Identifying and Exploiting Definitions in Wordnet Bahasa

David Moeljadi, Francis Bond
Division of Linguistics and Multilingual Studies,
Nanyang Technological University, Singapore

Global WordNet Conference 2016,
Romanian Academy Library, Bucharest

27 January 2016
Wordnet Bahasa

- A lexical database of the Malay languages: Indonesian and Standard Malay (Nurril Hirdana Mohamed Noor et al., 2011; Bond et al., 2014)
- Built based on Princeton Wordnet
- 40,493 synsets, 118,903 senses, 43,113 unique words
- 14,190 Indonesian definitions from the Asian Wordnet project (Riza et al., 2010)
  - crowd-sourced definitions
  - little quality control
- Add Indonesian definitions + clean up
  - Correcting, deleting and choosing definitions semi-automatically using Python 3.4
  - Extracting semantic relations between lemmas and definitions using Python 3.4 and the Natural Language Toolkit (NLTK) (Bird et al., 2009)
  - Ideas to improve Wordnet
Correcting definitions (1/3)

- **Misspellings, typos, abbreviations**

| 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             |

**Table:** Some examples of misspellings, abbreviations and typos, before and after the correction
Deleting hyphens between words

| Synset              | Definition                                                                 | Before correction | After correction |
|---------------------|-----------------------------------------------------------------------------|-------------------|------------------|
| 14118423-n ‘severe diabetes mellitus...’ | *diabetes-mellitus-tergantung-insulin*  
|                     | “diabetes mellitus depending on insulin”                                  |                   |                  |
|                     | *diabetes mellitus tergantung insulin*                                   |                   |                  |

Table: An example of a definition having hyphens, before and after the correction
## Correcting definitions (3/3)

- Deleting the first word which is the same as the lemma and brackets

| Synset         | Definition                                    | Before correction | After correction |
|----------------|-----------------------------------------------|-------------------|------------------|
| 09543673-n     | *Ghoul (roh jahat atau hantu)*               |                   |                  |
| ‘an evil spirit or ghost’ | “Ghoul (an evil spirit or ghost)”         |                   |                  |
|                | *roh jahat atau hantu*                       |                   |                  |
|                | “an evil spirit or ghost”                    |                   |                  |

**Table:** An example of a definition with lemma as the first word, before and after the correction
## Deleting definitions

### Definitions written in English or names

| 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:** Some examples of deleted definitions
Choosing the longest definition which includes others

| Synset            | Definition                          | Before cleaning up |
|-------------------|-------------------------------------|--------------------|
| 07904637-n        | *buah dari semak*                  |                    |
| ‘gin flavored     |  “fruit of the blackthorn”         |                    |
| with sloes        | *gin yang diberi rasa sloea*       |                    |
| (fruit of the     |  “gin flavored with sloes”         |                    |
| blackthorn)’      | *gin yang diberi rasa sloea*       |                    |
|                   |  *(buah dari semak)*               |                    |
|                   |  “gin flavored with sloes (fruit   |                    |
|                   |      of the blackthorn)”           |                    |
|                   | *gin yang diberi rasa sloea*       |                    |
|                   |  *(buah dari semak)*               |                    |

**Table:** An example of a synset with many parts of definition, before and after the cleaning up.
Choosing the best definition based on the English and Japanese definitions

| Synset                     | Definition                                      | Before correction | After correction |
|----------------------------|--------------------------------------------------|-------------------|------------------|
| 01711910-a                 | *kedinginannya menggigit ke tulang* “the coldness bites to bones” |                   |                  |
|                            | *kedinginannya menuisuk 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” |                   |                  |

**Table:** An example of a synset having many definitions, before and after the correction
Choosing definitions (3/3)

Correcting manually one or two words in the definition

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

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

- Open Multilingual Wordnet (1.2) hosted by NTU in Singapore (http://compling.hss.ntu.edu.sg/omw/)

Figure: A screenshot of synset 06254371-n ‘heliogram’
Extracting relations from definitions

- The basic method is based on Bond et al. (2004) which use a grammar to parse definitions and extract relations
- We used regular expressions in this paper
- Indonesian language:
  - strong tendency to be head-initial (Sneddon et al. 2010, pp. 160-162)
  - Noun + Adjective, Noun + Demonstrative, Noun + Relative clause
  - Numeral + Noun, Classifier + Noun
- We modified the definitions, so that the first word could be taken as a potential genus term
Some Indonesian definitions

(1)  09500625-n ‘Pegasus’
seekor *kuda* bersayap dalam mitologi Yunani
one-CL horse winged in mythology Greece

“a winged *horse* in Greek mythology”

(2)  05316175-n ‘ocular muscle’
satu dari *otot-otot* kecil pada mata...
one of muscle-RED small at eye

“one of the small *muscles* of the eye”
Modifying definitions

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

- words between brackets, relating to domain: e.g. *(Ilmu komputer)* “(Computer science)”
- numerals: e.g. *satu* “one”, *tiga* “three”, *5* “five”
- determiners: e.g. *setiap* “every”, *sejenis* “a kind of”, *salah satu* “one of”, *beberapa* “some”, *berbagai* “various”, *segala* “all”
- relativizer: *yang* “which”
- prepositions: e.g. *untuk* “for”, *dari* “of”, *dalam* “in”
- other stop words: e.g. *seperti* “like”, *tentang* “about”, *termasuk* “including”, *biasanya* “usually”

The plural (reduplicated) form of the head was changed to its singular (non-reduplicated) form: e.g. *otot-otot* “muscles” → *otot* “muscle”

Punctuations, e.g. / ; , dividing two words were replaced as a space
Extracting relations

Using Python 3.4 and NLTK (Bird et al., 2009), for each genus term (adjusted first word) in each definition,

- Check if it is in Wordnet
  - If it is not in Wordnet, check if it is in the Indonesian dictionary KBBI http://badanbahasa.kemdikbud.go.id/kbbi/ (Alwi et al., 2008)
  - If the genus term is in Wordnet,
    ★ If it is the same as a lemma → synonym
    ★ If it is in one of the hyponym synsets → hyponym
    ★ If it is in one of the hypernym synsets → hypernym
    ★ If it is in one of the instance hypernym synsets → instance hypernym
    ★ If no relations can be extracted, check manually
Results and discussion

Out of 14,190 original lines of definitions from the Asian Wordnet project,
- 1,522 definitions (10.7%) were deleted
- The remaining 12,668 definitions (89.3%) consist of:
  - 10,549 definitions for nouns
  - 1,663 definitions for adjectives
  - 409 definitions for verbs
  - 47 definitions for adverbs

Out of 10,958 definitions for nouns and verbs which were examined for relations,
- Relations could be extracted from 6,257 definitions (57.1%)
- Problems were found in the remaining 4,701 definitions (42.9%)
  - Relations could not be extracted from 3,943 definitions
  - The first word in 747 definitions could not be found in Wordnet
  - 11 definitions do not have explicit semantic relations with the lemmas
## Relations extracted from lemmas and definitions

| Relation         | Number of synsets | Synset     | Example                                      |
|------------------|-------------------|------------|----------------------------------------------|
| Hypernym         | 5,451             | 00021939-n | *suatu objek buatan manusia*                |
|                  |                   | artifact   | “a man-made object”                         |
| Instance h.      | 549               | 02956500-n | *gedung DPR di AS*                          |
|                  |                   | Capitol    | “the government building in the US”         |
| Synonym          | 252               | 00004475-n | *makhluk hidup yang dapat*                  |
|                  |                   | organism   | “a living thing that can ...”               |
| Hyponym          | 5                 | 00029677-n | *sebuah fenomena yang berkelanjutan*        |
|                  |                   | process    | “a sustained phenomenon”                    |
| **Total**        | **6,257**         |            |                                              |

**Table:** Four relations extracted from 6,257 definitions
Problems found in extracting relations (1/2)

| Problem                                      | Synset      | Example                                      |
|----------------------------------------------|-------------|----------------------------------------------|
| No correct synset in Wordnet Bahasa          | 14350206-n  | **inflamasi** pada **syaraf**...             |
|                                              | myelitis    | “inflammation of the spinal cord”           |
| No explicit relation                         | 14573846-n  | **kehadiran** suatu **virus**...             |
|                                              | viremia     | “the presence of a virus...”                |
| Relational noun                              | 07603411-n  | **singkatan** dalam bahasa **Inggris**...    |
|                                              | choc        | “colloquial British abbreviation...”         |
| Compound                                     | 14364217-n  | **bekas luka** dari **sayatan pedang**       |
|                                              | sword-cut   | “a scar from a cut made by a sword”          |
| Incomplete definition                        | 00046344-n  | **tidak biasa atau berbahaya**               |
|                                              | stunt       | “not usual or dangerous”                     |

Table: Five types of problem found in extracting relations in 3,943 definitions when the genus candidates are in Wordnet
| Problem | Synset | Definition |
|---------|--------|------------|
| 1 Word in online KBBI data | 13436063-n automatic data processing | **pemrosesan** data secara otomatis “automatic data processing” |
| 2 Word not in online KBBI data | 09603258-n Pluto | **karakter** kartun anjing… “a cartoon character…” |
| 3 Compound | 07865105-n chili dog | **hot dog** dengan daging sapi… “a hotdog with chili con carne…” |
| 4 Derived word with negation | 14099050-n visual aphasia | **ketidakmampuan** memahami… “inability to perceive…” |
| 5 Incomplete definition | 14155506-n cystic fibrosis | **disebabkan kerusakan suatu gen** “caused by defect in a single gene” |
| 6 Incorrect definition | 00662589-v insure | **membagikan kawasan untuk** “allot regions for soldiers” |

**Table:** Six types of problem found in extracting relations in 747 definitions when the genus candidates are not in Wordnet.
Summary and future work

Things to be improved in Wordnet Bahasa:

1. Edit the incomplete Indonesian definitions
2. Delete the incorrect Indonesian definitions
3. Add new lemmas from KBBI and possibly derived words with negation
4. Add existing lemmas to the correct synsets
5. Sense tag the definitions (project from English)

Things to be improved in Wordnet in general:

1. Make definitions more informative, possibly add the hypernyms
2. Standardize definitions, possibly make some guidelines
3. Quality control in the translation process
Acknowledgments

- Thanks to Hammam Riza for giving us permission to use the Indonesian definitions from Asian Wordnet project
- Thanks to Randy Sugianto and Ruli Manurung for helping us with the data
- This research was partly supported by the Singapore MOE ARF Tier 2 grant *That’s what you meant: A Rich Representation for Manipulation of Meaning* (MOE ARC41/13)