Supervised Learning Indonesian Gloss Acquisition

Gunawan, Member, IAENG, I Ketut Eddy Purnama, Mochamad Hariadi, Member, IAENG

***Abstract*—Aim of the research is to conduct automatic Indonesian synonym sets gloss extraction using the supervised learning approach. The main sources used are collections of web documents containing the gloss of the synonym sets. Three main phases of the proposed method are: preprocessing phase, features extraction phase, and classification phase. Preprocessing phase includes large scale fetch of web documents collection, extraction of raw text, text clean-up, and extraction of sentence from possible gloss candidates. Furthermore, in the features extraction phase, seven features are extracted from each of the gloss candidates: the position of a sentence in a paragraph, the frequency of a sentence in the document collection, the number of words in a sentence, the number of important words in a sentence, the number of characters in a sentence, the number of gloss sentences from the same word, and the number of nouns in the sentence. Lastly, in the classification phase, the supervised learning method will then accept or reject the candidate as a true gloss based on those seven features. It is shown in this paper that the proposed system was successful in acquiring 6,520 Indonesian synset glosses, with an average accuracy of 74.06% and 75.40% using the decision tree and backpropagation feedforward neural networks respectively. Thus, with the vast amount of successfully acquired glosses which is quite significant for Indonesian words, it is believed that the supervised learning approach used in this research will be useful to accelerate the process of lexical database formation such as WordNet for other languages.**

***Index Terms*—Gloss acquisition, Indonesian language, supervised learning, WordNet**

I. Introduction

LEXICAL database such as WordNet for natural language is absolutely necessary for the advancement of research on disciplines such as natural language processing or computational linguistic, information retrieval, as well as text and web mining for the language itself.

Today, there are millions of web pages in Indonesian (*Bahasa Indonesia*). With approximately 55 million Internet users, Indonesia is 8th country in the world with the most Internet users. Ironically, there is only a handful of the aforementioned researches for Indonesian.

Manuscript received February 11th, 2015; first revision received June 14th, 2015; second revision received August 14th, 2015.

Gunawan is with Computer Science Department, Sekolah Tinggi Teknik Surabaya, Surabaya, 60284, East Java, Indonesia (phone number: +62-818398761; e-mail: gunawan@stts.edu).

I Ketut Eddy Purnama is with Electrical Engineering Department, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia (email: ketut@ee.its.ac.id).

Mochamad Hariadi is with Electrical Engineering Department, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia (email: mochar@ee.its.ac.id).

Take the research collections in the Machine Translation category for instance. As of August 2014, there are more than 11,050 papers in MT-Archieve, a repository of articles, books and papers on topics in machine translation and computer-based translation tools (www.mt-archieve.info). However there have only been 20 text translation researches from or into *Bahasa Indonesia* in the span of almost 45 years (1970-2014). This value is clearly minuscule in comparison to researches from or into other Asian languages, e.g. Chinese, Korean, Japanese, Hindi, and Arabic where each of them in the range of hundreds each.

We believe that one fundamental cause is the scarcity of Indonesian lexical database that can be used freely by other researchers that need them. This is a stark contrast to similar researches in English. Hence, the Indonesian lexical database formation would definitely open the path for thousands of research and application developments in the said disciplines in Indonesian.

The internal structure of the lexicon in WordNet is aimed to support maximal access speed in the synset, gloss and all relationships –either through the library or the browser for online access– and it has been adopted by all languages in the world that have WordNet. This fast-growing adoption is certainly influenced by the free and open-source nature of WordNet. Therefore, this internal structure is unlikely to pose any problem when WordNet is proposed to be implemented into a new language.

The major challenge would be in the necessary work to acquire words, synset, gloss, and the complete relation of a language into WordNet. This challenge is inevitable when attempting to implement WordNet in Indonesian.

The complete WordNet development framework for Indonesian, which we have developed, was commenced with the project initialization focused on gaining Indonesian synonym set (*synset*) using the *Kamus Besar Bahasa Indonesia* (KBBI, The Indonesian Dictionary) as a monolingual lexical resources [1]. Synset itself can be viewed as the smallest unit in a lexical database that has the relationship between one another.

The next attempt is to acquire a semantic relation between hypernym-hyponym among the noun synset [2] and the last is to provide a graphical web-based tool to collaboratively fix the Indonesian lexical database [3].

The use of the KBBI as the monolingual lexical resource in the previous researches has actually provided the gloss of a synset [4]. As shown in Figure 1, the KBBI contains records of words, part of speech, word definitions, and examples of usage of words in the sentence [2]. However,

the majority of words in the KBBI do not provide complete definition for each of them. Taking Figure 1 as example, the first sense is indeed a fairly complete definition:

```
kayu (besi, batu, dsb) yang lebar dan tipis
(broad and thin wood (iron, rock, etc))
```

but the second sense, it is not a good definition, as only the synonym of those words are given:

```
tempat tinggal (home), rumah (house)
```

Obviously, to build a WordNet-like lexical database, the explanation of the second meaning cannot be utilized as a gloss of a word. Instead, it will be more useful for the acquisition of the synset of that word.

Papan *n*
1 kayu (besi, batu, dsb) yg lebar dan tipis (*broad and thin wood (iron, rock, etc)*)
2 tempat tinggal; rumah (*place to stay; house*)
Papan atas *n*
kelas utama; kelas tinggi; (*important class; high class*)

Fig. 1. Example of an entry in KBBI. The word 'Papan' in the Indonesian part of speech is n (noun) and this word has two different senses, each is shown through the explanation after the number 1 and 2 on this entry.

Although KBBI is available in digital format that enables information extraction, its usage in a lexical database such as WordNet that will be publicly availably must consider the ethical and usage authority aspects since the copyright of KBBI is held by the Language Center of the Department of Education and Culture of the Republic of Indonesia.

Noun
- S: (n) shrub, **bush** (a low woody perennial plant usually having several major stems)
- S: (n) **bush** (a large wilderness area)
- S: (n) scrub, chaparral, **bush** (dense vegetation consisting of stunted trees or bushes)
- S: (n) **Bush**, George Bush, George W. Bush, George Walker Bush, President Bush, President George W. Bush, Dubyuh, Dubya (43rd President of the United States; son of George Herbert Walker Bush (born in 1946))
- S: (n) **Bush**, Vannevar Bush (United States electrical engineer who designed an early analogue computer and who led the scientific program of the United States during World War II (1890-1974))
- S: (n) **Bush**, George Bush, George H.W. Bush, George H. W. Bush, George Herbert Walker Bush, President Bush (vice president under Reagan and 41st President of the US (born in 1924))
- S: (n) pubic hair, **bush**, crotch hair (hair growing in the pubic area)

Fig. 2. Results of a word search for *'bush'* in online WordNet Search 3.1. Results indicated only seven senses for part of speech *noun*. The same word search results for *verb* and *adjective* part of speech are not shown.

Besides the aforesaid, another consideration is that the dictionary, like the KBBI, may not provide a definition of a proper noun, while in fact, a lexical database like WordNet should provide such category. In Figure 2, the first section of the meaning of the word *bush* points at *part of wood, area, and hair,* whereas the other section points to proper noun for names of people: *George W. Bush, Vannevar Bush, etc.*

Even so, the goal that has been achieved through the research in this paper is acquiring a valid gloss from the available synset collection. For this reason the supervised learning has become our proposed methods to be used as a classification model which will accept or reject the glosses acquired from web.

II. Current Research

WordNet has been contributing significantly in numerous text mining, web mining, and information retrieval tasks such as query refinement, document recommender system, named-entity recognition, question-answering system, and conversational agents [5-10]. Moreover, utilization of WordNet itself is continuously becoming more widespread and is not limited to text-handling only. For example, when it is discovered that WordNet is able to improve the performance of a text classifier, it is then used to establish a life-like virtual character to express the emotions of a text [11]. When the hypernim-hyponym relationship –an *is-a* relationship– of a language can be acquired, it means that a complete ontology of a language is readily available and subsequently can be used for machine translation and semantic web [12][13]. It is impossible for a machine translation to have a perfect performance when it is not equipped with the lexical database of that language. In addition to that, definition extraction can also be used in other areas of computational linguistic, such as word sense disambiguation or word meaning representation using local context [14-16].

We use the term *gloss* rather than *definition*, on the basis that the end goal is to provide definitions for synonym sets, rather than a definition for a word. This research is a follow-up to the earlier one –which had successfully acquired the Indonesian synsets collection used as a resource in this research [1].

Next, when explanations of a proper noun needed in lexical databases such as WordNet is generated, the term *definition* would be less precise. All in all, the chosen approach does take into account the previous research on definition extraction, such that in this section, both terms can be considered as equal.

In their research, Chang and Zheng made the definition extraction in offline document that used decision tree, naive Bayes, and support vector machine with various features, such as character length, position of a sentence, number of terms –from single to bigram word– and part of speech [17]. Meanwhile, Cui et. al. acquired definition by utilizing soft pattern, generated by language model and Hidden Markov profile [18]. Both researches indicate that research methods for definition acquisition often used text patterns that sometimes cannot resolve the variations of the existing text pattern; thus, a soft pattern-matching analysis was put in place. Furthermore, more complex features were also used for the purposes of machine learning. For instance, the baseline grammar, word-class lattice, and compression-based minimum descriptive length of sentence definition that is represented in syntax tree [19-21].

To complete the utilization of machine learning methods for definition extraction, the random forest is also used here. A number of decision trees are combined to obtain a final decision for the definition determination through voting approach [22]. Although supervised learning methods from machine learning are more dominant, unsupervised learning can also be used for the definition extraction. One example is the effort to isolate patterns into *definition* and *non-definition* cluster with the patterns of n-grams, subsequence,

and dependency subtree features [23]. Borg et. al. combined the two approaches in evolutionary algorithms to obtain definition [24]. In this case, genetic programming is used for features extraction. The basic idea is actually grammar induction, that is discovering all the grammar rules that could cover the sentences containing the definitions. Meanwhile, genetic algorithms are used to rank the acquired definition candidates.

A number of studies about the extraction of text definition for non-English languages, such as Portuguese, Arabic, Polish, Slovenian, Japanese, and Chinese, have also been carried out [25-27] [22], [23].

In contrast to what was previously discussed about the purpose of extracting any word definition of a language, such extraction in a limited scope or for specific domains is also carried out, for example, for specific definitions in the field of medicine, law, and economy [28-30].

III. Indonesian Gloss Acquisition

In this section, we propose a supervised method for Indonesian gloss acquisition from Indonesian web pages. The proposed method consists of three fundamental phases: preprocessing, features extraction, and classification. These phases are illustrated in Figure 3.

It is crucial to explain about the resources in the beginning of this section since it will affect a number of patterns that can be used for the acquisition of text and features extracted from the raw text.

A. Resources

The first step in gloss acquisition is to get as many gloss sentences as possible from the collection of web pages based on the list of synsets that has been obtained [1]. There are more than 20,000 Indonesian nouns from which the gloss sentences must be acquired. Resources in the form of gloss sentences used for this purpose are derived from two sources: Wikipedia pages (both offline and online) and a collection of web pages that contain the gloss sentences, and obtained through the help of search engines like Google. We distinguish these two sources for the purpose of fetch process efficiency. It can be easily ascertained that Wikipedia pages that describe a word will contain the gloss of that word; consequently, a gloss search using search engines is only done for the list of synsets which are not yet covered by Wikipedia. The same consideration applies in the case of using offline and online Wikipedia. For Indonesian, the offline Wikipedia can be obtained periodically (once in a few years) in a few pieces of compact discs. However, these collections are yet to cover all items from the list of synset; therefore making the fetch from the online Wikipedia still necessary.

The storage of all the offline Wikipedia pages for the 2011 edition has covered 60% of more than 170,000 online pages, using the folders and subfolders structure, named using the first three letters of the word being searched. For example, to access an offline Wikipedia page for *kuda (horse)*, the address of that page is:

```
Wikipedia/articles/k/u/d/kuda.html
```

For the online Wikipedia, it can be accessed by fetching the Indonesian Wikipedia page, namely id.wikipedia.org and adding the suffix '/word '. The following is the address that should be used if the word *kuda* is not available in the offline version of Wikipedia:

```
http://id.wikipedia.org/Kuda
```

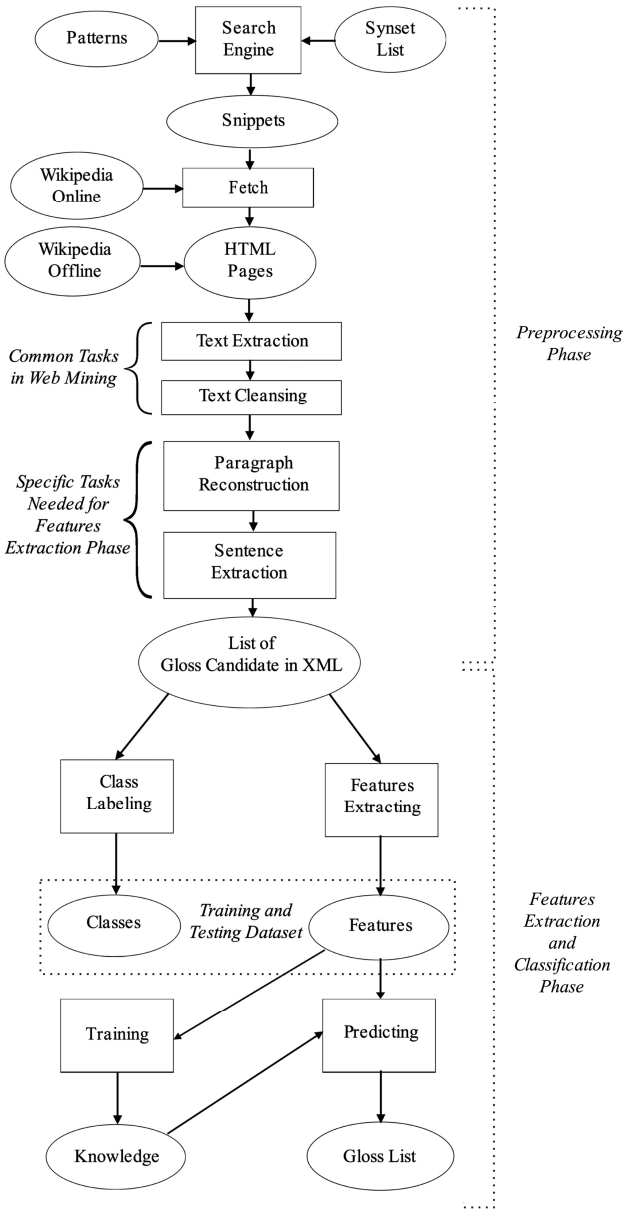

Fig. 3. Indonesian Gloss Acquisition Architecture System Overview. Generally, in this architecture there are three standard phases of pattern recognition: preprocessing, feature extraction, and classification.

Furthermore, accessing the page containing gloss sentence from search engines needs to consider the use of copula in Indonesian. Indonesian has three types of copula in a gloss sentence: *adalah*, *ialah* and *merupakan*. There are also a number of other copula, such as *yakni, yaitu* or *di mana*, which are not so commonly used in gloss sentence. In Indonesian grammar, the three copula above are actually used in different circumstances. For instance, the word *adalah* is more appropriately used to indicate "characteristically identical with" or "belongs to the group of", whereas the word *ialah* is used to confirm the details of the word described. Nevertheless, those circumstantial differences in the usage of the three copulas can be ignored

in the context of patterns preparation that will be passed on as the query string for the search engines. In Google –the search engine used in this study– to obtain the gloss from the word *q=hujan (rain)*, the following can be passed on:

```
?q = "hujan+adalah"
?q = "hujan+ialah"
?q = "hujan+merupakan"
```

Some examples of the results obtained through these patterns can be seen in Table 1.

TABLE I
SEARCH PATTERN FOR GLOSS SENTENCE ACQUISITION IN GOOGLE SEARCH ENGINE

Patterns	Example
noun+adalah	Ular adalah reptil yang tak berkaki dan bertubuh panjang.
noun+is	*Snakes are reptiles that are no-legged and long-bodied.*
noun+ialah	Republik Indonesia ialah sebuah negara kepulauan yang juga disebut sebagai Nusantara.
noun+is	*Republic of Indonesia is an archipelago country which is also known as the Nusantara.*
noun+merupakan	Hakim merupakan pejabat peradilan negara yang diberi wewenang oleh undang-undang untuk mengadili.
noun+is	*The judge is a state judicial officer authorized by law to prosecute.*

Some other parameters can also be set easily; such as the language must be Indonesian (hl = id) and the number of the desired result pages equals to n (num = n). In this research, the number of page results will help the setting of the number of pages to be fetched, i.e. from the first 30 links of the Google snippet list search results.

B. Preprocessing

The preprocessing phase begins with text cleansing to transform an HTML page into a raw text. Removal of a number of HTML tags can be done by using regular expression (regex). For example, the regex:

```
(http|ftp|https):\\/\\/[\\w\\-_]+
(\\.[\\w\\-_]+)+([\\w\\\\.,@?^=%&amp;
:/~\\+#]*[\\w\\-\\@?^=%&amp;/~\\+#])?
```

can be used to dispose a URL. With similar approach, some regex can be constructed to dispose of a wide range of HTML tags which are not required –from the email link to the recurring characters that are frequently encountered on the page that is predicted to contain the gloss.

Another example, the disposal of fullstops or repeated question marks can be done using the following regex:

```
\\.{2,}
\\?{2,}
```

The replacement of escape characters such as *&amp*, * *, or *&quot* also needs to be noted. For instance, """ will be transformed into the characters " (double quotation).

The entire process described earlier would almost certainly become the standard preprocessing for raw text extraction from web documents. Even so, the specificity of features extraction performed on the next stage demands some additional tasks (see Figure 3). To illustrate, when the sentence position or sentence phrase attribute in a paragraph is required (#1, see next section), the reconstruction of the paragraphs should also be done; lest, one problem may be encountered when a paragraph is scattered into several text lines, so that a line of a text represents neither a paragraph nor a valid sentence. The example in Figure 4 shows the reconstruction input and output of the required paragraphs.

INPUT:

```
Apel adalah jenis buah-buahan, atau buah yang
dihasilkan dari pohon buah apel. Buah apel
biasanya berwarna merah kulitnya jika masak
dan (siap dimakan), namun bisa juga kulitnya
berwarna hijau atau kuning. Kulit buahnya
agak lembek, daging buahnya keras. Buah ini
memiliki beberapa biji di dalamnya.
```

(Apple is a type of fruits, or a fruit produced from the apple tree. Apple skin is usually red when ripe and (edible), but it can also be green or yellow. The skin is rather soft, the flesh is hard. This fruit has few seeds in it.)

OUTPUT:

```
<PARAGRAPH id="1">
   <SENTENCE id="1">
    Apel adalah jenis buah-buahan, atau
    buah yang dihasilkan dari pohon buah
    apel.
   </SENTENCE>
   <SENTENCE id="2">
    Buah apel biasanya berwarna merah
    kulitnya jika masak dan (siap dimakan),
    namun bisa juga kulitnya berwarna hijau
    atau kuning.
   </SENTENCE>
   <SENTENCE id="3">
    Kulit buahnya agak lembek, daging
    buahnya keras.
   </SENTENCE>
   <SENTENCE id="4">
    Buah ini memiliki beberapa biji di
    dalamnya.
   </SENTENCE>
</PARAGRAPH>
```

Fig. 4. Paragraph Reconstruction Input and Output. XML as the output structure allows simple hierarchy presentation to show the relationship of some extracting sentences in each obtained paragraphs.

Similarly, when several attributes involving multiple values in a sentence (attr #3, attr #4, and attr #5) are required, a sentence extractor is absolutely necessary. It is also required to obtain information on a value that indicates the number of gloss sentence appearances in a document (attr #2). Take the first paragraph which is shown in Figure 2 as an example; two sentences will be extracted successfully.

At the end of the preprocessing phase, an XML file containing sentences-paragraph structure will be generated to hold the results of the preprocessing of an HTML page.

C. Features Extraction

In classification problem or supervised learning of the machine learning discipline, given examples F = {f_1, f_2, ..., f_n} from some instances and the set of class C = {c_1, c_2, ..., c_m} are given; this approach will seek to get a mapping function F → C which directs f_i into *one* of the class c_j. In other words, this approach will construct a model that can distil the knowledge in the given examples. This knowledge is subsequently used for the purposes of contents predictions of the class or target attribute of the unlabeled instance. In

general then, f_i is the *vector input*, whereas c_i or target (class) is a *single value*.

From the collection of documents stored in thousands XML files as the result of the preprocessing phase, we consider that it contains a number gloss candidates (all the texts between the tags <sentence id="n"> and </sentence>). However, not all of the gloss candidates will be accepted as good; some of them will be rejected. This applies when a synset has more than one correct gloss, only one will be accepted and the others will be rejected. For example, in the following pair of gloss:

1: Soekarno adalah presiden pertama Republik Indonesia (Sukarno was the first president of the Republic of Indonesia)

2: Soekarno adalah presiden Indonesia idolaku (Sukarno is my Indonesian president idol)

Although both gloss candidates use the same copula, it can be seen that the first sentence is a gloss of Soekarno, while the second is clearly not. The example above also shows that an Indonesian sentence that has a copula is not necessarily a gloss sentence, just like that in English.

Likewise in different instances for the following pair of candidates:

1: editor adalah orang di balik layar (editor is the person behind the screen)

2: editor istilah lain penyunting (editor, another term for *penyunting*)

3: editor adalah orang yang ditugasi untuk melakukan pengeditan atau penyuntingan suatu naskah (editor is the person assigned to edit a manuscript)

Although all three of them are the correct glosses of the word *editor*, we had to choose the 3rd candidate because of the completeness of its gloss; whereas the 1st and 2nd gloss must be rejected.

After the labeling for class C = {c_1, c_2} –where c1=accept and c2=reject– from the same XML collections, it is a must to extract F = {f_1, f_2, ..., f_n} for each gloss candidate. In this research, seven pieces of features will be extracted, all of which will contain positive integer values. Following is an explanation of each feature used.

Position of a sentence in the paragraph (POSINPAR): An attribute that states the position of a sentence or phrase in a paragraph, or in which line is the gloss candidate located in a paragraph. Our hypothesis is that a gloss sentence tends to be present in the beginning of the paragraph, such as the first or the second sentence after the title.

Frequency of sentence in a collection of documents (DOCFREQ): When there is more than one exactly identical gloss sentence in a collection of documents, only one will be taken as a candidate, although the number of its appearances will be recorded. This attribute states how many times the gloss sentence candidate appears in a collection of documents. This approach is the negation of the *idf (inverse document frequency)* in information retrieval.

Number of words in a sentence (NUMWORD): An attribute that states the number of words in the gloss sentence candidates. In cases when there are more than one candidate of the same word and there exists a gloss of a candidate which is described more completely and contains more words, this particular gloss will usually be selected.

Number of important words in a sentence (NUMIMPT): Logically, it makes sense that a better gloss should be linked with a number of other terms. Those linkages with other terms will have a positive strong correlation with the number of important words, which is represented in this research by the number of words that are not *stop words.*

Number of characters in a sentence (NUMCHAR): It is similar to NUMWORD, the number of words in a sentence, but it is calculated in units of characters composing the gloss sentence candidates.

Number of gloss sentences from the same word (CANDFREQ): This attribute contains the number of resulting gloss sentence candidates for the same word. In contrast to DOCFREQ –which calculates the number of documents, this attribute is more similar to *tf (term frequency)* in information retrieval area. It should be noted that a single document from the collection of available documents can contain more than one identical gloss sentence of the same word.

Number of nouns in the sentence (NUMNOUN): This attribute represents the appearance of nouns that is defined in the gloss sentence candidate. In this case, the synset –with part of speech of noun– that has been obtained through previous research [1] is used to obtain this feature.

```
<CANDIDATE id="1">
   <DEFINITION>
    udang-udang kecil yang sudah dikeringkan
   </DEFINITION>
   <PATTERNNO>3</PATTERNNO>
   <POSINPAR>1</POSINPAR>
   <DOCFREQ>1</DOCFREQ>
   <NUMWORD>5</NUMWORD>
   <NUMIMPT>2</NUMIMPT>
   <NUMCHAR>37</NUMCHAR>
   <CANDFREQ>1</CANDFREQ>
   <NUMNOUN>2</NUMNOUN>
</CANDIDATE>
```

Fig. 5. An example of an XML file that contains all of the features and its target class (accept/reject) of the word *'ebi ' (dried shrimp)*. In this example the gloss *ebi* accepted as the correct candidate.

Unlike manual labeling given to the binary class, regardless of whether a sentence candidate is accepted as true gloss, the acquisition of these seven features is certainly done with simple algorithms for the efficiency of the training set formation. Thus, in this section we have managed to establish some *examples* which are used later on as a training set of binary classification modeling. Figure 5 shows an example of an XML file that represents an instance used for the classification process.

D. Classification

Backpropagation feedforward neural networks (BPFFNN) and decision tree are selected as the model for classification. BPFFNN was chosen because, at this moment, along with Support Vector Machine, they can be considered as *the state of the art* of classification methods for non-linear separable problems. The architecture used in BPFFNN can be seen in Figure 6, which is a multilayer architecture with seven input nodes, a number of hidden nodes, and an output node that generates the accept or reject response.

We use single-layer architecture considering that the expected accuracy performance can be obtained through it. The seven features with positive integer values can be easily normalized into the bipolar range (-1 ...+ 1) by the formula:

$$\text{newI} = (\text{oldI} - \text{oldMin})\frac{\text{newMax} - \text{newMin}}{\text{oldMax} - \text{oldMin}} + \text{newMin} \quad (1)$$

before they are passed on into the input nodes. As for the output, the value +1 is used to represent accept and -1 to represent reject. Other properties like momentum and weight initialization using Nguyen-Widrow are also utilized to accelerate the training time. In this study we use the total square error <= 0.2%, the momentum = 0.2, the learning rate = 0.3, and the maximum number of epoch = 500.

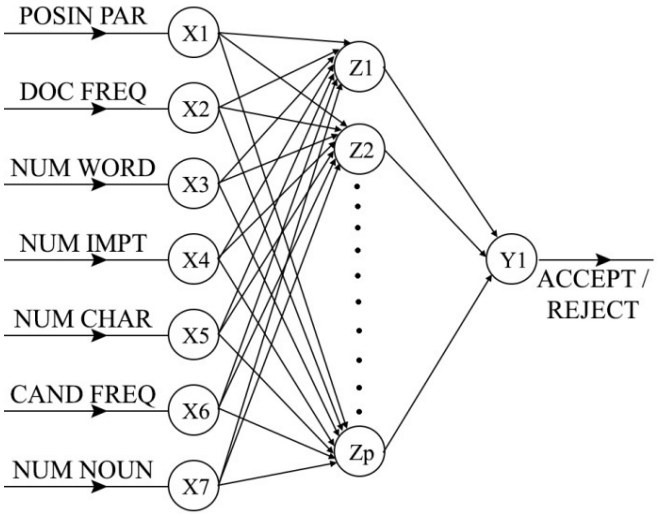

Fig. 6. BPFFNN Multilayer architecture for Gloss Candidate Classification. In this 7-P-1 architecture, the number of input nodes represents each attribute, whereas an output node is used for deciding the accept/reject prediction.

On the induction using the decision tree, since all of the features are of continuous type with positive integer, hence, the entire branch nodes in the decision tree will be a binary split, –each of which has a label <=n and >n (see Figure 7). In the calculation of gain, Δ, for the selection of the best split attribute, we use the formula:

$$\Delta = \text{I}(\text{parent}) - \sum_{j=1}^{k} \frac{\text{N}(v_j)}{\text{N}} \text{I}(v_j) \quad (2)$$

wherein I(t) is the impurity measure which is calculated through Gini with the formula:

$$\text{Gini}(t) = 1 - \sum_{i=0}^{c-1} [\,p(i \mid t)\,]^2 \quad (3)$$

Moving on, decision tree was chosen because of the visible nature of its knowledge. In Figure 7, it can be shown that NUMCHAR (the number of characters in a sentence) is a feature that is most instrumental in making accept or reject decision in the decision tree. Similarly, the seven features used only have significant roles in the first seven levels of the decision tree.

IV. Experimental Results and Discussion

In this section we are describing the results of the experiment that has been done; starting from the fetch results of the available Indonesian web resources, the preprocessing to obtain gloss sentence candidates, to the acquisition of the acceptable gloss through classification process.

There were no significant problems in the features extraction phase because the use of several simple algorithms provides a guarantee of a successful acquisition of the seven features from the gloss sentence candidates.

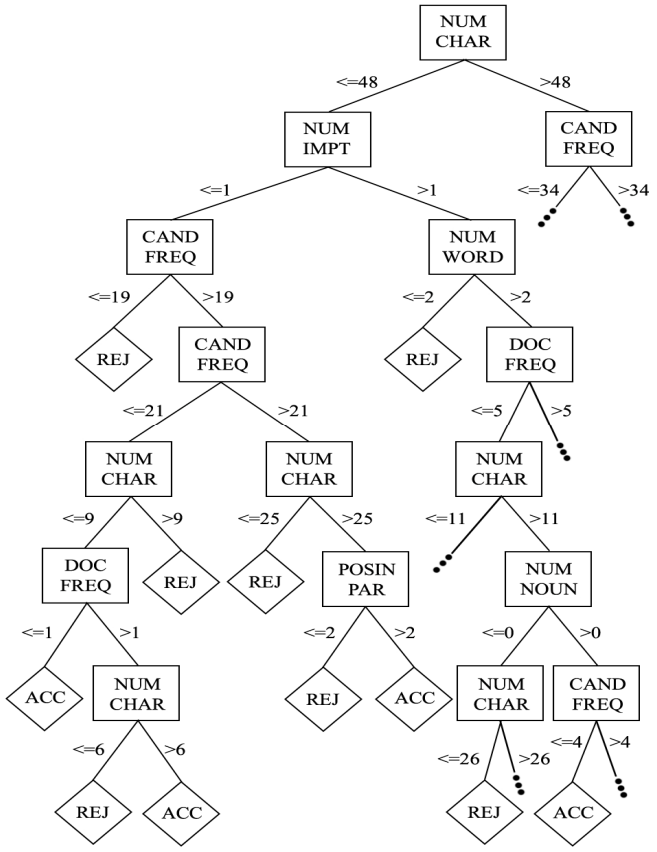

Fig. 7. Resulting Binary Split Decision Tree from the induction process. We just show several first levels to identify some attributes which have most significant effects.

The results of the experiments have been measured from two different perspectives, namely quantitatively and qualitatively. Quantitative performance is mainly used to measure how many of the gloss candidates and the accepted gloss were successfully retrieved. Table II summarizes the results of the experiments which have been carried out quantitatively on all phases. On the other hand, the qualitative performance was used to measure the accuracy and error rate as it is frequently used in the evaluation of the classification method.

We divide the entire 21,201 nouns into four groups based on initial letters (A-E, F-J, K-O, and U-Z). For the efficiency of the process, the group P-T is not addressed in this research, considering the relatively large collection of Indonesian nouns in this group, except for the letter Q which only has 17 Indonesian nouns. Nevertheless, the fetch was still done for other large collection of words, such as for nouns that start with the letters A-E (8,684 words) and K-O (6,606 words).

TABLE II
QUANTITATIVE PERFORMANCE IN WEB RESOURCES ACQUISITION, PREPROCESSING, AND CLASSIFICATION

Letter Group	# Words	Texts Acquired		Preprocessing		Gloss Candidate		Classification	
		# Fetched files	# Words without documents found	# Files		# Words (candidates found)		# Words (glosses found)	
A-E	8,684	1,448,646	2,872	760,238	52.48%	3,009	34.65%	2,161	71.82%
F-J	4,578	714,716	1,753	389,534	54.50%	1,759	38.42%	1,559	88.63%
K-O	6,606	1,495,577	726	837,524	56.00%	2,844	43.05%	2,176	76.51%
U-Z	1,333	313,317	239	169,518	54.10%	676	50.71%	624	92.31%
∑	**21,201**	**3,972,256**	**5,590**	**2,156,814**	**54.30%**	**8,288**	**39.09%**	**6,520**	**78.67%**

A. Web Resources Acquisition and Preprocessing

From the 21,201 nouns which the gloss is expected to be found, 8,288 (39.09%) gloss sentence candidates were obtained. In general, it is indeed reasonable that web resources do not provide as much gloss as what is obtainable from a list of words in the monolingual lexical resources in Indonesian, namely the KBBI and the Indonesian Thesaurus. We can see in table II that there are no web documents which provide gloss through the copula patterns for 5,590 (26.36%) words.

We identified two causes when we investigated this case. First, the web resources indeed never provides the gloss of a number of single words that are rarely used in everyday Indonesian, such as *eltor (name of stomach disease, a type of cholera), elung (arc), nahu (word syntax),* or *najam (star)*. Second, not all compound words, which are a narrow meaning of a single word, have its gloss available on the web. To illustrate, there are over thirty Indonesian compound words beginning with the word *emas (gold)*, see Figure 8.

[1]**emas** *n* **1** logam mulia berwarna kuning yg dapat ditempa dan dibentuk, biasa dibuat perhiasan spt cincin, kalung (lambangnya Au, nomor atomnya 79, bobot atomnya 196,9665); logam adi; aurum: *ia memakai cincin --bermata mutiara*; **2** *ki* uang; harta duniawi; **3** *ki* uang suap;
⋮
permukaan; **-- biru** *ki* semen; **-- cokelat** *ki cengkih*; **-- hidrotermal** emas yg digali dan ditambang; **-- hijau** *ki* komoditas hasil pertanian yang potensial untuk diekspor; **-- hitam** *ki* **1** pasir besi; **2** aspal; **3** batu bara; **-- juita** *ki*
⋮
-- tua emas yg tinggi mutunya (tinggi kadar emasnya), 24 atau 23 karat; **-- tulen** emas murni; **-- urai** emas yg berbutir-butir halus; **-- urung** emas yg tidak ada harganya;

Fig. 8. Some examples of compound word in KBBI (Indonesian Dictionary) beginning with the word *'emas'* (gold). The picture only shows a portion of 40 compound words beginning with *emas*.

Table III shows some examples of the circumstance where the gloss candidates for a single word could be obtained, but the glosses of some compound words beginning with same word could not be found through search engines. The first three examples in Table III show the compound words which is still having a relationship with or is narrowing the meaning of its single word. The last three examples, however, show cases where the compound words have absolutely no relationship with its single word meaning.

TABLE III
SOME EXAMPLES OF INDONESIAN COMPOUND WORDS WHOSE GLOSS CANDIDATES ARE NOT FOUND THROUGH SEARCH ENGINES

Single Word ***Gloss Candidates were Obtained***		**Compound Word** ***Gloss Candidates were not Obtained***	
edisi	*edition*	edisi revisi	*revised edition*
emas	*gold*	emas pukal	*lump of gold*
negara	*state, country*	negara teluk	*gulf states*
nangka	*jackfruit*	nangka belanda	*soursop*
ekstra	*extra*	ekstra hati-hati	*extremely careful*
nada	*tone*	nada-nadanya	*seemingly*

We were also able to explain another secondary cause, such as the capability of the fetching of web document files, which is only 54.30%. Several other secondary causes that we could identify were the size limitation of the file that can be fetched, corrupt files, broken links, and empty response because of non-standard HTML files like AJAX and document image files. Besides that, we need to consider the limitations on the performance of the information extraction from the web pages as well. For example, the use of <p> tag and header tags such as <h3>, whereas gloss sentences contained in the table tags (<table>, <tr>, and <td>) are ignored.

B. Supervised Methods Training and Testing

The measure of success of a classification is usually done qualitatively by testing fresh data with previously known label that are not used as training sets.

TABLE IV
CONFUSION MATRIX FOR BINARY CLASSIFICATION PROBLEM

Confusion Matrix		Predicted Class	
		Accept	Reject
Actual Class	Accept	f_{11}	f_{10}
	Reject	f_{01}	f_{00}

Confusion matrix for binary classification problem whose cells are shown in table IV can be used to assist in the calculation of accuracy and error rate with the following formula:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} = \frac{f_{11} + f_{00}}{f_{11} + f_{10} + f_{01} + f_{00}} \quad (4)$$

TABLE V
COMPARISON OF ACCURACY LEVEL AND ERROR OF DECISION TREE AND BACKPROPAGATION FEED FORWARD NEURAL NETWORK

Ratio Training:Testing	Training Set	Testing Set	Decision Tree		Backpropagation NN	
			Accuracy	Error	Accuracy	Error
80:20	6,249	1,562	76.06%	23.94%	80.03%	19.97%
70:30	5,468	2,343	72.17%	27.83%	74.26%	25.74%
60:40	4,687	3,124	73.94%	26.06%	71.90%	28.10%
		Average	74.06%	25.94%	75.40%	24.60%

TABLE VI
SOME EXAMPLE OF ACCEPTED AND REJECTED GLOSS CANDIDATES

Word	Accepted Glosses	Rejected Glosses
batik	Seni gambar diatas kain untuk pakaian yang dibuat dengan teknik resist menggunakan material lilin. *An art of drawing on the fabric for clothing that is made with materials using wax resist technique.*	Salah satu cara pembuatan bahan pakaian. *One way of making clothing materials.*
editor	Orang yang ditugasi untuk melakukan pengeditan atau penyuntingan suatu naskah. *Man tasked with editing or proofreading a manuscript.*	Orang di balik layar. *Person behind the scenes.*
detektif (*detective*)	Seseorang yang melakukan penyelidikan dan memecahkan suatu masalah (biasanya kasus kriminal) yang belum terungkap, menggunakan metode sistematis dan terencana berdasarkan bukti-bukti yang ada, dan merangkainya menjadi suatu fakta yang bisa dipertanggungjawabkan. *Someone who investigate and solve a problem (typically criminal cases) that has not been revealed, using a systematic and planned method based on available evidences, and crafted it into a fact which can be accounted for.*	Seseorang yang melakukan penyelidikan terhadap suatu kejahatan, baik sebagai detektif polisi maupun sebagai detektif swasta. *Someone who did the investigation of a crime, either as a police detective or a private investigator.*
jumat (*friday*)	Hari keenam dalam satu pekan. *The sixth day in a week.*	Hari terakhir pendaftaran pencalonan. *Last day of candidacy registration.*
marathon (*marathon*)	Nomor lari dalam cabang atletik yang menempuh jarak 42,195 kilometer (26 mil dan 385 yard). *A number in athletics branch, covering a distance of 42.195 kilometers (26 miles and 385 yards).*	Salah satu cabang olahraga atletik yang amat terkenal. *One branch in athletics that is very famous.*
yard	Satuan dasar untuk ukuran panjang dalam unit imperial. *Base unit for length in imperial units.*	Disingkat: yd *Abbreviated: yd*

$$\text{Error Rate} = \frac{\text{Number of wrong predictions}}{\text{Total number of predictions}} = \frac{f_{10} + f_{01}}{f_{11} + f_{10} + f_{01} + f_{00}} \quad (5)$$

On the calculation of accuracy and error rate, formulations such as the 80:20 ratio are normally used to show a comparison of the percentage (training set : set of testing). In this research, each was done with the ratio of 80:20, 70:30, and 60:40.

The training set used was the entire gloss sentence candidates that begin with the letter E and N. In addition to the large number of Indonesian words beginning with both letters, in our research, all of the glosses that were successfully obtained for words beginning with the letters E and N have been successfully acquired. The total number of the training sets, as shown in table V which reaches 6,811 training sets, consists of 3,899 candidates starting with the letter E and 3,012 candidates starting with the letter N. The entirety of the data has been previously labeled manually and will be used as training sets to calculate the accuracy of both the methods used, namely the decision tree and backpropagation feed forward neural networks.

From the experiments, it appears that the accuracy achievements for both the methods used are equivalent. Similarly, the average accuracy of the three comparisons for the number of different training and testing sets obtained, were up to 74.06% for decision tree and 75.40% for backpropagation feedforward neural networks (as shown in Table V).

Quantitatively, out of 8,288 words, we were able to obtain 78.67% or 6,520 nouns with gloss candidates through classification approach. Overall, from the 21,218 nouns targeted in this research, the gloss of 6,520 nouns was successfully obtained; this is represented as 30.76%.

Finally, from the experiment conducted, table VI shows a number of examples of the gloss candidates in Indonesian that were accepted or rejected by the proposed methods. It is depicted in this table that only the best glosses are accepted, whereas the rest which are not sufficiently good or those which are not actually a gloss are rejected.

When a word has several meanings (multi-sense), the relationship between each of these available meaning will determine whether the collection of meaning is a homonym or polysemy. Attempts to identify whether the collection of meaning of a word is a homonym or polysemy are actually quite hard to do, or that there is no clearly-defined limit to distinguish them [28]. In polysemy, each meaning actually still has a relationship (*sememes*) in the same semantic domain, whereas for homonyms, there is no relationship at all for every meaning available. Table VII and VIII show that our method can also accommodate both forms of these multi-sense words. In the case of multi-sense as shown in these tables, all the gloss has been accepted as the correct gloss.

TABLE VII
SOME EXAMPLES OF ACCEPTED GLOSS FOR POLYSEMY WORDS

Polysemy Word	Accepted Glosses	
Bendera (*flag*)	[1] Salah sebuah simbol dari kebudayaan yang dimiliki oleh suatu kelompok suku bangsa atau negara.	[1] *One symbol of a culture that is owned by an ethnic group or nation.*
	[2] (Kata serapan dari bahasa Portugis), *bandeira* adalah sepotong kain yang dipakai sebagai wahana untuk menggambarkan simbol sebuah negara.	[2] *(Derived word from Portuguese), bandeira is a piece of fabric that is used as a medium to describe a country's symbol.*
Dinasti (*dynasty*)	[1] Sistem reproduksi kekuasaan yang primitif karena mengandalkan darah dan keturunan dari segelintir orang.	[1] *A primitive reproductive system as it relies on blood relation and descendants of a handful of people.*
	[2] Keluarga raja yang memerintah suatu wilayah secara turun-temurun.	[2] *The royal family which ruled a region for generations.*
depot	[1] Kata umum yang dipakai warga Jawa Timur pada umumnya untuk menyebut warung makan.	[1] *Common words used by East Java residents in general to refer to food stalls.*
	[2] Tempat kegiatan penerimaan, penimbunan dan penyaluran kembali bahan bakar minyak (BBM).	[2] *Place for activities of receiving, stockpiling and redistribution fuel (BBM).*

TABLE VIII
SOME EXAMPLES OF ACCEPTED GLOSS FOR HOMONYM WORDS

Homonym Word	Accepted Glosses	
darah tinggi (*hypertension*)	[1] Penyebab utama penyakit serangan jantung (heart attack) lemah jantung (heart failure), kegagalan buah pinggang dan buta.	[1] *The main cause of coronary heart disease (heart attack) weak heart (heart failure), kidney failure and blindness.*
	[2] Film Indonesia yang dirilis pada tahun 1960 yang disutradarai oleh Lilik Sudjio.	[2] *An Indonesian film released in year 1960 which is directed by Lilik Sudjio.*
Debu (*dust*)	[1] Nama umum untuk sejumlah partikel padat kecil dengan diameter kurang dari 500 mikrometer (lihat juga pasir atau granulat).	[1] *The common name for a number of small solid particles with a diameter of less than 500 micrometers (see also sand or granules).*
	[2] Kelompok musik muslim yang pertama kali tampil tahun 2001 dan sekarang berkediaman di Jakarta, Indonesia.	[2] *Moslem music group who first appeared in 2001 and now resides in Jakarta, Indonesia.*
gladiator	[1] Sebutan para petarung yang di adu di arena ini pada masa lalu.	[1] *The title of the fighters who compete in this arena in the past.*
	[2] Sebuah film sejarah yang diluncurkan tahun pada 2000 dan diarahkan oleh Ridley Scott dan dibintangi oleh Russell Crowe, Joaquin Phoenix, Connie Nielsen, Djimon Hounsou dan Richard Harris.	[2] *A historical movie that was launched in the year 2000 and directed by Ridley Scott and starring Russell Crowe, Joaquin Phoenix, Connie Nielsen, Djimon Hounsou, and Richard Harris.*

V. CONCLUSION

This paper demonstrates supervised learning based method that acquires the gloss of the Indonesian synset collection from web pages. The utilization of copula as a simple pattern is capable of acquiring 6,520 gloss (30.76%) of the 21,201 Indonesian noun synsets.

The seven features used (the position of the sentence in the paragraph, the frequency of sentences in a document collection, the number of words in the sentence, the number of important words in the sentence, the number of characters in the sentence, number of gloss sentences of the same word, and the number of nouns in the sentence) have a significant role in the classification of gloss sentence candidates.

The decision making process to accept or reject a gloss sentence candidate through supervised learning is able to achieve an accuracy of up to 74.06% for decision tree and 75.40% for backpropagation feedforward neural networks.

The results for gloss sentences obtained through this research can complement the standard gloss that has been provided on a monolingual resource for the purpose of constructing a lexical database. More specifically, the methods offered are able to accommodate for some forms of multi-sense words: polysemy, homonym, and expansion of the sense of generic words used as proper nouns (movie titles, names of famous persons, names of music groups, etc.)

In conclusion, the supervised learning approach used in this research has been successfully acquired a vast amount of gloss with relatively high accuracy which is quite significant for Indonesian words ranging from A to Z. Thus, we believe that the method we offer can also be applied for the purposes of the acquisition of gloss on other natural world languages other than Indonesian.

ACKNOWLEDGMENT

The authors thank the reviewers for their valuable comments to improve on this paper. We also thank the members of Computational Linguistic Research Center Sekolah Tinggi Teknik Surabaya, especially for their contributions in manually labeling the accepted glosses.

REFERENCES

[1] Gunawan and A. Saputra, "Building Synsets for Indonesian WordNet with Monolingual Lexical Resources," in *Proceedings of the International Conference on Asian Language Processing (IALP)*, Harbin, China, December 2010, pp. 297-300.

[2] Gunawan and E. Pranata, "Acquisition of Hypernymy-Hyponymy Relation between Nouns for WordNet Building," in *Proceedings of the International Conference on Asian Language Processing (IALP)*, Harbin, China, December 2010, pp. 114-117.

[3] Gunawan, J.F. Wijoyo, I. K. E. Purnama, and M. Hariadi, "WordNet Editor to Refine Indonesian Language Lexical Database," in *Proceedings of the International Conference on Asian Language Processing (IALP)*, Penang, Malaysia, November 2011, pp. 47-50.
[4] *Indonesian Dictionary (Kamus Besar Bahasa Indonesia)*. Jakarta: Balai Pustaka, 2008.
[5] J. Nemrava, "Using WordNet Glosses to Refine Google Queries," in *Proceedings of the Dateso 2006 Annual International Workshop on Databases, Texts, Specifications and Objects*, Desna, Czech Republic, April 2006, pp. 85–94.
[6] B. Magnini and C. Strapparava, "Using WordNet to Improve User Modelling in a Web Document Recommender System," in *Proceedings of the North American Chapter of the Association for Computational Linguistics Workshop on WordNet and Other Lexical Resources*, Pittsburgh, June 2001, pp. 132-137.
[7] B. Magnini, M. Negri, R. Prevete, and H. Tanev, "A WordNet-based Approach to Named-Entites Recognition," in *Proceedings of the 2002 Workshop on Building and Using Semantic Networks*, Taipei, 2002.
[8] A. Y. K. Chua and S. Banerjee, "A Comparison of Quality, Speed, Scope and Usability between English and Chinese CQAs," *IAENG International Journal of Computer Science*, vol. 40, no. 2, pp. 110-116, 2013.
[9] M. Ramprasath, S. Hariharan, "A Survey on Question Answering System," in International Journal of Research and Reviews in Information Sciences (IJRRIS), vol. 2, no. 1, March 2012, pp. 171-178.
[10] O. S. Goh and C. C. Fung, "Trust Based Knowledge Acquisition for Conversation Agents," *IAENG International Journal of Computer Science*, vol. 33, no. 2, pp. 43-51, 2008.
[11] S. Sumpeno, M. Hariadi, and M. H. Purnomo, "Facial Emotional Expressions of Life-like Character Based on Text Classifier and Fuzzy Logic," *IAENG International Journal of Computer Science*, vol. 38, no. 2, pp. 122-133, 2011.
[12] R. Navigli, P. Velardi, and A. Gangemi, "Ontology Learning and Its Application to Automated Terminology Translation," *IEEE Intelligent Systems*, vol. 18, no. 1, January-February 2003, pp. 22-31.
[13] E. Daehnhardt and Y. Jing, "An Approach to Software Selection Using Semantic Web," *IAENG International Journal of Computer Science*, vol. 40, no. 4, pp. 238-249, 2013.
[14] R. Mihalcea and D. I. Moldovan, "An Iterative Approach to Word Sense Disambiguation," in *Proceedings of the 13th International Florida Artificial Intelligence Research Society Conference (FLAIRS)*, Orlando, Florida, USA, May 2000, pp. 219-223.
[15] W. Duan and A. Yates, "Extracting Glosses to Disambiguate Word Senses," in *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Los Angeles, California, June 2010, pp. 627–635.
[16] D. Huynh, D. Tran, and W. Ma, "Contextual Analysis for the Representation of Words," *IAENG International Journal of Computer Science*, vol. 41, no. 2, pp. 148-152, 2014.
[17] X. Chang and Q. Zheng, "Offline Definition Extraction Using Machine Learning for Knowledge-Oriented Question Answering," in *Proceedings of 3th International Conference on Intelligent Computing*, Qingdao, China, August 2007, pp. 1286-1294.
[18] H. Cui, M. Y. Kan, and T. S. Chua, "Soft Pattern Matching Models for Definitional Question Answering," *ACM Transactions on Information Systems*, vol. 25, no. 2, article 8, 2007.
[19] Ł. Degórski, M. Marcinczuk, and A. Przepiórkowski, "Definition Extraction Using a Sequential Combination of Baseline Grammars and Machine Learning Classifiers," in *International Conference on Language, Resources and Evaluation*, 2008.
[20] R. Navigli and P. Velardi, "Learning Word-Class Lattices for Definition and Hypernym Extraction," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, July 2010, pp. 1318-1327.
[21] M. Tsuchiya, S. Kurohashi, and S. Sato, "Discovery of Definition Pattern by Compressing Dictionary Sentences," in *Progress in Discovery Science*, ser. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2002, vol. 2281, pp. 284-295.
[22] Ł. Kobyliński and A. Przepiórkowski, "Definition Extraction with Balanced Random Forests," in *Lecture Notes in Computer Science: Proceedings of the 6th International Conference on Natural Language Processing, GoTAL 2008*, Gothenburg, Sweden, 2008, pp. 237–247.
[23] Y. Yan, C. Hashimoto, and K. Torisawa, "Pattern Mining Approach to Unsupervised Definition Extraction," 2012.
[24] C. Borg, M. Rosner, and G. Pace, "Evolutionary Algorithms for Definition Extraction," in *Proceedings of the 1st Workshop on Definition Extraction*, Borovets, Bulgaria, September 2009, pp. 26–32.
[25] H. G. Oliveira and P. Gomes, "Automatic Discovery of Fuzzy Synsets from Dictionary Definitions," in *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, Barcelona, Spain, July 2011, pp. 1801-1806.
[26] O. Trigui, L. H. Belguith, and P. Rosso, "An Automatic Definition Extraction in Arabic Language," in *Lecture Notes in Computer Science: Proceedings of the 15th International Conference on Applications of Natural Language to Information Systems, NLDB 2010*, Cardiff, United Kingdom, June 2010, pp. 240-247.
[27] D. Fišer, S.Pollak, and Š. Vintar, "Learning to Mine Definitions from Slovene Structured and Unstructured Knowledge-Rich Resources," in *Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010, pp. 2932-2936.
[28] J. L. Klavans and S. Muresan, "Evaluation of DEFINDER: A System to Mine Definitions from Consumer-oriented Medical Text," in *Proceedings of the 1st ACM/IEEE Joint Conference on Digital Libraries (JCDL'01)*, Virginia, USA, 2001, pp. 201-202.
[29] M. Curtotti, E. McCreath, and S. Sridharan, "Software Tools for the Visualization of Definition Networks in Legal Contracts," in *Proceedings of the 14th International Conference on Artificial Intelligence and Law (ICAIL'13)*, June 2013, pp. 192-196.
[30] P. Velardi, R. Navigli, and P. D'Amadio, "Mining the Web to Create Specialized Glossaries," *IEEE Intelligent System*s, vol. 23, no. 5, September-October 2008, pp. 18-25.
[31] T. McArthur, *Concise Oxford Companion to the English Language*. Oxford University Press, 2005.

Gunawan received both diploma and bachelor degrees in Computer Science from Sekolah Tinggi Teknik Surabaya, Surabaya, East Java, Indonesia in 1991. He received his Master of Technology from Sekolah Tinggi Teknik Surabaya, Surabaya, Indonesia in 2002. He joined Computer Science Department as a lecturer in Sekolah Tinggi Teknik Surabaya, Surabaya, Indonesia since 1992. His research interest includes Computational Linguistic, Data and Web Mining, and Information Retrieval. He is working toward the Ph.D. degree at Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia since 2008. He is a member of IAENG.

I Ketut Eddy Purnama received the bachelor degree in Electrical Engineering from Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia in 1994. He received his Master of Technology from Institut Teknologi Bandung, Bandung, Indonesia in 1999. He received Ph.D degree from University of Groningen, the Netherlands in 2007. Currently, he is a staff of Electrical Engineering Deparment of Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia. His research interest is in Data Mining, Medical Image Processing and Intelligent System.

Mochamad Hariadi received the B.E. degree in Electrical Engineering Department of Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia, in 1995. He received both M.Sc. and Ph. D. degrees in Graduate School of Information Science Tohoku University Japan, in 2003 and 2006 respectively. Currently, he is a staff of Electrical Engineering Deparment of Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia. He is the project leader in joint research with PREDICT JICA project Japan and WINDS project Japan. His research interest is in Video and Image Processing, Data Mining and Intelligent System. He is a member of IEEE, IEICE, and IAENG.