thereby), etc

Location sentence refers to a location of the activity. On the data sources, a type of sentence can be identified to form this type of sentence.

LocationType-1

Sentence Pattern: Event + Passive Verb (Passive VBT) or Intransitive Verb (VBI) + location-keyword + Location + Explanation (optional)

Example: Perekaman objek permukaan bumi (event) dilakukan (passive verb) di (location-keyword) luar angkasa (location)

The recording of object of the earth surface(event) is conducted (passive verb) in (location-keyword) the outer space (location) (Sebuah gempa bumi berkekuatan 6,9 magnitud (event) terjadi (Intransitive Verb) di (location-keyword) barat-daya Pichilemu, Region O'Higgins (location)

6.9 magnitude earthquake (event) occurs (Intransitive Verb) in (location-keyword) south-west of Pichilemu, O'Higgins Region (location))

Location-keywords: Menuju (toward), di (at), ke (to), dari (from), pada (on), dalam (within/in), etc

Entity sentence relates to utterance towards sides like institution, personal, and particular group. This type of sentence can be formed through three ways.

EntityType-1

Sentence Pattern: Object + Passive Verb (Passive VBT) + entity-keyword + Subject + Preposition (optional) + Explanation (Optional)

Example: Kopi pertama kali (object) ditemukan (passive verb) oleh (entity-keyword) Bangsa Etiopia (subject) sebagai (preposition) minuman berkhasiat dan berenergi (explanation) *(Coffee was first (object) discovered (passive verb) by (entity-keyword) the Ethiopians (subject) as (preposition) nutritious and energizing drinks (explanation))*

EntityType-2

Sentence Pattern: Entity-subject + Active Verb (VBT) + Entity-object + Preposition (optional) + Explanation (Optional)

Example: Isaac Newton (subject) mengembangkan (active verb) teori matematika (object) yang (preposition) pada akhirnya berkembang menjadi kalkulus (explanation) *(Isaac Newton (subject) developed (active verb) mathematical theory (object) which (preposition) eventually evolved into calculus (explanation))*

Entity-keyword: Oleh (By)

Last is Date-Time sentence. It relates to time from an event; further, there are two ways to identify this type of sentence.

TimeType-1

Sentence Pattern: Time-keyword + Date + Event

Example: Pada tanggal (time-keyword) 10 Oktober 1991 (date) pecah pemberontakan di kota industri Wu Can (event) *(On (time-keyword) October 10th, 1991 (date) there was an uprising in industrial town of Wu Can (event))*

TimeType-2

Sentence Pattern: Event + time-keyword + Date

Example: Elvis membayar $4 untuk merekam dua buah lagu di perusahaan rekaman Sun Studios sebagai hadiah ulang tahun bagi ibunya (event) pada tahun (time-keyword) 1953 (date) *(Elvis paid $4 to record two songs at the Sun Studios record label as a birthday gift for his mother(event) on (time-keyword)1953 (date))*

time-keyword: pada (on) + {Primary Numeral (angka) (number)), januari s/d desember (January to December), tanggal (date), hari (day), pekan (weekend), minggu (week), bulan (month), triwulan (trimester), caturwulan (quadmester), tahun (year), musim (season), windu (windu), dekade (decade), dasawarsa (decade), abad (century), milenium (millennium), periode (period), saat (when)}, terjadi ketika (occur when), terjadi saat (occur at), terjadi pada masa (occur during), terjadi ketika (occur while), ketika (while), sejak (since), etc

The main problem of identifying the type of sentences is the appearance of the same keyword in among the types of sentences. For example, the Definition sentence can contain a keyword of Date-Time sentence; further, the Entity sentence can also contain a keyword for Date-Time sentence. Such situation also occurs in other types of sentences. Thus, the identification is done only to refer to the main characteristics of each sentence type and the issue will be resolved in the question generation section.

Table 1: The Datasets of Sentences

No	Coarse Type	Fine Type	Num. of Sentences	Total
1.	Definition	DefinitionType-1	42	173
		DefinitionType-2	43	
		DefinitionType-3	43	
		DefinitionType-4	45	
2.	List	ListType-1	121	121
3.	Cause-Effect	CauseEffectType-1	50	127
		CauseEffectType-2	37	
		CauseEffectType-3	40	
4.	Reason	ReasonType-1	50	153
		ReasonType-2	55	
		ReasonType-3	48	
5.	Method	MethodType-1	79	152
		MethodType-2	31	
		MethodType-3	42	
6.	Entity	EntityType-1	78	106
		EntityType-2	28	
7.	Location	LocationType-1	35	35
8.	Date-Time	TimeType-1	52	111
		TimeType-2	59	
Total Dataset				978

3. AUTOMATIC QUESTION GENERATION

3.1. Data Collection and Transformation

Figure 1 presents the overall system of open domain Indonesian question used in this research. The system starts from collecting eBooks form my school subjects. The high school subjects used in this research include Biology, Physics, Geography, History, Indonesian Literature, and Sociology. Meanwhile in vocational school, several subjects cover Puppetry Sciences, Engineering Antenna Systems, Artificial Feed Production, Agribusiness Poultry Feed, Soil and Water Conversion, and Fish Hatchery Techniques. These eBooks were downloaded from http://bse.kemdikbud.go.id/ and Wikipedia.

The next stage is Data Preprocessing. This stage consists of Text Transformation, Text Segmentation and Filtering, and Features Extraction. The text documents used are in many formats, and thus Text Transformation is necessarily required to convert the documents into PDF. Text Segmentation is necessary for detecting the sentences that remain in paragraph forms and then takes each sentence as the sources for the succeeding stage. The result of text segmentation comes in the collection of words, characters, and punctuations that are not relevant, such as: book titles, subtitles, table of contents, the list of figures, headers, footers, comments in brackets, and patterns of particular characters that do not have any meaning. In the further step, the process of Filtering was carried out according to the rules and dictionaries that are able to identify declarative sentences and eliminate the remaining segmentation results.

The next stage is feature extraction for the classification of sentences. Feature identification was performed by observing any kinds of sentences that become differentiator of one type of sentence to others. The extracted features used for coarse class classification (1st classification) are divided into training data and testing data. There are 8 coarse classes and 19 fine classes. The first phase was coarse-class classification based on a machine learning and the second phase was done based on the rules.

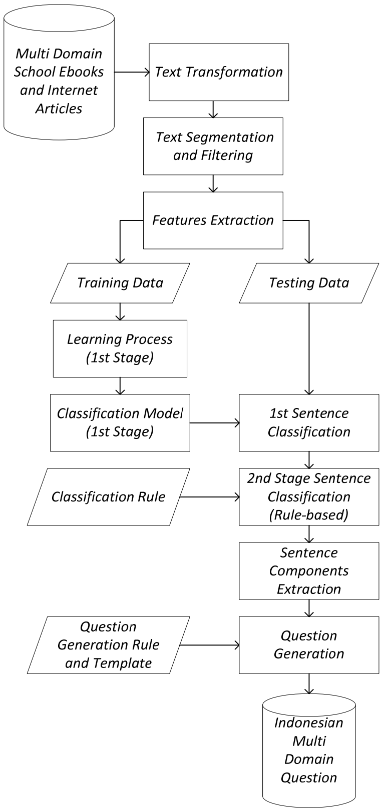
Multi Domain School Ebooks and Internet Articles
Text Transformation
Text Segmentation and Filtering
Features Extraction
Training Data
Testing Data
Learning Process (1st Stage)
Classification Model (1st Stage)
1st Sentence Classification
Classification Rule
2nd Stage Sentence Classification (Rule-based)
Sentence Components Extraction
Question Generation Rule and Template
Question Generation
Indonesian Multi Domain Question

Figure 1: The Architecture of Open Domain for Indonesian Question Generation

3.2. Coarse Sentence Type Classification

This section describes the features of the first classification phase along with the justification. The justification is determined based on manual observations in all declarative sentences.

The features used to identify Definition sentences are:

Table 2: Classification Features for Definition Sentences

No	Features	Values
1	The availability of a definition-keyword in a sentence *(All definition sentences must contain a definition-keyword)*	yes/no
2	The position of a definition-keyword in a sentence *(On the DefinitionType-1 sentence, a definition-keyword generally resides on the first five words)*	1/2/3/4/5
3	The availability of a comma before the definition-keyword *(The DefinitionType-2 sentences commonly contain a comma before the definition-keyword)*	yes/no
4	The availability of a colon or a semi-colon on the first 5 words *(On the DefinitionType-3 sentences, the existence of a definition-keyword can be replaced by colon or semi-colon.)*	yes/no
5	The availability of definition-keyword on the 5 last words *(On the DefinitionType-4 sentences, the definition-keyword commonly appears at the end of the sentence)*	yes/no
6	The availability of a preposition on the right of the definition-keyword *(The DefinitionType-1 and DefinitionType-2 sentences generally contain a preposition on the right of the definition-keyword which functions as explanation)*	yes/no
7	The availability of a preposition on the left of the definition-keyword *(The DefinitionType-4 sentences generally contain a preposition on the left of the definition-keyword which gives explanation)*	yes/no
8	The availability of all types of the Cause Effect-keywords or Reason-keywords on the right or left of the definition-keyword *(Definition sentence often contains keyword that refers to other types of sentences)*	yes/no

The features used to identify List sentences are:

Table 3: Classification Features for List Sentences

No	Features	Values
1	The availability of a comma on the left of the List-keyword *(List sentence generally contains a comma just before the List keyword which indicates the start of the exemplification)*	yes/no
2	The availability of a colon on the right of the List-keyword *(Most of the List sentences contain a colon after the List-keyword which indicates the beginning of exemplification)*	yes/no
3	The availability of a comma, a conjunction, or a disjunction on the right of the list-keyword *(The existence of the exemplification is marked by the present of a comma, a conjunction, or a disjunction after the list-keyword)*	yes/no
4	The availability of the Primary Numerals (CDP) before the List-keyword *(List sentence is generally preceded by the Primary Numeral (CDP))*	yes/no
5	The availability of a dividing keyword on the left of the List-keyword *(List sentences generally contain a dividing keyword before the list-keyword)*	yes/no

The features used to identify Cause-Effect sentences are:

Table 4: Classification Features for CauseEffect Sentences

No	Features	Values
1	The availability of the CauseEffect-1-keyword *(The CauseEffectType-1 sentence must contain a CauseEffect-1-keyword)*	yes/no
2	The availability of the CauseEffect-2-keyword *(CauseEffectType-2 sentence must contain a CauseEffect-2-keyword)*	yes/no
3	The availability of the CauseEffect-3-keyword *(CauseEffectType-3 sentence must contain a CauseEffect-3-keyword)*	yes/no
4	The availability of a comma just before the CauseEffect-3-keyword *(On the Effect side of the CauseEffectType-3 sentence, it probably contains a comma just before the CauseEffect-3-keyword)*	yes/no

The features used to identify Reason sentences are:

Table 5: Classification Features for Reason Sentences

No	Features	Values
1	The availability of a reason-keyword-1, a reason-keyword-2, or a reason-keyword-3 *(Reason sentences must contain a reason-keyword-1, a reason-keyword-2, or a reason-keyword-3)*	yes/no
2	The availability of a comma just before a reason-keyword-2 *(ReasonType-2 sentences probably have a comma just before the reason-keyword-2)*	yes/no
3	The availability of a reason-keyword-3 in the first sentence *(ReasonType-3 sentences are identified by the presence of a reason-keyword-3 at the beginning of the sentence)*	yes/no
4	Availability of comma just after the reason-keyword-3 *(Most of the ReasonType-3 sentences contain a comma just after the reason-keyword-3)*	yes/no

The features used to identify Method sentences are:

Table 6: Classification Features for Method Sentences

No	Features	Values
1	The availability of a method-keyword-1 *(The MethodType-1 sentences contain the method-keyword-1)*	yes/no
2	The availability of a passive verb (passive VBT) or an intransitive verb (VBI) in a sentence *(The MethodType-1 sentences often contain a passive VBT or a VBI before the method-keyword-1)*	yes/no
3	The availability of a method-keyword-2 *(The MethodType-2 sentences may contain the method-keywrod-2)*	yes/no
4	The presence of an active verb after the method-keyword-2 *(The MethodType-2 sentences commonly contain an active verb after the method-keyword-2)*	yes/no
5	The availability of a method-keyword-3 *(The MethodType-3 sentences commonly contain the method-keyword-3)*	yes/no
6	Availability of comma after method-keyword-3 *(MethodType-3 sentences may contain a comma just after the method-keyword-3)*	yes/no

The features used to identify Entity sentences are:

Table 7: Classification Features for Entity Sentences

No	Features	Values
1	The availability of the entity-keyword after a Passive Verb (Passive VBT) in a sentence *(The EntityType-1 sentences contain the entity-keyword after a Passive Verb)*	yes/no
2	The POSTag of a word after the entity-keyword *(The Entity-keyword of the Entity-Type-1 sentences is generally followed by a Noun (NN) or a Proper Noun (NNP))*	yes/no
3	The availability of an Active Verb *(The Entity-Type-1 sentences generally contain an Active Verb)*	yes/no
4	The POSTag of a word just before an Active Verb *(The EntityType-2 sentences may contain Proper Noun (NNP) or Noun (NN))*	yes/no
5	The POSTag of a word after an Active Verb *(The EntityType-2 sentences may contain Proper Noun (NNP) or Noun (NN))*	yes/no

The feature used to identify Location sentences is:

Table 8: Classification Features for Location Sentences

No	Features	Values
1	The availability of the location-keyword after a Passive Verb (Passive VBT) or an Intransitive Verb (VBI) *(Generally, the Location sentences contain the location-keyword before a Passive Verb)*	yes/no

The features used to identify Date-Time sentences are:

Table 9: Classification Features for Date-Time Sentences

No	Features	Values
1	The availability of the Time-keyword on a sentence *(The TimeType-1 sentences always contain the Time-keyword)*	yes/no
2	The POSTag of a word after the time-keyword *(The TimeType-1 sentences have a Primary Numerals (CDP) or a Proper Noun (NNP), which is placed after the time-keyword)*	yes/no

3.3. Fine Sentence Type Classification

Fine-class classification was carried out by using rules on the sentences which are classified at the previous stage. The classification on this phase was solely conducted on 6 sentence classes out of 8 (2 sentence classes do not have a fine-class that are List and Location). The rules were established by observing the characteristics of each fine-class out of 19 classes in total. The following pseudo-code below describes the set of rules established for fine-class classification, (1) definition class: The existence of definition-keywords in the first 5 words of a sentence is a character from DefinitionType1 and DefinitionType2 (if there is a comma before the keyword, it is then categorized as DefinitionType-2. If not, then DefinitionType-1). If the sentence contains ":" sign then it can be categorized as DefinitionType-3. If it is not categorized as both, then it is categorized as DefinitionType-4, (2) entity class: EntityType-1 is identified by the presence of an active verb, while EntityType-2 is identified by the presence of Noun or Proper Noun, (3) method class: Type MethodType-1 is identified by the existence of method-keyword-1 in the sentence. MethodType-2 is characterized by the method-keyword-2 at the beginning of the sentence. If it does not belong to both, then it is included into MethodType-3, (4) reason class: ReasonType-1 is identified by the existence of reason-keyword-1 in the middle of the sentence. ResonType-2 is known from the existence of reason-keyword-2. If it is not categorized as both, then it is included into ReasonType-3, (5) cause-effect class: The CauseType-1 type is identified by the presence of cause-keyword-1. CauseType-2 is identified by the presence of cause-keyword-2. If it is not categorized as both, it is then categorized as CauseType-3, (6) date-time: Checking time-keyword positions in sentences; a sentence is categorized as TimeType-2 if there is a relative at the beginning of a sentence; meanwhile, the TimeType-1 is categorized if there is a relative at the end of a sentence.

3.4. Question Generation

The function of this component is to generate the question from the sentence that has been classified in the previous stage. This component consists of two sub-components: The Extraction of Components and the Generation of Question. In addition, these components contain the generation rules to generate questions automatically. The function of Component Extraction is to take some components of a sentence. The extraction was conducted by using rules adjusted to the sentence class. Extraction rules are different for each sentence class. The sub-components of Question Generation are responsible for arranging the sentence's component that has been extracted into question.

Generating Definition Questions

Pattern 1: Question Expression for Definition + entity

Generated Question: Apakah yang dimaksud dengan (Question Expression for Definition) Lokasi relatif (entity) (*What is the meaning of (Question Expression for Definition) relative location (entity)*)

Pattern 2: Explanation + definition-keyword + question-mark "?"

Generated Question:

Lokasi sesuatu objek yang nilainya ditentukan oleh objek-objek lain di luarnya (explanation) disebut sebagai (definition-keyword) ? (question-mark) *(Location of object which the value is determined by other objects (explanation) is called as? (definition-keyword) ? (Question-mark))*

/ Pseudo Code for Generating Definition Question*/*

```
Procedure DefinitionGeneration
Input:
    FineClassificationResult
    DefinitionType-n Pattern
    Declarative Sentence
Output:
    GeneratedDefinitionQuestion
start
  extract:
      entity
      definition-keyword
      explanation
generating definition questions according Pattern-1 and Pattern-2
end
```

Question Expression for Definition: Apakah yang dimaksud, Apa yang dimaksud, Jelaskan, Jelaskan yang dimaksud dengan, dsb *(What is the meaning, What the meanings, Explain, explain what the is meaning of, etc).*

Generating List Questions

Pattern 1: List-statement + ":"

Generated Question: Untuk mencegah terjadinya erosi (pernyataan-perincian) + ":" (*To prevent the erosion (statement-specification) + ":"*)

Pattern 2: Question Expression for Definition + entity-1 … entity-n

Generated Question: Apakah yang dimaksud (Question Expression for Definition) dengan lapisan kedap (impermeable) (entity) (*What is the meaning (Question Expression for Definition) of impermeable layer (impermeable) (entity)*)

/ Pseudo Code for Generating List Question*/*

```
Procedure ListGeneration
Input:
    FineClassificationResult
    List Pattern
    Declarative Sentence
Output:
    GeneratedListQuestion
 start
   extract:
        list-statement
        entity-1, entity-2, … entity-n
    generating list questions according Pattern-1 and Pattern 2
end
```

Generating Cause-Effect Questions

Pattern 1: Question Expression for Cause + Effect

Generated Question: Apakah penyebab dari (Question Expression for Cause) perbedaan kandungan zat larutan atau organisme yang ada di dalam laut tersebut (Effect) (*What is the cause of (Question Expression for Cause) differences of substances or organisms in the sea (Effect)*)

Pattern 2: Question Expression for Effect + Cause

Generated Question: Apa akibat dari (Question Expression for Effect) perbedaan warna air laut (Cause) (*What is the result of (Question Expression for Effect) different colors in the sea (Cause)*)

Question Expression for Cause: Apakah akibat dari (*What is the result of*)

Question Expression for Effect: Apakah sebab dari (*What is the cause of*)

/ Pseudo Code for Generating Cause-Effect Question*/*

```
Procedure CauseEffectGeneration
Input:
    FineClassificationResult
    CauseEffectType-n Pattern
    Declarative Sentence
Output:
    GeneratedCauseEffectQuestion
start
    extract:
        cause
        effect
    generating cause-effect questions according to Pattern-1 and Pattern-2
end
```

Generating Method/Stage Questions

Pattern 1: Question Expression for Method-1 + target-method + Question Mark (?)

Generated Question: Bagaimana cara (Question Expression for Method-1) reproduksi aseksual (target-method)? *(How (Question Expression for Method-1) does asexual reproduction works (target-method)?)*

Pattern 2: Detail-method + Question Expression for Method-2 + Question Mark (?)

Generated Question: Membentuk tunas eksternal atau tunas internal (detail-method) adalah proses dari (Question Expression for Method-2)? *(Forming external or internal buds (detail-method) is the process of (Question Expression for Method-2)?)*

Question Expression for Method-1: Bagaimana cara (*How/How to*)

Question Expression for Method-2: Adalah method dari (*Is the method of*)

/ Pseudo Code for Generating Method Question*/*

```
Procedure MethodGeneration
Input:
    FineClassificationResult
    MethodType-n Pattern
    Declarative Sentence
Output:
    GeneratedMethodQuestion
start
    extract:
        target-method
        detail-method
    generating method/stage questions according to Pattern-1 and Pattern-2
end
```

Generating Reason Questions

Pattern 1: Question Expression for Reason-1 + fact + Question Mark (?)

Generated Question: Mengapa (Question Expression for Reason) Mesir Pertengahan mengalami kemunduran (fact) ? (*Why (Question Expression for Reason) was Middle Egypt lost ground (fact)?*)

Pattern 2: Question Expression for Reason-2 + reason + Question Mark (?)

Generated Question: Apa dampak dari (Question Expression for Reason-2) Indonesia terletak di daerah sekitar khatulistiwa (reason) ? (*What is the impact of (Question Expression for Reason 2) Indonesia's location that lies along the equator (reason)?*)

Question Expression for Reason-1: Mengapa (*Why)*

Question Expression for Reason-2: Apa dampak dari (*What is the impact of*)

/ Pseudo Code for Generating Reason question*/*

```
Procedure ReasonGeneration
Input:
    FineClassificationResult
    ReasonType-n Pattern
    Declarative Sentence
Output:
    GeneratedReasonQuestion
start
    extract:
        reason
        fact
    generating reason questions according to Pattern-1 and Pattern-2
end
```

Generating Entity Questions

Pattern 1: Question Expression for Entity + Event + Question Mark (?)

Generated Question: Siapakah pelaku (Question Expression for Entity) Respirasi Anaerob (event)? *(Who is the doer of (Question Expression for Entity) Anaerobic Respiration (event)?)*

/ Pseudo Code for Generating Entity Question*/*

```
Procedure EntityGeneration
Input:
    FineClassificationResult
    EntityType-n Pattern
    Declarative Sentence
Output:
    GeneratedEntityQuestion
start
    extract:
        event
    generating Entity questions according to Pattern-1
end
```

Question Expression for Entity: Siapakah pelaku, siapakah (*Who is the doer, who is)*

Generating Location Questions

Pattern 1: Question Expression for Location + Event

Generated Question: Dimanakah (Question Expression for Location) perekaman objek permukaan bumi dilakukan (event)? (*Where is (Question Expression for Location) the recording of object of the earth surface conducted (event)?)*

/ Pseudo Code for Generating Location Question*/*

```
Procedure LocationGeneration
Input:
    FineClassificationResult
    Location Pattern
    Declarative Sentence
Output:
    GeneratedLocationQuestion
start
  extract:
    event
  generating location questions according to Pattern-1
end
```

Question Expression for Location: Dimanakah (*Where)*

Generating Date-Time Questions

Pattern 1: Question Expression for Time + Event + Question Mark (?)

Generated Question: Kapankah (Question Expression for Time) pecah pemberontakan di kota industri Wu Can, yang kemudian merembet ke kota-kota lain (event) (*When did (Expression for Question Time) the uprising in the industrial city of Wu Can that later spread out to other cities happen (event)*

/ Pseudo Code for Generating Date-Time Question*/*

```
Procedure DateTimeGeneration
Input:
    FineClassificationResult
    TimeType-n Pattern
    Declarative Sentence
Output:
    GeneratedDateTimeQuestion
start
  extract:
    Event
  generating date-time questions according to Pattern-1
end
```

Question Expression for Time: Kapankah, kapan (*When is, when)*

The question generation process from a text is shown in the example below.

Hutan hujan tropika

Hutan hujan tropis di Amazon

Hutan hujan tropika atau sering juga ditulis sebagai hutan hujan tropis adalah bioma berupa hutan yang selalu basah atau lembap, yang dapat ditemui di wilayah sekitar khatulistiwa; yakni kurang lebih pada lintang 0°–10° ke utara dan ke selatan garis khatulistiwa. Hutan hujan tropis bisa juga diartikan sebagai hutan yang terletak di daerah tropis yang memiliki curah hujan tinggi. Maka dari itu, disebut Hutan Hujan Tropis.Hutan-hutan ini didapati di Asia, Australia, Afrika, Amerika Selatan, Amerika Tengah, Meksiko dan Kepulauan Pasifik. Dalam peristilahan bahasa Inggris, formasi hutan ini dikenal sebagai lowland equatorial evergreen rainforest, tropical lowland evergreen rainforest, atau secara ringkas disebut tropical rainforest.

SENTENCE DETECTION:

(sentence detection and paragraphs segmentation)

[LINE 1] *[TITLE]* Hutan hujan tropika

[LINE 2] *[SUBTITLE]* Hutan hujan tropis di Amazon

[LINE 3] Hutan Hujan Tropika (HHT) atau sering juga ditulis sebagai hutan hujan tropis adalah bioma berupa hutan yang selalu basah atau lembap, yang dapat ditemui di wilayah sekitar khatulistiwa.

[LINE 4] HHT berada pada kurang lebih pada lintang 0°–10° ke utara dan ke selatan garis khatulistiwa.

[LINE 5] Hutan hujan tropis bisa juga diartikan sebagai hutan yang terletak di daerah tropis yang memiliki curah hujan tinggi.

[LINE 6] Maka dari itu, disebut Hutan Hujan Tropis.Hutan-hutan ini didapati di Asia, Australia, Afrika, Amerika Selatan, Amerika Tengah, Meksiko dan Kepulauan Pasifik.

[LINE 7] Dalam peristilahan bahasa Inggris, formasi hutan ini dikenal sebagai lowland equatorial evergreen rainforest, tropical lowland evergreen rainforest, atau secara ringkas disebut tropical rainforest.

SENTENCE FILTERING:

(Filtering, in this example, to eliminate the titles and subtitles)

[LINE 1] Hutan Hujan Tropika (HHT) atau sering juga ditulis sebagai hutan hujan tropis adalah bioma berupa hutan yang selalu basah atau lembap, yang dapat ditemui di wilayah sekitar khatulistiwa.

[LINE 2] HHT berada di kurang lebih pada lintang 0°–10° ke utara dan ke selatan garis khatulistiwa.

…..

[LINE 5] Dalam peristilahan bahasa Inggris, formasi hutan ini dikenal sebagai lowland equatorial evergreen rainforest, tropical lowland evergreen rainforest, atau secara ringkas disebut tropical rainforest.

COARSE-CLASS CLASSIFICATION:

(only take three sentences)

[LINE 1] Hutan Hujan Tropika (HHT) atau sering juga ditulis sebagai hutan hujan tropis adalah bioma berupa hutan yang selalu basah atau lembap, yang dapat ditemui di wilayah sekitar khatulistiwa. ***(Classified as Definition)***

[LINE 2] HHT berada di kurang lebih pada lintang 0°–10° ke utara dan ke selatan garis khatulistiwa. ***(Classified as Location)***

…..

[LINE 5] Dalam peristilahan bahasa Inggris, formasi hutan ini dikenal sebagai lowland equatorial evergreen

rainforest, tropical lowland evergreen rainforest, atau secara ringkas disebut tropical rainforest. ***(Classified as Definition)***
……

FINE-CLASS CLASSIFICATION:
(just take one example: LINE 1)
[LINE 1] Hutan Hujan Tropika (HHT) atau sering juga ditulis sebagai hutan hujan tropis adalah bioma berupa hutan yang selalu basah atau lembap, yang dapat ditemui di wilayah sekitar khatulistiwa. ***(Classified as DefinitionType-1)***

COMPONENT EXTRACTION:
Sentence Pattern for DefinitionType-1: Entity + definition-keyword + explanation
Hutan Hujan Tropika (HHT) atau sering juga ditulis sebagai hutan hujan tropis (Entity) adalah (definition-keyword) bioma berupa hutan yang selalu basah atau lembap, yang dapat ditemui di wilayah sekitar khatulistiwa (explanation)

QUESTION GENERATION:
Pattern 1: Question Expression for Definition + entity
Generated Question:
Apakah yang dimaksud dengan (Question Expression for Definition) Hutan Hujan Tropika (HHT) atau sering juga ditulis sebagai hutan hujan tropis (entity)
What is (Question Expression for Definition) a Tropical Rain Forests (HHT) or often also written as tropical rain forests (entity)
Pattern 2: Explanation + definition-keyword + question-mark "?"
Generated Question:
Bioma berupa hutan yang selalu basah atau lembap, yang dapat ditemui di wilayah sekitar khatulistiwa (explanation) disebut sebagai (definition-keyword)? (question-mark)
A form of forest biome is always wet or damp, which can be found in the region around the equator (explanation) known as (definition-keyword)? (question-mark)

4. RESULT AND DISUCUSSION

In this part, we performed evaluation tests of three components of the system namely coarse class classification, fine class classification, and question generation. The classification evaluation for both coarse class and fine class is focused on the classification accuracy. For question generation evaluation, the achievement is measured by comparing the generated question syntactic structure with question pattern that we have defined in the previous section. There are three categories of generated question results namely TRUE, UNDERSTANDABLE, and FALSE. If the generated questions follow the question pattern, the they will be categorized as TRUE. Otherwise, they will be categorized as FALSE. However, there are some of generated questions that do not match with the question pattern, but they can be accepted as TRUE questions since the readers are able to understand the meaning of that questions. This type of question is categorized as UNDERSTANDABLE.

4.1. The Accuracy of Sentence Classification

The experiments were carried out on two levels of classification incorporating coarse-class classification by using a machine learning that is followed by fine-class classification (not all coarse classes have a fine class). This classification was conducted on a dataset of 233 sentences representing all types of sentences. The highest accuracy value is 81.73% obtained by using SMO algorithm. The lowest value is 73.47% obtained by using Naïve Bayes algorithm and Random Forest. The Detailed accuracy values are shown in Table 10 below.

Table 1: The Accuracy of Coarse-Class Classification

Classifier	SMO	Naïve Bayes	Ibk	J48	Random Forest
Accuracy (%)	81.73	73.47	77.82	80.86	73.47

The misclassifications in this stage are dominated by Date-Time sentence with 9 sample sentences. From 9 sentences, there are 4 types of misclassification that should be Date-Time sentences, yet it is classified as Location sentence and the rest are 5 various mistakes. Classification error of Date-Time into Location occurred because 4 sampling test sets have the characteristics of Date-Time and Location. One of the test data is "In 2012, either Mandarin or English Wikipedia could access in China, except for politic article". This example could consider as sort of Date-Time sentences which is identified with Date-Time as "In 2012" and Location as "In China". As a solution, we could add a checking rule to each of classification result such as Location and Time sentence. If the result of classification is Location sentence and has the characteristic of Time Keyword, then it could be classified as Date-Time. Using classifier SMO could make the accuracy of addition this rule increased to 83.26%.

The result of coarse-class classification then forwarded to the rule-based fine-class classification stage. The experiment was conducted to the same dataset from the previous stage classification process to measure the accuracy of the rule that has been built. Because not all the coarse class have fine class, thus class classification was merely conducted to 6 sentence classes, such as: Definition, Cause-Effect, Reason, Method, Entity, and Date-Time. As a result, there are 169 test data used. The highest accuracy is obtained on the rule to classify the Entity sentence that reaches 100% and the lowest is the rule to classify the Time sentence that reaches 83%. The average accuracy of Fine-Class Classification for all types of sentences are 92%.

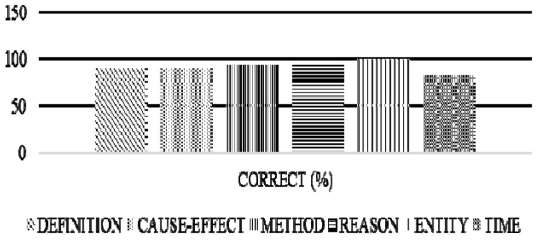

Figure 2: The Accuracy of Fine-Class Classification

4.2. The Accuracy of Question Generation

The accuracy of question generation was carried out to the sentences that have been classified in the previous stage. Generation process was preceded by the component extraction of each sentence. The accuracy measurement of all the question was conducted by comparing each generated question with a pattern of rules. The accuracy of question generation is categorized into three: TRUE, FALSE, and UNDERSTANDABLE. Generation is rated as TRUE when the question produced in accordance with the pattern of question that refers to the rules. Generation is rated as FALSE if the question produced is not understandable and does not follow the rules. The last, when the question generated is not in accordance with the rules yet the meaning of the question is UNDERSTANDABLE, the generation could be rated as UNDERSTANDABLE.

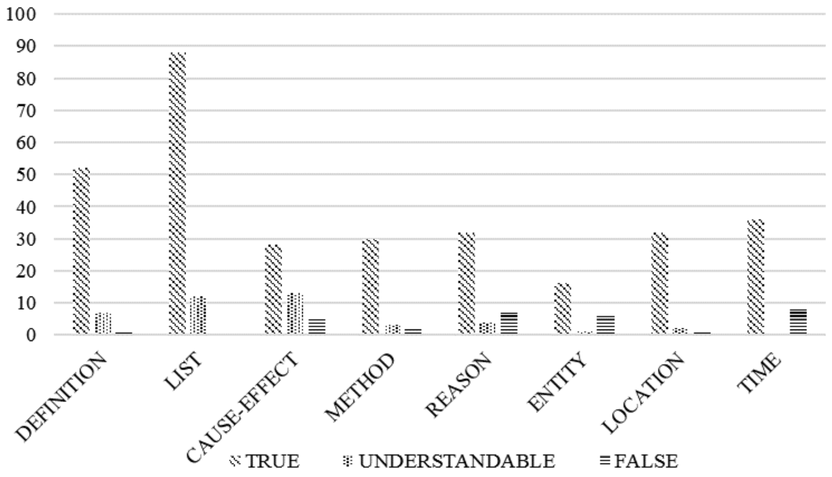

Figure 3: The Distribution of Question Generation for three categories TRUE, FALSE, and UNDERSTANDABLE (%)

Table 10: The Accuracy Comparison of Resurrected Question

	DEFINITION	LIST	CAUSE-EFFECT	METHOD	REASON	ENTITY	LOCATION	TIME
TRUE	52	88	28	30	32	16	32	36
UNDERSTANDABLE	7	12	13	3	4	1	2	0
FALSE	1	0	5	2	7	6	1	8
TOTAL OF QUESTION	60	100	46	35	43	23	35	44
TRUE + UNDERSTANDABLE ACCURACY (%)								
ACCURACY	98.33	100.00	89.13	94.29	83.72	73.91	97.14	81.82

Here are examples of generated questions for all categories. The questions categorized as TRUE are (Definition): “Bagian laut yang terletak di antara garis air surut sampai kedalaman 200 m disebut?” (Which part of the sea that lies between the low tide lines to depths of 200 m?) (List): “Jelaskan berturut-turut nama-nama planet yang masuk susunan matahari?” (Explain consecutive names of the planets that include to the composition of the sun?) (Cause-Effect): “Apakah akibat dari penambahan satu kromosom pada kromosom ke 13?” (What is the effect of the addition of one chromosome to chromosome number 13?) (Method): “Bagaimana cara membentuk senyawa 3c, yaitu asam laktat, tanpa melepaskan co2?” (How to form compound 3c, namely lactic acid, without releasing co2?) (Reason): “Apa dampak dari banyaknya pembukaan hutan?” (What is the impact of Deforestation?) (Entity): “Siapakah mark zuckerberg?” (Who is Mark Zuckerberg?) (Location): “Dimanakah merekam objek permukaan bumi bisa dilakukan?” (Where can we record the object of earth’s surface?) (Time): “Kapan 'world wide web' sudah dapat dijalankan dalam lingkungan CERN?” (When has 'World Wide Web' been able to run in the CERN environment?).

The questions categorized as FALSE are, (Definition): “Apakah akibat dari maka disebut karnivora atau konsumen sekunder?” (What is the effect of being called carnivores or secondary consumers?) (List): “Apakah saja akibat dari protoplas akan keluar menjadi badan yang disebut auksospora?” (What are the consequences of the protoplasts that come out into a body called auksospora?) (Cause-Effect): ”Apakah sebab dari hal ini ?” (What is the cause of this?) (Method): “Sejarah akan mudah dicerna, diingat dan melekat adalah proses dari?” (History will be easy to digest, remember and attach is the process of?) (Reason): N/A (Entity): “Siapakah jenis hewan dan jenis makanan hewan tersebut?” (What is the type of that animal and their food?) (Location): “Dimanakah Adanya minimarket berjaringan menyebabkan toko tradisional?” (Where is the minimarket networked cause traditional stores?) (Time): “Kapankah kemudian masuk?” (When does it then get in?).

Lastly, the questions categorized as UNDERSTANDABLE are (Definition): “Iklim berdasarkan tinggi tempat dan jenis tanaman yang tumbuh baik disebut?” (What kind of climate based on the place elevation and types of well-grown plants?) (List): “Apakah yang dimaksud keanekaragaman hayati 93 ada dua macam upaya pelestarian keanekaragaman hayati di Indonesia?” (What is biodiversity 93 that has two kinds of conservative efforts in Indonesia?) (Cause-Effect): “Apakah sebab dari aberasi kromosom yang?” (What are the causes of chromosomal aberrations?) (Method): “Dapat diperkecil keragaman antar anggota populasi, karena telah terjadi pengelompokkan sebelumnya adalah proses dari?” (It could be reduced diversity among members of the population because it happened before, it is the grouping process of?) (Reason): “Mengapa terjadilah perbedaan penyinaran di muka bumi?” (Why is there a different radiation on the earth?) (Entity): “Siapakah timberners-lee yaitu pada tahun 1991?” (Who was Tim Berners-Lee in 1991?) (Location): “Dimanakah Kesigapan tim pemadam kebakaran membuat api tidak merambat?” (Where does the alertness of firefighting team make a fire would not spread?) (Time): N/A.

4.3. Discussion

Each kind of sentence could be formed into one or more questions depending on the components that can be extracted. The experiments generated 386 questions. Figure 3 shows the distribution of generated questions along with the amount of the questions categorized as TRUE, UNDERSTANDABLE, and FALSE. In general, the distribution is dominated by TRUE questions. The accuracy is taken from the percentage of generated TRUE and UNDERSTANDABLE questions. The justification used is that the UNDERSTANDABLE questions basically are TRUE question, but the pattern does not exactly match the structure of well-developed questions. The quality of the generated questions is assessed by looking at the structure of the generated questions. The structure of standard Indonesian questions refers to the eight types of question sentences. The generated questions that meet such structure will be declared as TRUE; otherwise, it will be declared as FALSE. In other hands, UNDERSTANDABLE is basically a type of TRUE sentence but there are minor errors such as an error on affixes, pronouns, question word/expression, conjunction, and so forth. Table 11 shows the detail of the distribution of generated questions. The highest score is achieved from the rule to generate the LIST questions which reaches 100%. Meanwhile, the lowest score is obtained from generating the Entity questions that merely reaches 73.91%. However, the rules of question generation can be categorized successfully to form questions since the average success of all generated questions is 88.66%.

From the generation results, there are some limitations that the proposed method fails to generate the TRUE questions. There are two factors causing the error. The first factor is strongly influenced by the result of the previous process that is coarse and fine-class classification. If there is an error in coarse-class classification, the next step such as fine-class classification, sentence component extraction, and question generation will be wrong. If an error appears in fine-class classification, the next step including the sentence component extraction and the question generation will be wrong too. These classification errors may occur because there are sentences containing of two or more keywords for different types of sentences. The second factor is influenced by a wide variety of characters of declarative sentences such as co-reference, pronouns, complex sentences, etc.

The previous researches of Indonesian AQG focused on domain specific questions as in [17][18][19]. They built the AQG in the medical field. One of our reference proposed open domain Indonesian AQG [20] to generate three types of questions namely definition question, reason question, and method question that reaches 55.95% accuracy in generating the questions. In our research, we not only utilize more variative data sources (school eBooks), but also our proposed method able to generate more question types. Our method outperforms the previous work that achieve the generation accuracy of 88.66%. It shows that our technique used is potentially to implement in the real situation.

5. CONCLUSION

This research aims to develop Automatic Question Generation for Open Domain of Indonesian Questions. The proposed technique was carried out based on the syntactical approach concerning on the structure of each declarative sentence. The technique consists of three main modules, namely: Coarse-Class Classification, Fine-Class Classification, and Question Generation. Based on the evaluation of all stages, the accuracy of techniques reaches 80%. The accuracy of Coarse-Class Classification reaches 81.73% by using the SMO classifier, and the average accuracy of Fine-Class Classification reaches 92%. Further, the average accuracy of the question generation reaches 88.66%. The obtained results show that the proposed method is potential to implement in generating the open domain Indonesian questions automatically.

For further development, it is paramount to provide solutions related to these two issues: Firstly, it is necessary to accommodate declarative sentences containing two or more types of sentence's keywords. This condition needs to be solved because it is impossible to limit the writing content that merely contains one type of sentence's keyword. This causes error in Coarse-Class Classification. Secondly, there are various characters of declarative sentences in the text documents, for instance: the presence of co-reference contents, pronouns, and complex sentences makes the rules fail to extract the sentence's components. Overcoming these issues will improve the accuracy of the Coarse-Class Classification and Fine-Class Classification and increase the number of TRUE and UNDERSTANDABLE generated questions as well.

REFRENCES:

[1] A. E. Awad, "Automatic Generation of Question Bank Based on Pre-defined Templates", *International Journal of Innovations & Advancement in Computer Science (IJIACS)*, vol. 3, no. 1, 2014, pp. 80–87.

[2] N. Afzal and R. Mitkov, "Automatic generation of multiple choice questions using dependency-based semantic relations", *Soft Computing*, vol. 18, no. 7, 2014, pp. 1269–1281.

[3] I. E. Fattoh, "Semantic Attributes Model for Automatic Generation of Multiple Choice Questions", *International Journal of Computer Applications*, vol. 103, no. 1, 2014, pp. 18–24.

[4] N. Afzal, M. Clinic, P. Arterial, D. View, and N. Afzal, "Automatic Generation of Multiple Choice Questions using Surface-based Semantic Relations", *International Journal of Computational Linguistics*, vol. 6, no. 3, 2015, pp. 26–44.

[5] M. Liu, R. A. Calvo, V. Rus, R. A. Calvo RAFAELCALVO, P. Piwek, and K. Elizabeth Boyer, "G-Asks: An Intelligent Automatic Question Generation System for Academic Writing Support", *Dialogue and Discourse*, vol. 3, no. 2, 2012, pp. 101–124.

[6] H. Ali, Y. Chali, and S. A. Hasan, "Automation of Question Generation From Sentences", *Proceedings of the Third Workshop on Question Generation*, Pittsburgh, USA, 2010, pp. 58–68.

[7] Y. Chali and S. a. Hasan, "Towards Automatic Topical Question Generation," *Proceedings of*

COLING 2012: Technical Papers, no. December, 2012, pp. 475–492.

[8] S. Curto, A. C. Mendes, and L. Coheur, “A minimally supervised approach for question generation: what can we learn from a single seed?”, *15th Portuguese Conference on Artificial Intelligence*, 2011, pp. 832-844.

[9] M. Heilman and N. Smith, “Extracting simplified statements for factual question generation”, *QG2010: Proceedings of the Third Workshop on Question Generation*, Pittsburgh, Pennsylvania, USA, June 18, 2010, pp. 11–20.

[10] H. Hussein, M. Elmogy, and S. Guirguis, “Automatic English Question Generation System Based on Template Driven Scheme”, *International Journal of Computer Science Issues (IJCSI)*, vol. 11, no. 6, 2014, pp. 45–53.

[11] M. Liu, R. A. Calvo, V. Rus, and I. Engineering, “Hybrid Question Generation Approach for Critical Review Writing Support”, *Proceedings of the 20th International Conference on Computers in Education*, Singapore, 2012, pp. 2–4.

[12] K. Mazidi and R. D. Nielsen, “Automatically Generating Questions From Text”, University of North Texas, 2014.

[13] S. Ou, C. Orasan, D. Mekhaldi, and L. Hasler, “Automatic question pattern generation for ontology-based question answering”, *Proceedings of the Twenty-First International FLAIRS Conference*, 2008, pp. 183–188.

[14] W. Wang, T. Hao, and W. Liu, “Automatic Question Generation for Learning Evaluation in Medicine”, *ICWL 2007: Advances in Web Based Learning*, 2007, pp. 242–251.

[15] H. D. A. D. Ali, “Thesis: Automatic question generation : a syntactical approach to the sentence-to-question generation case”, University of Lethbridge, Canada, 2012.

[16] K. Mazidi and R. D. Nielsen, “Linguistic considerations in automatic question generation”, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers)*, Baltimore, Maryland, USA, June 23-25, 2014, vol. 2, pp. 321–326.

[17] W. Suwarningsih, I. Supriana, and A. Purwarianti, “Toward a Framework for Indonesian Medical Question Generator”, *TELKOMNIKA*, Vol.13, No.1, March 2015, pp. 357-363.

[18] W. Suwarningsih, I. Supriana, and A. Purwarianti, “Pattern discovery using QG for question-answering pairs”, *International Journal on Electrical Engineering and Informatics*, vol. 8, no. 2, 2016, pp. 237–252.

[19] W. Suwarningsih, I. Supriana, and A. Purwarianti, “Indonesian Medical Sentence Transformation for Question Generation”, *IEEE Asia Pacific Conference on Wireless and Mobile (APWiMob)*, 2015, pp. 68–71.

[20] M. Fachrurrozi, N. Yusliani, J. Teknik, I. Universitas, J. Teknik, and I. Universitas, “Sistem pembangkit pertanyaan otomatis dengan metode template-based”, *Journal Research Computer Science & Application*, vol. 2, no. 1, 2013, pp. 24–29.