Workshop/Hackathon for the Wordnet Bahasa

Francis Bond
Computational Linguistics Group
Linguistics and Multilingual Studies,
Nanyang Technological University

2014-10-25

Wordnet Bahasa Boleh!

NANYANG TECHNOLOGICAL UNIVERSITY

Overview

- Thanks for coming
  Everyone introduce themselves
- Goals of the workshop
- Schedule
- Current state of the Wordnet Bahasa (WB)
- More detailed discussion of goals

NANYANG TECHNOLOGICAL UNIVERSITY

Goals of the workshop

- Strengthen the WB community so that more people are confident to enhance the resources
- Improve and make accessible the infrastructure
- Discuss improvements in the content of the wordnet Bahasa including links to external resources
- Come up with a (totally non-binding) roadmap

NANYANG TECHNOLOGICAL UNIVERSITY

The Wordnet Bahasa

- Merge of five sources
  - Wordnet Bahasa (NTU: made from KAMI+FEM using pivot)
  - Wiktionary (NTU: OMW links to wiktionary/CLDR)
  - Malay Wordnet (MMU: hand aligning KIMD+PWN; Wikipedia)
  - Indonesian Wordnet (AWN: PWN+Eng-Ind lexicons; some revision)
  - Some hand additions from the NTU-Multilingual Corpus

NANYANG TECHNOLOGICAL UNIVERSITY

Numbers

Wordnet	Lang	Synsets	Words	Senses
Indonesian Wordnet	ind	27,506	30,358	57,560
Malaysian Wordnet	zsm	23,953	23,833	48,996
Wordnet Bahasa	ind	19,316	19,522	48,111
	zsm	19,347	19,572	48,181
Combined	ind	48,689	58,541	133,005
	zsm	38,736	45,664	114,025

The combined wordnet has 8,200 definitions;

85,315 of the senses are shared between Indonesian and Malay.

NANYANG TECHNOLOGICAL UNIVERSITY

Some stuff not fully merged

- New data from Lian Tze
- Clean up by Helmy, David (mainly removing bad POS)
- Derivational links
- Pronouns
- New words from NTU-MC
- Links to SIL Semantic Fields

NANYANG TECHNOLOGICAL UNIVERSITY

Applications

- Cross-lingual crossword (needs code release)
- NLTK
- Malay tweets
- Lumen Robot Friend Knowledge Base (http://lumen.hendyirawan.com)
- Malay Semantic Processing

- Malay/Indonesian NAF (NLP Annotation Format)
  - text to segmented POS tagged lemmatized text (di-, -nya, reduplication)
  - tagged text to concepts
  - concepts to disambiguated sense (UKB)
  - We have rough code, would like to merge with state-of-the-art
- Indonesian HPSG (David)
  Lexical and Structural Semantics

NANYANG TECHNOLOGICAL UNIVERSITY

Infrastructure

- Is sourceforge ok for everyone? Should we move to github?
- Current flow: .tab $\rightarrow$.sql $\rightarrow$ LMF/Lemon
  - Need to release conversion tools (NTU)
  - Need to release wn-gridx.cgi (NTU)
- How should we add new words
  - Go to sql as source?
  - Database dumps in version control? Or distributed copies?

NANYANG TECHNOLOGICAL UNIVERSITY

Content Improvements

- Can we have a single wordnet with lemmas marked as Malay/Indonesian/Other dialects?
  - It will still be possible to compile out separate wordnets (e.g. for language teaching)
  - How can we improve the Malay/Indonesian classification? In DPB/KBBI? In corpora (currently we use wikipedia)? FYP project?
    Can we link to DPB/KBBI entries?

- Non nvar entries
  - Pronouns (added)
  - Classifiers
    - Add as domain-usage of 06308436-n 'a word or morpheme used in some languages in certain contexts (such as counting) to indicate the semantic class to which the counted item belongs';
      Is this kind of meta-reference OK? is there a better link type
    - Link the classifier to the synset classified
      assume hyponyms classified unless marked (so mark shadows)

* Also link senses as necessary
* Cross-link between languages ($orang \approx nin$)?

- Add definitions: can we generate Malaysian and Indonesian?
  - generate from abstract?
  - translate?
  - Are there definition guidelines
- Can we legally pull in from KBBI/DBP/Wiktionary?
  - add to multidict (NTU)
  - Show only?
- Could we get some kind of funding for this?
  - Or work with translation departments (make it homework)?
  - Or crowd source?

- Things not in English
  - Add them (but with English gloss — ILI talk)
- Illustrations/Examples
- Derivational links (important in Malay)
  - We have some (not in yet)
  - hard to get the senses right
- Loan words/historical links
- Corpus Annotation

NANYANG TECHNOLOGICAL UNIVERSITY

- Quality control (still many errors)
  mark confidence (need to do — in source but not in converted wn)

Anything else?

- Have a roadmap discussion at the end: David will list ideas