International Journal of Multimedia and Ubiquitous Engineering
Vol.11, No.8 (2016), pp.13-28
http://dx.doi.org/10.14257/ijmue.2016.11.8.02

Semi-Automatic Indonesian WordNet Establishment: From Synset Extraction to Visual Editor

Gunawan[1,2], I Ketut Eddy Purnama[2] and Mochamad Hariadi[2]

[1]*Computer Science Department, Sekolah Tinggi Teknik Surabaya, Surabaya, East Java, Indonesia*
[2]*Electrical Engineering Department, Institut Teknologi Sepuluh November, Surabaya, East Java, Indonesia*
gunawan@stts.edu, ketut@ee.its.ac.id, mochar@ee.its.ac.id

Abstract

In this study, we have developed an Indonesian WordNet through four main phases: synonym set extraction (synset) as the smallest entity of lexical database from a natural language, semantic relation establishment between synsets (hypernym-hyponym and holonym-meronym), gloss extraction for synset collection, and the visual editor creation. The Semi-automatic term refers to the three initial phases which are automatically done using a number of machine learning approaches, while using visual editor to collaboratively complement the results collected from the previous phases. A number of raw data used on synset acquisition, semantic relations and glosses come from Kamus Besar Bahasa Indonesia (Great Dictionary of the Indonesian Language, abbreviated as KBBI) and Tesaurus Bahasa Indonesia (Indonesian Language Thesaurus), large collection of web pages from search engines, Wikipedia, and even Princeton WordNet for mapping purpose. This study shows that the proposed system successfully achieve 37,485 synsets, 24,256 hypernym-hyponym relations, 11,044 holonym-meronym relations and 6,520 gloss synsets. Similar approach is believed to accelerate lexical database development like WordNet for other languages.

Keywords*: WordNet, Indonesian language, synonym set, semantic relation, gloss, visual editor*

1. Introduction and Background

WordNet is a lexical reference system that contains the information, class and definition from all words collection of a language. These three are integrated into a single entity and each entity will relate to one another. The smallest unit in WordNet itself is not a word, but a synonym set (synset). For example, the latest English version WordNet of Princeton University has 117.659 synsets. There are two types of relations in WordNet: semantic relations and lexical relations. Semantic relations occur between synsets like hypernym-hyponym and holonym-meronym relations, whereas lexical relations occur between lexical units in a language like antonym relations.

Lexical reference system for a natural language is absolutely necessary for the advancement of research in the disciplines of natural language processing / computational linguistics, information retrieval, and web mining of that language. Since it was first developed by a number of psycholinguist at Princeton University in 1985, WordNet has helped significantly on several tasks from question-answering system [1] to machine translation [2]. In addition, when all semantic relations and lexical relations are available in a language, then a complete ontology is available for that language [2].

Research on the development of Indonesian WordNet had started since 2008. Using the approach on mapping by Princeton WordNet (PWN), it generated less than 5000 synsets [3]. One of the steps taken is the use of Latent Semantic Analysis for mapping English

ISSN: 1975-0080 IJMUE

words into Indonesian words, using English-Indonesian parallel corpus [4]. The acquisition of Indonesian synsets was also demonstrated by using synset assignment.

In the evaluation of this study, the results of synset assignment for bilingual dictionaries, such as Thai-English, Indonesian-English, and Mongolian-English were compared. Although it was shown that the obtained synsets have already consisted of several different classes (noun, verb, adjective, and adverb), the number of synsets obtained through these study is still small, thet is less than 1000 synsets [5].

The large number of synsets obtained in Indonesian has led to a collaboration between Indonesian WordNet and Asian WordNet (AWN) [6]. Other efforts on creating a single WordNet for two languages at once, namely Malay and Indonesian, were done by combining information from several lexical resources [7]. However, there has been no progress at all on AWN research and single WordNet Malay-Indonesian in the last four years.

We emphasized that in all of these studies, there were no information on the number and the extent of gloss synset semantic relations which were successfully acquired. Thus, efforts are limited to extraction of Indonesian synsets and the obtained synsets are currently unavailable for public access.

With more than 55 million users and being the 8th most internet users in the world, naturally the number of researches in the disciplines of Natural Language Processing or Computational Linguistics, Information Retrieval, Text and Web Mining via Indonesian Language should be proportional to the number of internet users. Nevertheless, in reality, this is not the case. One of the main causes is the unavailability of Indonesian lexical database that can be used freely. Hence, we believe that with the establishment of Indonesian lexical database that can be used publicly, the door for researches and application developments in a variety of disciplines that use Indonesian will be opened.

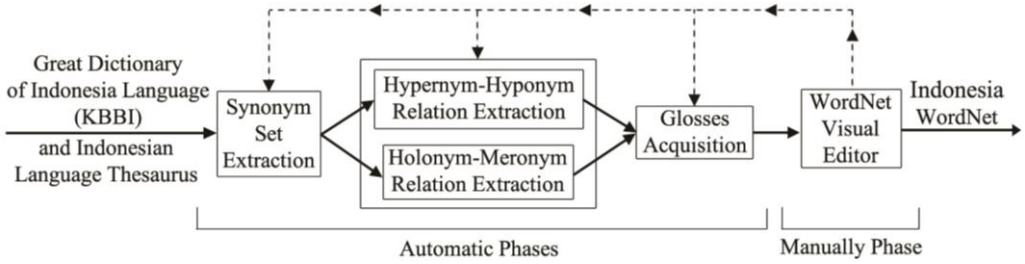


Figure 1. Semi-automatic Indonesian WordNet Architecture

In the past few years, we have been constructing an Indonesian WordNet where the initialization is obtained from synonym sets (synset) of Indonesian nouns using *Kamus Besar Bahasa Indonesia (KBBI)* as monolingual lexical resources [8]. Next, we attempted to obtain the semantic relation of hypernym-hyponym [9] and holonym-meronym between noun synsets. After all synsets and relationships are obtained, gloss will be extracted from all of the synsets [10]. The last, is an attempt to provide a web-based graphical tool for the purpose of improving Indonesian lexical database collaboratively [11]. Thus, as shown in Figure 1, the number of stages in our proposed method in this paper is presented according to that task sequence.

2. Current Research

The focus of our research in the development of Indonesian WordNet consists of four phases; those are synset extraction, semantic relations establishment between synset extracts, synset gloss extraction, and visual editor development. Henceforth, a number of studies on the latest researches that we show will be divided according to each phase.

In synsets extraction, the approach taken is the use of regular expressions (RE) for information extraction from monolingual resources and then it is perfected through the mapping of the PWN.

The matching or regular expression has been known to be useful for information extraction from a collection of documents [12]. Similar to our approach on information extraction from online dictionary and thesaurus document, the same approach can be proven to work for information extraction from clinical reports and physician notes. Turchine *et al.*. used more than 50 regular expressions and regular definitions to extract blood pressure information tags (patterns that indicate the presence of blood pressure information in a sentence), integer values of blood pressure, up to the information on intensification of anti-hypertensive treatment [13].

The main idea of mapping is to utilize the synset collections of a language and the semantic relations between synsets of available lexical reference system, that already existed in the new target language. At this point, there is no doubt that WordNet is a well-recognized lexical reference system for English language. Thus, WordNet is a valuable reference in the mapping approach into a number of languages; for example, into Korean language and Romanian [14-15].

The purpose of mapping is to link a synset of the target language with the right WordNet synset. In the case, the information from the bilingual dictionary of English language and the target language is used for this purpose. Firstly, to link the first synset s_{t1} of a word in the target language, to WordNet synset s_w, translation results w_e, in English are obtained from dictionary. In the instance w_e have five synsets in WordNet, namely s_{w1}, s_{w2}, …, s_{w5}, it means that s_t has five synset candidates, s_{w1}, s_{w2}, …, s_{w5}.

The main problem in mapping is to choose the correct synset out of s_{w1}, s_{w2}, …, s_{w5} and linking s_{t1} to it. A number of semantic ambiguities are encountered during automatic mapping, therefore a mechanism is needed for word sense disambiguation. Lee *et al.* offer six heuristics, where each produces a value. Those six values combined will be used as features input in the decision tree (or other supervised learning methods) which will select one synset candidate out of s_{w1}, s_{w2}, …, s_{w5} as the winner [14]. For the same mapping purpose, Barbu and Mitielu offer four word sense disambiguation heuristics for this. They propose a set of meta-rules, which evaluate the pros and cons of each heuristic, to replace supervised learning method that absolutely requires training and manual labeling beforehand [15].

Mapping plays an important role in our study because it is used in both synset extraction and semantic relation extraction between synsets. However, several other approaches were also used in this study, especially for semantic relation extraction. Ruiz-Casado *et al.* showed an attempt to identify lexical pattern that provide semantic relation between synsets by experimenting on English Wikipedia. The experiment used part-of-speech arrangement of sentences which contains those semantic relations [16]. Van Hage *et al.* identify these patterns in the types of commonly used words in the English language *(containing, found in, such as, in, is found in, etc.)* so that the part-of-speech tagger is not necessary for the approach they offer [17].

In gloss synset extraction, Chang *et al.*, on his study, performed definition extraction from offline documents that utilize a number of classification methods –such as decision tree and naïve Bayes– and support vector machine with various features –for example, character length, sentence position, the number of terms, single and bigram word, and part of speech [18]. Complementary, to identify documents that may contain definition sentences, Cui *et al.* offer a number of patterns which are mostly in English copula [19].

Attempts to visualize an ontology have been performed by several researchers. Some examples of ontology visualization that are general purpose and can be applied to any domain are models of DNA, Bohr model for atoms and WordNet. Visualizers that are not domain-specific become possible when a standardized structure is predetermined, for instance Resource Description Framework (RDF) Data through XML [20].

Specific visualizers for WordNet –as a giant semantic network– have actually already existed; for example, WordNet Explorer that uses 2-dimentional radial layout of a node-link diagram and Tree Map that shows the hyponym-hypernym relation between synsets [21]. Other WordNet visualizers are Synonym that utilizes force-directed graphs (code.google.com/p/synonym/) and VisuWord that is web-based and is equipped with gloss information from synsets (www.visuwords.com). Both Synonym and VisuWord used PWN 3.0 as reference. Hence, the fundamental difference between all other researches and ours is the capability of expansion from visualizer to graphical editor that can be used collaboratively.

3. Synonym Sets Extraction

Synset is a set composed of one or more words that have a synonym relationship, where each member can be used interchangeably without changing its meaning. Here is an example of synset and gloss in the PWN:

Animal, animate being, beast, brute, creature, fauna -- (a living organism characterized by voluntary movement)

From this example, it can be seen that all the words *animal, animate being, beast, brute, creature* and *fauna* are one synset with the meaning of a living organism that has the characteristics of conscious movement.

Through the perspective of WordNet as a large semantic network, a synset is represented as a vertex. This is in contrast with the general lexical reference system of a language, like monolingual dictionary with word as the smallest element. The acquisition of a large collection of Indonesian synset is what we will describe in this section. Based on its principle, the acquisition of a collection of synset is done through two methods: information extraction from monolingual lexical resources and mapping of PWN.

In the first method, monolingual lexical resources that will be used are KBBI and *Tesaurus Bahasa Indonesia* (Indonesian Thesaurus). Both resources are available in PDF format and are freely downloadable from the official site of the Ministry of Education and Culture of Indonesia.

KBBI is the most used official monolingual dictionary in Indonesia. It provides definitions of every sense from a lemma and examples of usage in sentences. Next, similar to other languages thesaurus, Indonesian thesaurus is a resource that lists a group of words that share the same meaning. Additionally, thesaurus also provides a number of relationships, such as the antonym, synonym, and hyponym; though the differences between hyponyms and synonyms are not shown explicitly like the differences between synonyms and antonyms.

Until this study is done, no ideal corpus (*e.g.* encyclopedias) in Indonesian can be found or is available.

(a)	(b)
bunga *n* **1** bagian tumbuhan yg akan menjadi buah, biasanya elok warnanya dan harum baunya; kembang; **2** jenis bagi berbagai-bagai bunga; **3** gambar hiasan (pd kain, pamor ukiran, dsb); **4** *ki* sesuatu yg dianggap elok (cantik) spt bunga; **5** bunga uang; rente; **6** tambahan untuk memperindah; **7** tanda-tanda baik;	[1]**bunga** *n* **1** kembang, kesuma, kujarat, puspa, puspita, sari, sekar; **2** *ki* dekorasi, hiasan, ornamen; **3** *ki* perempuan cantik, primadona; **4** *ki* bumbu, embel-embel, komplemen, pelengkap, suplemen, tambahan; [2]**bunga** *n* anak uang, anakan, riba;

Figure 2. Sample of Lemma Entry on KBBI and Thesaurus

In Figure 2, a sample entry for a lemma is shown, one from KBBI (a) and another from Indonesian Thesaurus (b). There are seven homonyms in KBBI for the word bunga (flower), whereas in English Thesaurus there are two words of the same sense. A pair of senses shown in thesaurus for *bunga* are *kembang* (flower) and *anak uang* (saving interest); each is shown on the 1st and 5th holonyms in KBBI.

In the first method, we are doing the extraction of monolingual lexical resources. The structure of an entry in each of these resources can be seen in regular definitions in Figure 3.

Regular Definition for KBBI

DefinitionPart	=	(Definition ("(" Explanation ")")? (: ExampleSentence)? (";")?)+
LemaDef	=	(HomonimNumber)? LemaPartOfSpeech (SenseNumber)? DefinitionPart
SublemaDef	=	(SublemaPartOfSpeechDefinitionPart)*
FirstForm	=	LemaDef <tab> SublemaDef
SecondForm	=	LemaPartOfSpeech "," SublemaDef <newline>
KBBIEntry	=	(FirstForm | SecondForm) <newline>

Regular Definition for Tesaurus Bahasa Indonesia

Sublema	=	<tab> LemaPartOfSpeechSynonymPart ";" <newline> AntonymPart <newline>
AntonymPart	=	<tab> "ant" AntonymSenseNumberLema
SynonymPart	=	((SenseNumber)? (lema)(,lema)* ";")+
Entry	=	LemaPartOfSpeechSynonymPart ";" <newline> (AntonymPart <newline>)* (sublema <newline>)*

Figure 3. Regular Definitions for KBBI and Indonesian Thesaurus

Both regular definitions will be used as a reference on information extraction mechanism.

However, there are two main problems that we identified in the processing of KBBI and Indonesian Thesaurus, those are:

1. In PWN, synonym relation should be bidirectional, meaning that if a wordw_1 has synonyms with w_2, then w_1 must be a synonym of w_2. Bidirectional relationship like this does not always happen in English.
2. Similar entries on KBBI and Indonesian Thesaurus have no links with one another. For example, the word w_1 has four senses in KBBI while in the thesaurus, it might only have a single sense. Furthermore, it is unknown which sense in KBBI is connected to that of the thesaurus.

Unlike Roget's Thesaurus for English language, Indonesian Thesaurus does not provide categorization for the entry and the sense. In addition, an entry is found to have a single sense (monosemous) but has some different semantic classes.

Considering the aforementioned problems, we set the stages for synset extraction as follow:

1. Extracting synset candidates of the entire Indonesian Thesaurus entries.
2. Adding synset candidates of the entire KBBI entries.
3. Eliminating all redundant (similar) synset candidates.
4. Combining all synset candidates that have similar meaning by utilizing clustering.

4. Semantic Relations Construction

There are two types of semantic relations that have been constructed to connect synsets collection acquired from the method discussed in the previous section. The first relation is an *is-a* relation (hypernym-hyponym) and the second one is a *part-of* relation (holonym-meronym).

The other relation is antonym relation, which is more appropriate to be categorized in lexical relation. Since this relation will connect adjective synsets and not noun synsets, thus, antonym will not be a part of relation that needs to be reconstructed in our research.

4.1. Is-a (Hypernym-Hyponym) Construction

Machine readable format of KBBI is used as the main source to acquire is-a relations. Information extraction into records of lemma, part-of-speech, and definition can be done through the regular definition in KBBI as shown in Figure 2. Following is an example of construction of *is-a* relation from an entry in KBBI:

kapal *n kendaraan pengangkut penumpang dan barang di laut (sungai dsb)*

(ship *n transportation vehicle for passengers and baggages on the sea (river, etc))*

From this extracted record, an *is-a* relation can be acquired, that is the word *kapal (ship)* as the hyponym of *kendaraan (vehicle)*, and *kendaraan* as the hypernym of *kapal* conversely. In most examples, it can be seen that the noun, which acts as the hypernym *(kendaraan)* is a part of a definition of the noun explained in an entry *(kapal).* Nevertheless, this condition turns out to be problematic since synonym and hypernym which are parts of a lemma in KBBI cannot be differentiated explicitly. Take the following entry for instance:

anjing *- n - binatang menyusui yg biasa dipelihara untuk menjaga rumah, berburu, dsb; Canis familiaris; (***dog** – noun - mammal which is cared to guard houses or hunt, *etc*.; Canis familiaris*)*

From the first part, it can be determined that the word *binatang (animal)* is the hypernym of *anjing (dog)*. However, *Canis Familiaris* is the synonym of *anjing*, not the hypernym of *anjing*.

kapal *n* kendaraan pengangkut penumpang dan barang di laut (sungai, dsb)
(*transportation vehicle for passengers and baggages on the sea (river, etc)*)

kendaraan	n	*something used for a ride (such as horse, cart, car)*
pengangkut	n	*device (ship, car, etc) for conveying*
penumpang	n	*person who boards on or rides (cart, ship, etc)*
Dan	p	*a conjunction (words, phrases, clauses, and sentences)*
barang	n	*general thing (everything which has a form or tangible)*
di	p	*a prefix to specify a location*
Laut	n	*salt water (in a large amount at a large area) which covers and divides land into continents and islands*
sungai	n	*large stream of river (usually created by nature)*

Figure 4. Definition Exploration of Every Word in the Word *kapal (ship)* Definition

With those conditions in mind, the construction of hypernym-hyponym pairs that we propose consists of three processes:

1. Word Sense Disambiguation (WSD) every word acquired in a definition to find the precise part-of-speech for each of them. In this case we use Lesk algorithm as the simple WSD process in this stage [22].
2. Definition simplification so that only the part of the definition that contains information of hypernym of lemma is extracted. The main task in this process is to discard the synonym part of a lemma entry.
3. Acquisition of hypernym from the simplified definition. A very important heuristic that we use in this process is that the first noun-phrase from the definition has high probability to be hypernym candidate of a synset even though definition exploration of every word in the definition still needs to be done (see Figure 4).

4.2. Is part-of (Holonym-Meronym) Construction

A mapping approach is used in the construction of holonym-meronym relationship or frequently referred as part-whole relationship in our research. For every Indonesian noun synset s_i, an English translation synset s_e can be obtained. Holonym and meronym synset from s_e, s_{eh} and s_{em} respectively, are acquired through the use of PWN. Furthermore, by translating s_{eh} and s_{em} into Indonesian, a set of synset s_{ih} and s_{im} is created. Finally, s_{ih} and s_{im} will be linked with s_i, each as holonym and meronym synset in Indonesian.

The main difficulty in this approach is the ambiguity in both translation to the target language, whether to Indonesian or to English. When translating s_i to English, a set of English synset, S_e, is produced instead of a single synset. Thus, WSD is needed to choose the appropriate s_e out of the member of S_e. Similar case can be seen as well when translating s_{eh} and s_{em} into Indonesian. Figure 5 shows the approach that we offer for synset level.

We also have tried to do similar mapping approach for word level. This word mapping can be done by simplifying the algorithm shown in Figure 5; by removing all WSD tasks required to get the appropriate synset, whether in English (step 3) or in Indonesian (step 7 and 12) since those are not yet required. Nevertheless, the post processing for assignment to the Indonesian synsets (Section 3) that have been acquired accurately is still needed.

```
1.  For Each s_i in AllIndonesianSynset
2.      S_e←TranslateIndonesianToEnglish(s_i)
3.      s_e←GetWSDMappedSynset(S_e, s_i)
4.      s_eh←GetAllPWNHolonym (s_e)
5.      For Each h in s_eh
6.          S_i←TranslateEnglishToIndonesian(h)
7.          s_ih←GetWSDMappedSynset(S_i , h)
8.          GenerateHolonymMeronymRelationInIndonesian(s_i,s_ih)
9.      s_em←GetAllPWNMeronym (s_e)
10.     For Each m in s_em
11.         S_i←TranslateEnglishToIndonesian(m)
12.         s_im←GetWSDMappedSynset(S_i , m)
13.         GenerateHolonymMeronymRelationInIndonesian(s_i,s_im)
14. Return
```

Figure 5. Algorithm for *Part-of* Semantic Relations Construction with Synset Mapping

5. Gloss Acquisition

In this third phase, the gloss word resources are retrieved from Wikipedia pages (both offline and online) and web page collections from search engines like Google. Online Wikipedia and thousands of web pages are used to complete synset and gloss collections since not all Indonesian synsets are covered in the offline Wikipedia.

Glosses retrieval from Wikipedia is done by accessing a specific web page. For instance, the following steps are taken to acquire gloss from the word *rumah (house)*. For offline Wikipedia, where r/u/m is the folder name, the acquisition is done at the page:

Wikipedia/articles/**r/u/m**/rumah.html

For the online version, where the id subdomain shows the Indonesian Wikipedia, the acquisition is done at the page:

http://id.wikipedia.org/rumah

While for the gloss acquisition through Google, three types of Indonesian copula, *e.g. adalah*, *ialah*, *merupakan*, are used in the query strings that will be passed on to the search engine. For instance:

?q=”rumah+adalah”

?q=”rumah+ialah”

?q=”rumah+merupakan”

Online Wikipedia or search engine is only utilized when the word *rumah (house)* is not available in the offline Wikipedia.

The problem faced in this gloss acquisition process can be explained through two gloss acquisition results below, for the words *predator* and *Joko Widodo* respectively, in which the multi sense homonym are as follows:

predator:

[1] binatang yang hidupnya dari memangsa binatang lain; hewan pemangsa hewan lain *(animal that lives by preying on other animals; animal which preys on other animals)*

[2] sebuah film sains fiksi yang diluncurkan pada tahun 1987 dan diarahkan oleh John McTiernan dan dibintangi oleh Arnold Schwarzenegger, ….. *(a science fiction movie that was launched in the 1987 and directed by John McTiernan and starred by Arnold Schwarzenegger, ...)*

We can understand that (1) is the original definition of the word *predator* (noun). Moreover, although it is inaccurate, (2) is still a definition since it is an explanation about a movie title (proper noun) that uses the same word. Both of the definitions must be accommodated (accepted) by Indonesian WordNet as PWN has also accommodated many similar cases; like the word *bush* for several senses, such as: *a low woody perennial plant, a large wilderness area, and George Walker Bush (U.S. President)*.

Joko Widodo:

[1] Joko Widodo adalah presiden ketujuh Republik Indonesia *(Joko Widodo is the seventh president of the Republic of Indonesia)*

[2] Joko Widodo adalah sosok pemimpin harapan rakyat *(Joko Widodo is public's expectation figure)*

[3] Joko Widodo adalah pemimpin yang bijaksana *(Joko Widodo is a wise leader)*

Although both of the glosses use the same copula to explain the word *Joko Widodo* (the seventh president of Indonesia), it is understandable that only (1) must be accepted as the right gloss, whereas (2) and (3) must be rejected since it is only an expression of a web author’s opinion towards *Joko Widodo*.

In cases like this, we use supervised method from machine learning to acquire Indonesian gloss from Indonesian web pages. Clarification models will be formed in order to separate the glosses that must be accepted and rejected.

After resources acquisition that was explained earlier in this section, the proposed method is divided into three phases: preprocessing, features extraction, and classification.

5.1. Preprocessing

The preprocessing phase is started with text cleansing used to transform a HTML page into a raw text. A number of regular expressions are utilized for this purpose. Next, sentence extraction and paragraph reconstruction are carried out since features extraction needs some information related with the word sequence or frequency in a paragraph.

5.2. Features Extraction

In order to generate the binary class C = {accept, reject} from a gloss word, the features F = {f_1, f_2, ..., f_n} from some instances must be known first if this problem is going to be solved with supervised learning or classification. In this research, we use seven features, all of which will contain positive integer. The following are the explanation of each feature used:

- Position that is indicated by the gloss candidate word sequence in a paragraph.
- Frequency or the number of gloss candidate word appearance in the document collections.
- Number of words that exist in a gloss candidate word.
- Number of important words *(non-stop words)* contained in a candidate word.
- Number of characters contained in a gloss candidate word.
- Number of all identical gloss candidate words generated.
- Number of noun contained in a gloss candidate word.

5.3. Classification

Backpropagation Feedforward Neural Networks (BPFFNN) and decision tree are chosen to be the model for classification phase.

BPFFNN is chosen because of the non-linear separable problem capability similar to Support Vector Machine. We used BPFFNN 7-P-1, a multi-layer architecture with seven input nodes, P hidden nodes, and one output node for accept or reject response. Momentum and Nguyen-Widrow initialization will be utilized to accelerate training time. In this research, we use total square error <= 0.2%, momentum = 0.2, learning rate = 0.3, and maximum epoch number = 500.

Decision tree is chosen because of its knowledge results visibility. Our experiments have shown that NUMCHAR (the number of character in a word) is the most important feature in decision making process by decision tree.

6. WordNet Visual Editor

It is nearly impossible to create a perfect lexical database for a language in a fully automatic way. We can see this fact from a number of proposed methods in the current research until this day, including several approaches we proposed in the previous three sections for Indonesian. The main problem encountered is the very high accuracy requirement in a large volume of data for each target language. If PWN has 82,115 noun synsets from a total of 117,659 synsets for all part of speech, then each of those synsets must have minimally one gloss.

An approach to build synset, gloss, and relation collection is indeed required as a significant foundation so that the effort (cost and time) needed is not as much as when it is done completely manually. Nevertheless, considering the fact that the accuracy of automatic construction will never be perfect because of errors and incompleteness, it is clear that manual editing is still needed. In this perspective, we call our proposed method as semi-automatic Indonesian WordNet building.

WordNet can be seen as a huge semantic network that connects hundreds of thousands of complete synsets from a natural language. If the nodes that represent synsets are the atomic element of a language, then the edges are the connection of each node that serves

as the semantic relations like hypernym, hyponym, meronym, and holonym. The gloss itself is just an attribute of the synset node. Therefore, for the convenience of manual editing, a web-based graph editor that can be accessed collaboratively is needed.

Adding antonym relation manually as the connection between word nodes is also included in this editing effort. The antonym relation is handled distinctively because antonym relation is a lexical relation that connects word nodes, that is different from semantic relations that connect synset nodes. A handful of volunteers have tried to fix the Indonesian WordNet and their works have been validated by an administrator through the same web based graph editor.

This graph editor is developed through three phases as follows:

Converting PWN Data and Index Files to Relational Database-preparing a relational database which acts as the graph reference that will be edited by users. The main consideration to use relational database is the ease of access of the server-side program to supply resources needed to visualize WordNet data and manage the user editing results before validation process. Files like .dat and .idx for each part of speech of PWN are used to optimize program or WN browser speed in accessing (reading) the synsets inside and not to changing (writing) the data. Therefore, it is impossible to use the data and index directly as the graph reference for users to edit.

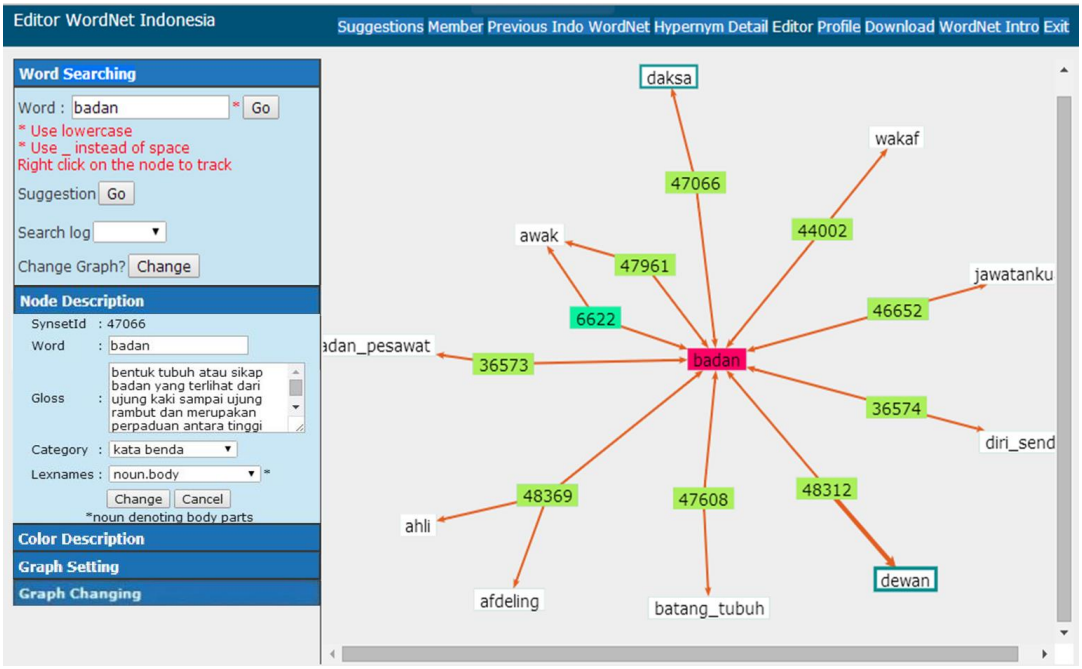
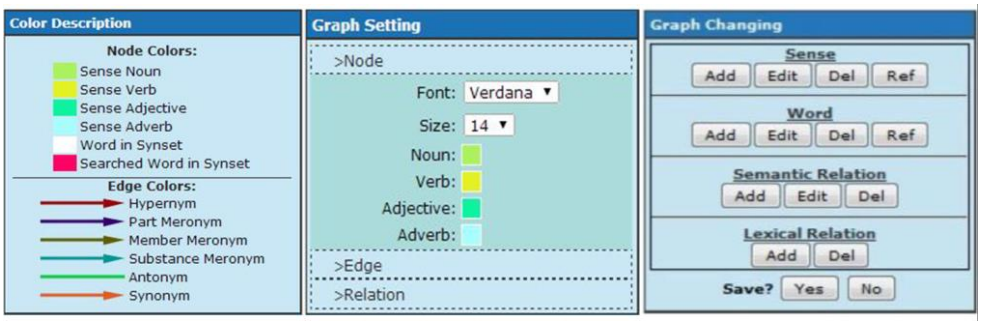


Figure 6. Web Based Graph Editor Overall Layout

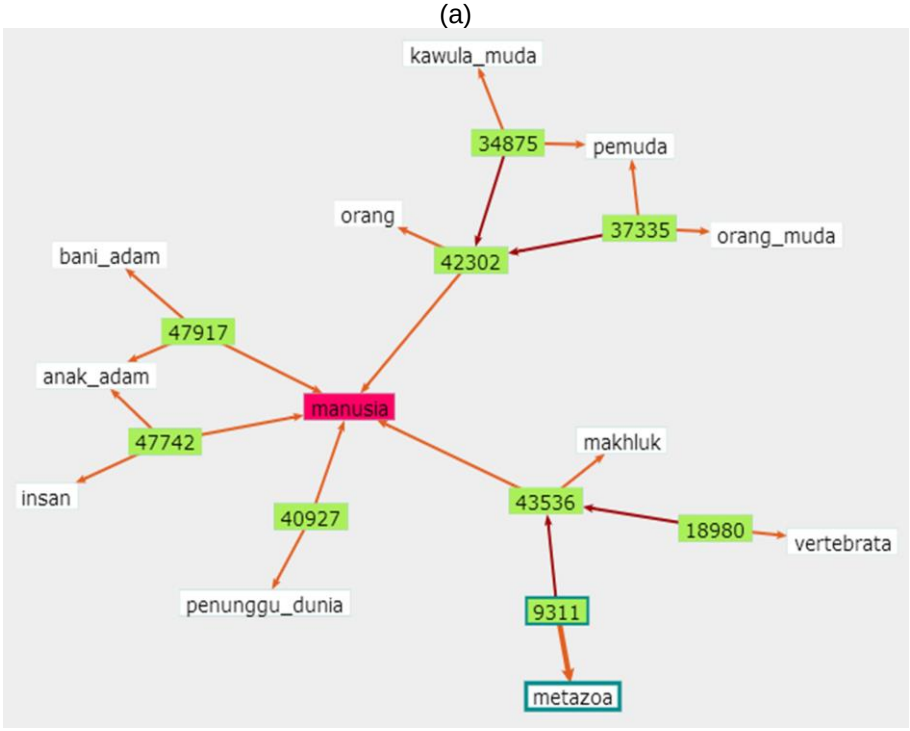
Visualization and Editing-are done using a force-directed graph, a method used to draw graph with a purpose to minimize the overlapping edges [23]. This method is simulated as if it was a physical system by assigning attractive and repulsive forces between nodes. This process is done continuously to find the optimal layout by minimizing the energy of the system. For the implementation of this method, we use *springy* and *springui* library.

Figure 6 shows the two main areas in the graph editor that we developed; the left area is used to show the editing menu options and the right area is used for the drawing canvas. Aside from the word searching and graph setting menu as shown in Figure 6, the other editing menu options are colors description, node description, graph changing, and add new sense as shown in Figure 7.a.

In the right area that is used for graph drawing that implements a force-directed graph (Figure 6), nodes are used to represent synsets accompanied with their respective numbers. Nonetheless, to help user understand the context of the editing focus easily, a number of Indonesian words, that act as the members of a particular synset are also shown similarly like a node but without outlinks. For instance, the word *manusia (human)* is visualized in Figure 7.b. The edge color is used to differentiate relation types between synsets.



(a)



(b)

Figure 7. Alternative Display of Graph Editor. (a). Other Facilities: Color Description, Graph Setting and Graph Changing; (b) Another Example of Graph Display for *Manusia (human)* Synset and Its Relations

Postprocessing-the last phase of WordNet Visual Editor development is to validate the proposed changes of synset, relation, and gloss from volunteer and to build the Indonesian WordNet lexical database. Ideally, this phase is done by several linguists.

7. Experimental Results

Up to the moment this research is put into writing, the effort to acquire a number of relations and glosses as well as manual correction using WordNet Visual Editor that we have developed is only done for noun synset category.

For synonym set extraction using Indonesian monolingual resources (KBBI and thesaurus), we successfully extracted 25,587 noun synsets that have only a single member or single word. For noun synsets that have multiple members, we managed to extract 11,898 synsets [8]. The effort to combine a number of similar multiple members synsets into a single synset has also been done using clustering technique.

In the construction of semantic relations that is done using mapping approach, the main problem faced when acquiring part-of (holonym-meronym) relation is the dependence of this approach on the translation mechanism *(see Figure 4, step 2, 6 and 11)*. There are 68% of all Indonesian words that do not have English translation in *Kamus Elektronik Bahasa Indonesia (KEBI)* and conversely, there are 56% of all English words that do not have Indonesian translation.

For the hypernym-hyponym construction, our approach has successfully acquired 24,256 pairs from 54,395 possible pairs in 91,029 records in KBBI [9]. There are two approaches done for the construction of holonym-meronym relation, namely word mapping and synset mapping. Both of these mapping approaches respectively yield 8,555 pairs for noun word mapping and 2,489 pairs for noun synset mapping. Aside from the incomplete Indonesian to English translation and vice-versa, the performance of the mapping approach used is also affected by the performance of WSD algorithm used [9].

In the phase of gloss acquisition from Indonesian noun synset collections, we were able to acquire 6,520 correct glosses with the help of copula –as the simple pattern– to obtain web page collections that are presumed to contain the gloss. This value is equal to 78.67% of 8,288 gloss candidates which were successfully acquired after being tested using supervised learning model, with the help of seven features that we extracted from each gloss candidate. Based on three experiments with 80:20, 70:30, and 60:40 –each of which shows the ratio of training set and testing set from labeled gloss candidate, we are confident that this method has an accuracy rate of 74.06% for decision tree and 75.40% for backpropagation feedforward neural networks [10].

On the use of WordNet Visual Editor that we have developed to repair synsets, all kinds of relations between synsets and words, as well as glosses, the focus is mainly on the manual correction to all relations that we have acquired. In contrast with the gloss correction –with proven accuracy in the range of 75%, the number of relations that were successfully constructed is quantitatively too small. In our last experiment, 3,852 out of 9,427 (40.86%) hypernym-hyponym relations have been checked manually for its accuracy. For the correction of holonym-meronym relations, the number of correct relations for each category are 1,141 out of 2,101 (40.86%) for part relation, 146 out of 179 (81.56%) for substance relation, and 180 out of 209 (86.12%) for member relation. Through this editor, the acquisition of 78 antonym relations between Indonesian words that cannot be done automatically with a particular method, can be realized to complement all other semantic relations [11].

8. Conclusion and Future Works

Through this paper, we have successfully showed that the effort to build Indonesian lexical database can be done semi-automatically. The automatic part was done in the attempt to acquire synset, gloss, and each of the relations. The non-automatic or manual part was done in the refinement of automatic results, in terms of quantity or quality.

In spite of the help of web-based WordNet Visual Editor that allows more than 85 volunteers to work collaboratively, our experience working on both sides –automatic and manual sides– shows that more than half of our total effort in building this Indonesian

lexical database is assisted by the automatic part. Based on all experiments that have been done, we can get some valuable notes for future works.

Firstly, from our proposed approach, it can be seen that semantic relation extraction is done first and then followed by gloss synset extraction. In order to increase the number of semantic relations to be extracted, we believe that the acquired gloss synset can be re-used as word sense disambiguation references in the mapping technique used in relation extraction.

Secondly, noun synsets are indeed the largest part and they require the greatest effort for the construction of synset and its relations in a lexical database like PWN. Even so, verb, adjective, and adverb synsets and their relations must still be built. Therefore, besides manual collaborative editing of noun synsets semantic relations, the construction of synsets and glosses from verb, adjective, and adverb along with their relations automatically becomes our main focus as well.

Thirdly, after all synsets, glosses, and their relations are completely available, the construction of image database, which is connected with Indonesian WordNet synset hierarchy, becomes possible to be realized in our last stage as our final future work. ImageNet, which was developed by Stanford University in conjunction with Princeton University, now provides more than 14 million images that are connected with 21,481 noun synsets in WordNet 3.0. These valuable resources have become the state-of-the-art dataset for computer vision research –through deep learning, which will then be utilized in Ubiquitous or Pervasive Computing. Thus, it will be a great achievement when millions of images from ImageNet can be mapped into Indonesian WordNet eventually.

Finally, we believe that our proposed methods can be implemented for the purpose of building WordNet in other natural languages in the world besides Indonesian. Furthermore, for Indonesian itself, we believe that the presence of Indonesian WordNet that we are currently working on will stimulate more and more further researches in text mining, web mining, and computational linguistic as well as deep learning and ubiquitous computing for Indonesian.

Acknowledgements

The authors would like to appreciate the thirteen undergraduate alumni from the Computer Science Department of Sekolah Tinggi Teknik Surabaya for independently doing the research on WordNet in Indonesian language as well as English during these last 17 years.

References

[1] M. Ramprasath and S. Hariharan, “A Survey on Question Answering System”, International Journal of Research and Reviews in Information Sciences (IJRRIS), vol. 2, no. 1, **(2012)**, pp. 171-178.
[2] R. Navigli, P. Velardi and A. Gangemi, “Ontology Learning and Its Application to Automated Terminology Translation”, IEEE Intelligent Systems, vol. 18, no. 1, **(2003)**, pp. 22-31.
[3] D. D. Putra, A. Arfan and R. Manurung, “Building an Indonesian WordNet”, the 2nd International Malindo Workshop (MALINDO2008), Selangor, Malaysia, **(2008)**.
[4] E. Margaretha and R. Manurung, “Comparing the Value of Latent Semantic Analysis on Two English-to-Indonesian Lexical Mapping Tasks”, Proceeding of Australasian Language Technology Association Workshop, Hobart, Australia, vol. 6, **(2008)**, pp. 88-96.
[5] S. T. Charoenporn, C. Mokarat, H. Isahara, H. Riza and P. Jaimai, “Synset Assignment for Bi-lingual Dictionary with Limited Resource”, Proceedings of the Third International Joint Conference on Natural Language Processing, Hydearabad, India, vol. 2, **(2008)**, pp. 673-678.
[6] H. R. Budiono and C. Hakim, “Collaborative Work on Indonesian Wordnet through Asian Wordnet”, Proceedings of the 8th Workshop on Asian Language Resources, Beijing, China, **(2010)**, pp. 9-13.
[7] N. H. M. Noor, S. Sapuan and F. Bond, “Creating the Open Wordnet Bahasa”, Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation (PACLIC25), Singapore, **(2011)**, pp. 255–264.

[8] Gunawan and A. Saputra, "Building Synsets for Indonesian WordNet with Monolingual Lexical Resources", Proceedings of The 2010 International Conference on Asian Language Processing (IALP 2010), Harbin, China, **(2010)**, pp. 297-300.
[9] Gunawan and E. Pranata, "Acquisition of Hypernymy-Hyponymy Relation between Nouns for WordNet Building", Proceedings of The 2010 International Conference on Asian Language Processing (IALP 2010), Harbin, China, **(2010)**, pp. 114-117.
[10] Gunawan, I. K. E. Purnama and M. Hariadi, "Supervised Learning Based Indonesian Gloss Acquisition", IAENG International Journal of Computer Science (IJCS), vol. 42, no. 4, **(2015)**, pp. 337-346.
[11] Gunawan, J. F. Wijoyo, I. K. E. Purnama and M. Hariadi, "WordNet Editor to Refine Indonesian Language Lexical Database", Proceedings of The 2011 International Conference on Asian Language Processing (IALP 2011), Penang, Malaysia, **(2011)**, pp. 47-50.
[12] G. G. Zweig and M. Padmanabhan, "Information Extraction from Documents with Regular Expression Matching", U.S. Patent 6 842 796 B2, **(2005)**.
[13] A. Turchin, N. S. Kolatkar, R. W. Grant, E. C. Makhni, M. L. Pendergrass and J. S. Einbinder, "Using Regular Expressions to Abstract Blood Pressure and Treatment Intensification Information from the Text of Physician Notes", Journal of the American Medical Informatics Association (AMIA), vol. 13, no. 6, **(2006)**, pp. 691-695.
[14] C. Lee, G. Lee, and J. Seo, "Automatic WordNet Mapping using Word Sense Disambiguation", Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, Hongkong, **(2000)**, pp. 142-147.
[15] E. Barbu and V. B. Mititelu, "Automatic Building of Wordnets", Journal of Recent Advances in Natural Language Processing IV: Selected Papers from RANLP 2005, Current Issues in Linguistic Theory, vol. 292, **(2007)**, pp. 217-226.
[16] M. R. Casado, E. Alfonseca, and P. Castells, "Automatic Extraction of Semantic Relationships for WordNet by Means of Pattern Learning from Wikipedia", Lecture Notes in Computer Science 3513, Natural Language Processing and Information Systems, Edited Andrés Montoyo, Rafael Muńoz, and Elisabeth Métais, Springer International Publishing, **(2005)**, pp. 67-79.
[17] W. R. van Hage, H. Kolb, and G. Schreiber, "A Method for Learning Part-Whole Relations", Lecture Notes in Computer Science 4273, The Semantic Web – ISWC 2006, Edited Isabel Cruz *et al.*, Springer International Publishing, **(2006)**, pp. 723-735.
[18] X. Chang and Q. Zheng, "Offline Definition Extraction Using Machine Learning for Knowledge-Oriented Question Answering", Communications in Computer and Information Science Series, Advanced Intelligent Computing Theories and Applications: With Aspects of Contemporary Intelligent Computing Techniques, Edited D.-S. Huang, L. Heutte and M. Loog, Springer Berlin Heidelberg, vol. 2, **(2007)**, pp. 1286-1294.
[19] 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, **(2007)**.
[20] Z. Nevěřilová, "Visual Browser: A Tool for Visualizing Ontologies", Proceedings of I-KNOW '05, Know-Center in coop. with Graz Uni, Joanneum Research and Springer Pub. Co. Graz, Austria, **(2005)**, pp. 453-461.
[21] C. Collins and G. Penn, "WordNet Explorer: Applying Visualization Principles to Lexical Semantics", Technical Report Series of Knowledge Media Design Institute (KMDI-TR-2007-2) University of Toronto, **(2007)**.
[22] M. E. Lesk, "Automatic Sense Disambiguation using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream cone", Proceedings of the 5th Annual International Conference on Systems Documentation (SIGDOC '86), New York, USA, **(1986)**, pp. 24-26.
[23] Y. F. Hu, "Efficient and High Quality Force-Directed Graph Drawing", the Mathematica Journal, vol. 10, no. 1, **(2005)**, pp. 37-71.

Authors

Gunawan, He received diploma, bachelor, and master degrees in Computer Science from Sekolah Tinggi Teknik Surabaya (STTS), Surabaya, East Java, Indonesia. His research interest includes Computational Linguistic, Data and Web Mining, and Information Retrieval. Currently, he is a staff of Computer Science Department of STTS, Surabaya, Indonesia. He is working toward the Ph.D. degree at Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia.

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

Mochamad Hariadi, He received the B.E. degree in Electrical Engineering Department of Institut Teknologi Sepuluh Nopember (ITS), Surabaya, 1995; 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 Department of ITS, Surabaya, Indonesia. His research interest is in Video and Image Processing, Data Mining and Intelligent System.

