2nd Wordnet Bahasa Workshop: expanding to other SEA languages

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

2016-01-15

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

NANYANG TECHNOLOGICAL UNIVERSITY

The Wordnet Bahasa

Wordnet	Lang	Synsets	Words	Senses
2014	ind	48,689	58,541	133,005
	zsm	38,736	45,664	114,025
2016	ind	52,051	70,621	143,152
	zsm	42,770	57,074	119,535

The combined wordnet has definitions for 12,663 synsets

- New corpora tagged with senses (sentiment to come)
  - Speckled Band
  - Kumo-no-Ito
- What are good Indonesian stories?

Infrastructure

- Presented at ACL demo session
  code still not quite released — rebuilding databases
- Current flow: .tab →.sql → 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?

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
  - Exclamatives
- Share definitions: can we generate Malaysian and Indonesian?
  - generate from abstract?
  - translate?
- Can we legally pull in from KBBI/DBP/Wiktionary?
- Could we get some kind of funding for this?

- 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
- Quality control (still many errors)

Anything else?

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

NANYANG TECHNOLOGICAL UNIVERSITY

Last Roadmap

A non-binding list of things we will be trying to do

X Infrastructure release by Luis et al., 25 Dec 2014
make the interface code available for everyone; need to coordinate witht he work on the ILI

O Merge new things by David and Lim, end of Nov 2014
plus new data from UI; defininitions already done

N Dialects by Luis and Lim, Nov 2014 present the wordnet as a single language

X Check frequency of words in Indonesian/Malay corpus by Ruli

X Shared definitions by Ruli

X Classifiers by David
first round done; not yet released

N Multidict by Mike

? Derivational links (human annotators + freq) ???

O Parsing definitions ??? (David)

O Reduplication (reduplicatable) in HPSG grammar

? Machine translation for Asian languages (work with the ASEAN project: Enya Kong)

- Look for funding: Singapore's Malay Heritage, Indonesia's Pusat Bahasa (Language Centre), maybe also also from Malaysia and Brunei

Thanks

- Thank our supporters
  - ICT Virtual Organization of ASEAN Institutes and NICT (ASEAN IVO)
  - Fuji-Xerox Corporation
  - Centre for Liberal Arts and Social Sciences (CLASS)

- Thanks to David and Takayuki for their tireless work in setting things up