Identifying and Exploiting Definitions in Wordnet Bahasa

David Moeljadi **Francis Bond**
Division of Linguistics and Multilingual Studies
Nanyang Technological University
Singapore
D001@ntu.edu.sg, bond@ieee.org

Abstract

This paper describes our attempts to add Indonesian definitions to synsets in the Wordnet Bahasa (Nurril Hirfana Mohamed Noor et al., 2011; Bond et al., 2014), to extract semantic relations between lemmas and definitions for nouns and verbs, such as synonym, hyponym, hypernym and instance hypernym, and to generally improve Wordnet. The original, somewhat noisy, definitions for Indonesian came from the Asian Wordnet project (Riza et al., 2010). The basic method of extracting the relations is based on Bond et al. (2004). Before the relations can be extracted, the definitions were cleaned up and tokenized. We found that the definitions cannot be completely cleaned up because of many misspellings and bad translations. However, we could identify four semantic relations in 57.10% of noun and verb definitions. For the remaining 42.90%, we propose to add 149 new Indonesian lemmas and make some improvements to Wordnet Bahasa and Wordnet in general.

1 Introduction

A lexical database with comprehensive data about words, definitions, and examples is very useful in language research. In Princeton Wordnet, nouns, verbs, adjectives, and adverbs are grouped into sets of cognitive synonyms (synsets) which are interlinked through a number of semantic relations (Fellbaum, 1998; Fellbaum, 2005). Since its creation, many other wordnets in different languages have been built based on Princeton Wordnet (PWN) (Bond and Paik, 2012; Bond and Foster, 2013). One of them, Wordnet Bahasa, is built as a lexical database of the Malay language. At present, it consists of two language variants: Indonesian and Standard Malay. It combines data from several lexical resources: the French-English-Malay dictionary (FEM), the KAmus Melayu-Inggeris (KAMI), and wordnets for English, French and Chinese (Nurril Hirfana Mohamed Noor et al., 2011, p. 258).

We added Indonesian definitions from the Asian Wordnet project (Riza et al., 2010) to Wordnet Bahasa. To the best of our knowledge, the Asian Wordnet project is the only project that translated the English definitions of some synsets in PWN into Indonesian. However, the definitions were crowd sourced and had little quality control so not all of 14,190 definitions could be directly transferred. Many of the definitions had problems and needed to be cleaned up. The definitions for nouns and verbs which had been cleaned up were exploited to extract relations, such as synonym, hyponym, hypernym and instance hypernym, between lemmas and definitions. The method of extracting these relations was done in Bond et al. (2004) to build an ontology. We used Python (3.4, Python Software Foundation) and the Natural Language Toolkit (NLTK) (Bird et al., 2009) to process the data.

This paper is organized as follows: Section 2 describes the process of cleaning up the definitions, Section 3 explains the process of extracting hypernyms and other relations from the definitions. Section 4 presents the results and discussion and Section 5 concludes.

2 Cleaning up the definitions

As mentioned in Section 1 above, the definitions we had available were not clean. Many infelicities were found, such as misspellings, definitions using abbreviations, typos, synsets having more than one similar definitions, definitions written in English, improper use of hyphens, and lemmas written as the first word in the definitions. Each error is illustrated in the following subsections.

2.1 Correcting and deleting definitions

Words in the definitions which are not spelled correctly according to standard Indonesian, such as *dimana* "where" and *lain lain* "others", as well as typos such as *enerji* "energy" and *bagain* "part", were semi-automatically corrected. Since the typos are many and scattered throughout the file, we may have missed some. Abbreviations, most of them are prepositions, such as *dgn* "with" and *utk* "for", were also normalized to their full forms (see Table 1).

Before correction	After correction	Meaning	Number of hits
(double space)	(single space)		416
dimana	*di mana*	"where"	313
dengans	*dengan*	"with"	121
dgn	*dengan*	"with"	93
utk	*untuk*	"for"	52
kpd	*kepada*	"to"	25
pd	*pada*	"at"	23
lain lain	*lain-lain*	"others"	21
enerji	*energi*	"energy"	12
bagain	*bagian*	"part"	12
spt	*seperti*	"like"	12
dr	*dari*	"from"	10
thdp	*terhadap*	"toward"	10
sst	*sesuatu*	"something"	3

Table 1: Some examples of misspellings, abbreviations and typos, before and after the correction

Definitions which are obviously written in English or just names, were deleted (see Table 2).

Synset	Definition
03491491-n	Hanging Gardens of Babylon
09164241-n	ho chi minh city
10875910-n	George Herbert Walker Bush
11252392-n	rain in the face
13615557-n	a unit of measure for capacity officially adopted in the British Imperial System

Table 2: Some examples of deleted definitions

Some definitions had hyphens separating the words. In this case, the hyphens were deleted (see Table 3).

Synset	Definition	
14118423-n 'severe diabetes mellitus with an early onset'	*diabetes-mellitus-tergantung-insulin* "diabetes mellitus depending on insulin"	Before correction
	diabetes mellitus tergantung insulin	After correction

Table 3: An example of a definition having hyphens, before and after the correction

For definitions in which the first word is the same as the lemma with the real definition placed between brackets afterwards, the first word and the brackets were deleted (see Table 4).

Synset	Definition	
09543673-n 'an evil spirit or ghost'	*Ghoul (roh jahat atau hantu)* "Ghoul (an evil spirit or ghost)"	Before correction
	roh jahat atau hantu "an evil spirit or ghost"	After correction

Table 4: An example of a definition with lemma as the first word, before and after the correction

2.2 Choosing definitions

Some synsets have two or more different definitions as shown in Table 5. The longest one which includes other definitions, is assumed to be the correct one and automatically selected as the best definition.

Synset	Definition	
07904637-n 'gin flavored with sloes (fruit of the blackthorn)'	*buah dari semak* "fruit of the blackthorn"	Before cleaning up
	gin yang diberi rasa sloea "gin flavored with sloes"	
	gin yang diberi rasa sloea (buah dari semak) "gin flavored with sloes (fruit of the blackthorn)"	
	gin yang diberi rasa sloea (buah dari semak)	After cleaning

Table 5: An example of a synset with many parts of definition, before and after the cleaning up

However, if the definitions are all completely different and one of them was considered good based on the English and Japanese definitions, that one was chosen to be the correct one (see Table 6). This manual checking was done by the first author who has a good command of Indonesian, English, and Japanese.

If we found no satisfying definition after checking and comparing with the English and Japanese definitions, one or two of the words in the definitions were manually corrected (see Table 7).

After the cleaning up process, we made the Indonesian definitions available in the Open Multilingual Wordnet (1.2) hosted by Nanyang Technological University in Singapore (`http://compling.hss.ntu.edu.sg/omw/`). Figure 1 shows a screenshot of synset 06254371-n 'heliogram' with its Indonesian definition.

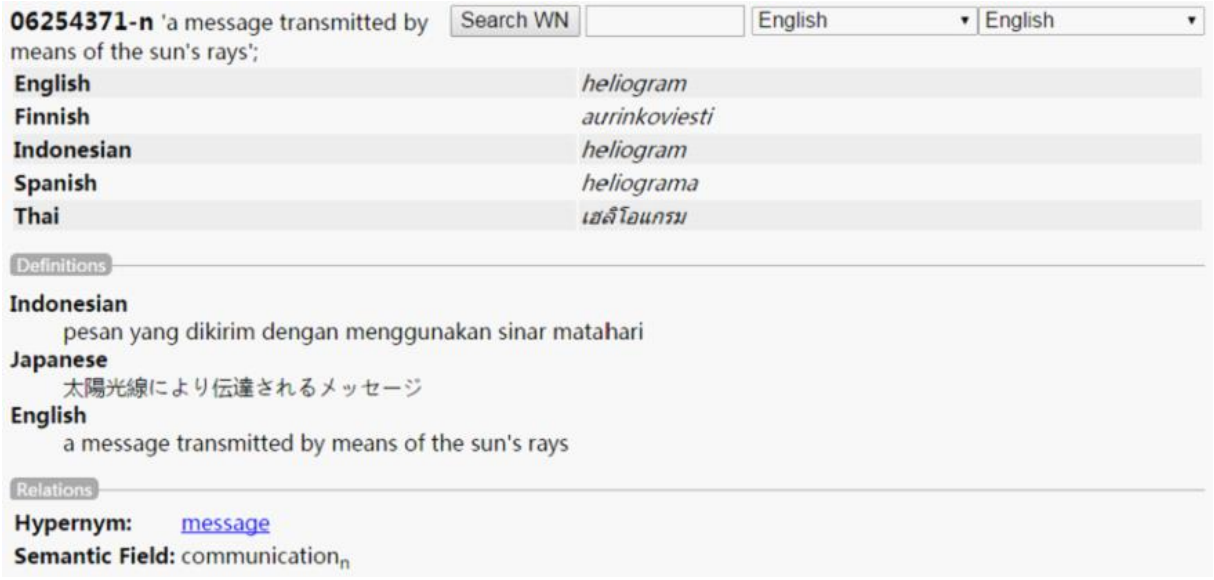

Figure 1: A screenshot of synset 06254371-n 'heliogram'

Synset	Definition	
01711910-a 'causing a sharply painful or stinging sensation'	*kedinginannya menggigit ke tulang* "the coldness bites to bones"	Before correction
	kedinginannya menusuk ke tulang "the coldness stings to bones"	
	sejuk hingga menggigit ke tulang "cool biting to bones"	
	sejuk hingga menusuk ke tulang "cool stinging to bones"	
	sejuk hingga menusuk ke tulang "cool stinging to bones"	After correction

Table 6: An example of a synset having many definitions, before and after the correction

Synset	Definition	
00731471-a 'supported by both sides'	*didukung oleh dua negara* "supported by both countries"	Before correction
	didukung oleh dua partai "supported by both parties"	
	didukung oleh dua pihak "supported by both sides"	After correction

Table 7: An example of a synset having two definitions, before and after the correction

3 Extracting relations from the definitions

Unlike Bond et al. (2004) who parsed the definition sentences using a grammar before extracting hypernyms and other relations, we simply used regular expressions. Indonesian has a strong tendency to be head-initial (Sneddon et al., 2010, pp. 160-162). In a noun phrase with an adjective, a demonstrative or a relative clause, the head noun precedes the adjective, the demonstrative or the relative clause. Typically numerals and classifiers precede the head noun (Alwi et al., 2014, pp.251-255).

Example (1) shows the Indonesian definition of synset 09500625-n 'Pegasus', the head of which is preceded by a numeral prefix *se-* "one" and a classifier *ekor* (lit. "tail") and followed by an attributive verb *bersayap* (lit. "having wings") and a prepositional phrase.

(1) *seekor **kuda** bersayap dalam mitologi Yunani*
one-CL horse winged in mythology Greece

"a winged horse in Greek mythology"

Example (2) contains a part of the Indonesian definition of synset 05316175-n 'ocular muscle'. Its head *otot-otot* "muscles" is in the plural (reduplicated) form, preceded by *satu dari* "one of" and followed by an adjective *kecil* "small".

(2) *satu dari **otot-otot** kecil pada mata...*
one of muscle-RED small at eye

"one of the small muscles of the eye"

We assume that after modifying the definitions, relations between lemmas and definitions can be extracted from the first lexical word (i.e. the head) in the definitions.

3.1 Modifying the definitions

For each definition for nouns and verbs, we removed the following words at the beginning:

(i) words which are written between brackets, such as *(Ilmu komputer)* "(Computer science)" relating to domain

(ii) numerals, such as *satu* "one", *tiga* "three", and *5* "five"

(iii) determiners, such as *setiap* "every", *sejenis* "a kind of", *semacam* "a sort of", *sembarang* "any kind of", *salah satu* "one of", *suatu* "a (for thing)", *sebuah* "a (for thing)", *seorang* "a (for person)", *seekor* "a (for animal)", *selembar* "a piece of", *sekelompok* "a group of", *beberapa* "some", *berbagai* "various", and *segala* "all"

(iv) relativizer *yang* "which"

(v) prepositions, such as *untuk* "for", *dari* "of", and *dalam* "in"

(vi) other stop words, such as *seperti* "like", *tentang* "about", *termasuk* "including", and *biasanya* "usually"

We also changed the plural (reduplicated) form of the head to its singular (non-reduplicated) form, for example *otot-otot* "muscles" was changed to *otot* "muscle" and *daun-daunan* "foliage, a cluster of leaves" was changed to *daun* "leaf". Punctuations such as slashes (/), semicolons (;), and commas (,) dividing two words were replaced as a space. After we made these changes, the first word in the definition was taken as a potential genus term.

3.2 Extracting relations

The first step was to check whether each first word of the definitions is in Wordnet or not. If it is not in Wordnet, we checked whether it is in *Kamus Besar Bahasa Indonesia* (KBBI) "The Great Dictionary of the Indonesian Language of the Language Center" or not. KBBI is published by the language institute who provides support for the standardization and propagation of Indonesian. Its third edition has been made online to public and has an official site (`http://badanbahasa.kemdikbud.go.id/kbbi/`) (Alwi et al., 2008).

The next step was to check whether the lemma synset is the same as the synset of the first word in the definition. This allows us to identify when the same word is used to define the lemma. Besides synonyms, hyponyms can also be employed to define the lemma. In order to confirm this, the lemma synset was compared to the hyponyms of the first word in the definition.

The next important step was to check whether the hypernym is used to define the lemma by comparing the hypernyms and instance hypernyms of the lemma synset with the synsets of the first word in the definition. If a lemma does not have any hypernym in Wordnet, we checked whether it has instance hypernym. Finally, lemmas having neither hypernyms nor instance hypernyms were checked by hand.

4 Results and discussion

The definition file which originally has 14,190 lines of definitions was cleaned up and 1,522 definitions (10.7%) were deleted. The remaining 12,668 definitions consist of 10,549 definitions for nouns, 1,663 definitions for adjectives, 409 definitions for verbs, and 47 definitions for adverbs. Although these definitions are considered quite clean, they may still contain small errors as mentioned in Section 2.1. Since adjectives and adverbs do not have relations such as hypernym in Wordnet, we only examined nouns and verbs. Out of 10,958 definitions for nouns and verbs, we could extract four relations from 6,257 definitions (57.10%) as shown in Table 8. The remaining 4,701 definitions (42.90%) have problems, such as words which could not be found in Wordnet and lemmas without explicit relations as shown in Table 9.

Most of the relations we extracted (95.89%) are hypernym and instance hypernym. The remaining are synonym and hyponym as shown in Table 8 for synset 00004475-n and 00029677-n. Synset 00004475-n has six Indonesian lemmas. One of these lemmas, i.e. *makhluk* "being", is used as the head of its definition and thus we regard the lemma is synonymous with the definition. Synset 00029677-n has *proses* "process" as one of its lemmas, which is the hypernym of the head of the definition *fenomena* "phenomenon".

Out of the 4,701 definitions for which we could not find the relations, most of them (83.88%) have hypernyms which are different from the first word in the definitions. We found four patterns for this problem (see Table 9):

1. The genus term is correct but Wordnet Ba-

Relation	Number of synsets	Example	
		Synset	Definition
Hypernym	5,451	00021939-n artifact	*suatu* ***objek*** *buatan manusia* "a man-made object"
Instance hypernym	549	02956500-n Capitol	***gedung*** *DPR di AS* "the government building in the United States"
Synonym	252	00004475-n organism	***makhluk*** *hidup yang dapat mengembangkan kemampuan bertindak independen* "a living thing that can develop the ability to act independently"
Hyponym	5	00029677-n process	*sebuah* ***fenomena*** *yang berkelanjutan* "a sustained phenomenon"
Total	6,257		

Table 8: Relations extracted from lemmas and definitions

Problem	Number of synsets	Example	
		Synset	Definition
No match	3,943	14350206-n myelitis	***inflamasi*** *pada syaraf tulang belakang* "inflammation of the spinal cord"
		14573846-n viremia	***kehadiran*** *suatu virus di dalam aliran darah* "the presence of a virus in the blood stream"
		13251154-n clobber	***istilah*** *informal untuk harta pribadi* "informal terms for personal possessions"
		07603411-n choc	***singkatan*** *dalam bahasa Inggris untuk coklat* "colloquial British abbreviation for chocolates"
		14364217-n sword-cut	***bekas*** *luka dari sayatan pedang* "a scar from a cut made by a sword"
		00046344-n stunt	*tidak biasa atau berbahaya* "not usual or dangerous"
Word not in Wordnet			
- Word in KBBI	252	13436063-n automatic data processing	***pemrosesan*** *data secara otomatis* "automatic data processing"
		07865105-n chili dog	***hot*** *dog dengan daging sapi diberi cabai bubuk* "a hotdog with chili con carne on it"
- Word not in KBBI	495	14099050-n visual aphasia	***ketidakmampuan*** *memahami kata-kata tertulis* "inability to perceive written words"
		09603258-n Pluto	***karakter*** *kartun anjing ciptaan Walt Disney* "a cartoon character created by Walt Disney"
		14155506-n cystic fibrosis	*disebabkan kerusakan suatu gen* "caused by defect in a single gene"
		00662589-v insure	*membagikan kawasan untuk kawalan tentara* "allot regions for soldiers"
No explicit relations	11	01773734-v grudge	*terpaksa menerima atau mengakui* "accept or admit unwillingly"
Total	4,701		

Table 9: Problems found in extracting relations

hasa does not have the right synset for the lemma. For example, synset 14350206-n 'myelitis' has 14336539-n 'inflammation' as its hypernym, which is also the first word in the English and Indonesian definitions. Wordnet Bahasa does have *inflamasi* "inflammation" but only in a different synset.

2. The semantic relation is not written explicitly in the definition. For example, synset 14573846-n 'viremia' has *kehadiran* "presence" as the first word in the English and Indonesian definitions which has nothing related with the semantic relation.

3. The genus candidate is a relational noun. For example, synset 13251154-n has *istilah* "terms" and synset 07603411-n has *singkatan* "abbreviation" as the first word in the definition. To get the real genus term requires more parsing.

4. Compounds were not extracted. For example, although the head of the definition of synset 14364217-n, was *bekas luka* "scar" (lit. "former wound"), we extracted only the first word *bekas* "former, past"

5. The definition is incomplete. For example, the Indonesian definition for synset 00046344-n lacks the head noun *usaha* "feat"

The second problem we found is that the first word in 747 definitions (15.89%) is not in Wordnet. In this case, we checked whether the word is in the Indonesian dictionary (KBBI) or not as mentioned in the previous section. We found 252 definitions having 149 unique words (the heads) which are in KBBI but not in Wordnet. Some of them are compounds as in synset 07865105-n with the definition *hot dog dengan daging sapi diberi cabai bubuk* "a hotdog with chili con carne on it". We did not distinguish compounds and thus, failed to extract *hot dog* as the head. The word *hot* does exist in KBBI as an adjective meaning 'sexually excited or exciting'.

The remaining 495 definitions have 235 unique words which are not in KBBI. We examined three patterns for this:

1. Derived words with negation are not listed as lexical items in KBBI. For example, the word *ketidakmampuan* "inability" (lit. "not able-ness") has the stem *tidak mampu* "not able" with a circumfix *ke-...-an* to nominalize. Including in this group are *ketidakadaan* "absence" (lit. "not present-ness") and *ketidaksempurnaan* "imperfection" (lit. "not perfect-ness").

2. The online KBBI data is not perfect, it does not include all Indonesian words listed in the paper dictionary. For example, the word *karakter* "character" is listed in the paper dictionary but not in the online version.

3. The Indonesian definition is incomplete. For example, the Indonesian definition for synset 14155506-n lacks the head noun *penyakit* "disease".

4. The Indonesian definition is incorrect. For example, the Indonesian definition for synset 00662589-v.

We found 11 lemmas have no explicit semantic relations with the definitions. They are all verbs: 01773734-v 'grudge', 00616857-v 'neglect', 01336635-v 'overlay', 01767949-v 'strike', 01944252-v 'hover', 02086805-v 'stampede', 02119241-v 'ignore', 02150510-v 'watch', 02413480-v 'work', 02581477-v 'prosecute', and 02673965-v 'stand out'.

5 Summary and future work

We have presented the process of cleaning up the definitions and extracting relations from the definitions. While doing the relation extraction, we spotted errors such as incompleteness and incorrectness in the definitions which we could not detect only by cleaning up the definitions. The reason why there are errors is probably because of little quality control in the translation process. In addition, we found things to be improved in Wordnet Bahasa and Wordnet in general. Based on our findings, we propose to:

1. Edit the incomplete Indonesian definitions. For example, definitions for synset 00046344-n which lacks the head noun *usaha* "feat" and 14155506-n which lacks the head noun *penyakit* "disease", as mentioned in Section 4

2. Delete the incorrect Indonesian definitions. For example, definitions for synset 00662589-v 'insure' which has the Indonesian definition *membagikan kawasan untuk kawalan tentara* "allot regions for soldiers"

3. Add 149 new lemmas from KBBI and possibly derived words with negation to Wordnet Bahasa

4. Add existing lemmas in Wordnet Bahasa to the correct synsets. For example, *inflamasi* to be added to synset 14336539-n 'inflammation'

5. Edit definitions in Wordnet to make them more informative, possibly add the hypernyms. For example, instead of having definition *jenis dari genus Soleidae* "type genus of the Soleidae" for synset 02664136-n 'Solea', we propose *jenis ikan dari genus Soleidae* "type of fish from the Soleidae genus"

6. Standardize the definitions in Wordnet, possibly make some guidelines for definitions. For example, regarding the numerals, some of them are written alphabetically, as in synset 09506337-n 'Fury' *tiga monster berambut ular...* "three snake-haired monsters...", but some of them are written in numbers, as in synset 09549416-n 'Hyades' *7 putri Atlas...* "7 daughters of Atlas...". Another problematic case is circular definitions.

For example, for synset 04658942-n 'inhospitableness' *memiliki sifat tidak ramah* "having an unfriendly and inhospitable disposition" and synset 04657876-n 'unfriendliness' "an unfriendly disposition"

Acknowledgments

Thanks to Hammam Riza who gave us permission to use the Indonesian definitions from Asian WordNet project. Thanks to Randy Sugianto and Ruli Manurung for their help. This research was supported in part by the MOE Tier 2 grant *That's what you meant: a Rich Representation for Manipulation of Meaning* (MOE ARC41/13).

References

Hasan Alwi, Dendy Sugono, and Sri Sukesi Adiwimarta. 2008. *Kamus Besar Bahasa Indonesia Dalam Jaringan (KBBI Daring)*. 3 edition.

Hasan Alwi, Soenjono Dardjowidjojo, Hans Lapoliwa, and Anton M. Moeliono. 2014. *Tata Bahasa Baku Bahasa Indonesia*. Balai Pustaka, Jakarta, 3 edition.

Steven Bird, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

Francis Bond, Eric Nichols, Sanae Fujita, and Takaaki Tanaka. 2004. Acquiring an ontology for a fundamental vocabulary. In *20th International Conference on Computational Linguistics (COLING-2004)*, pages 1319–1325, Geneva.

Francis Bond, Lian Tze Lim, Enya Kong Tang, and Hammam Riza. 2014. The combined wordnet bahasa. *NUSA: Linguistic studies of languages in and around Indonesia*, 57:83–100.

Christiane Fellbaum. 1998. *WordNet: an electronic lexical database*. MIT Press, Cambridge.

Christiane Fellbaum. 2005. WordNet and wordnets. In *Encyclopedia of language and linguistics*, pages 665–670. Elsevier, Oxford, 2 edition.

Nurril Hirfana Mohamed Noor, Suerya Sapuan, and Francis Bond. 2011. Creating the open Wordnet Bahasa. In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation (PACLIC 25)*, pages 258–267, Singapore.

Hammam Riza, Budiono, and Chairil Hakim. 2010. Collaborative work on Indonesian WordNet through Asian WordNet (AWN). In *Proceedings of the 8th Workshop on Asian Language Resources*, pages 9–13, Beijing, China. Asian Federation for Natural Language Processing.

James Neil Sneddon, Alexander Adelaar, Dwi Noverini Djenar, and Michael C. Ewing. 2010. *Indonesian Reference Grammar*. Allen & Unwin, New South Wales, 2 edition.