Extending Wordnet Bahasa with External Resources

Lim Lian Tze[1] and Tang Enya Kong[2]

[1]KDU College Penang, Malaysia (liantze@gmail.com)

[2]Linton University College, Malaysia (enyagkong1@gmail.com)

WordNet Bahasa Hackathon/Workshop

Topics

Topics

Aligning Bilingual Dictionary to Princeton WordNet

UTMK@USM linguists; Lim and Hussein (2006)

Kamus Inggeris-Melayu Dewan (KIMD)

dot *n small round spot*, titik; (*appearing in large numbers on dress, leaf, etc*) bintik: ...

KIMD senses (manually) aligned to WordNet 1.6 senses

kimd (dot, n, 1, [small round spot, (appearing in large numbers on dress, leaf, etc)], ⟨*titik, bintik*⟩).
wordnet (110025218, 'dot', n, 1, [a very small circular shape]).

Malay WordNet synset

(***titik, bintik***; [a very small circular shape]).

Malay WordNet Prototype

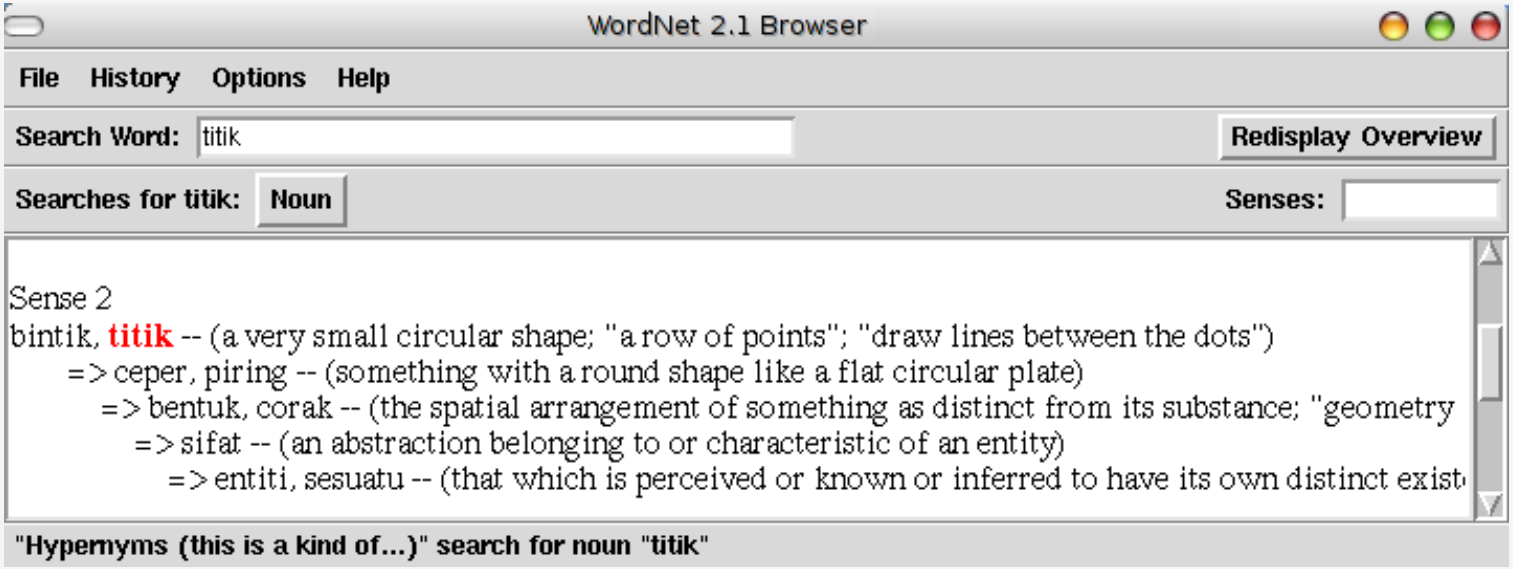
Nouns

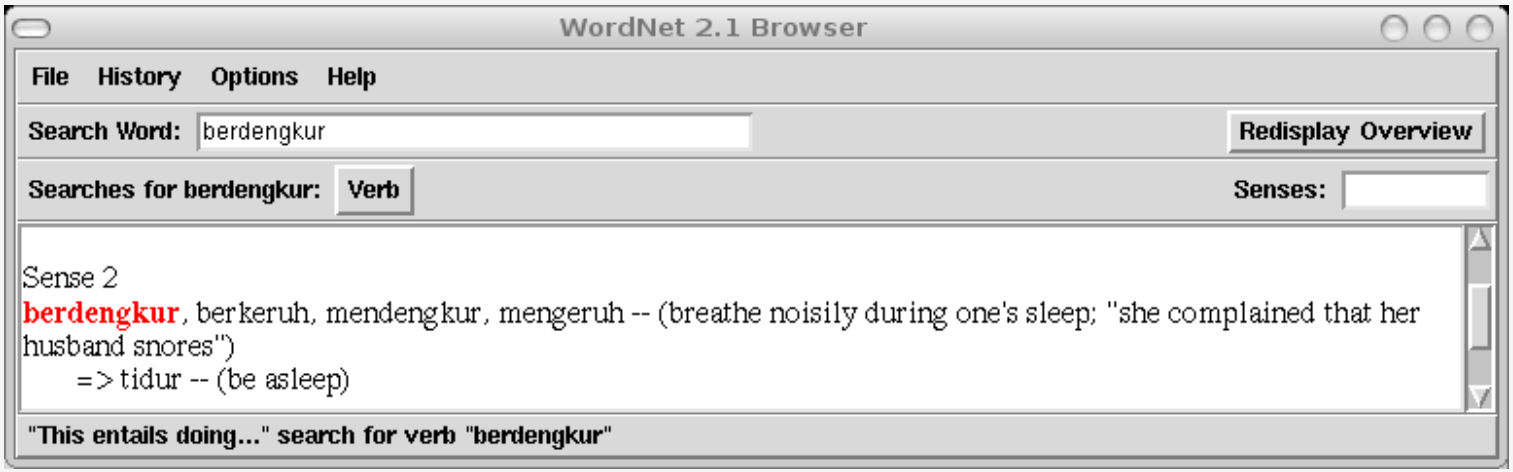
- 12429 synsets
- hypernymy/hyponymy
- holonymy/meronymy
  - part-of
  - member-of
  - substance-of

Verbs

- 5805 synsets
- hypernymy/troponymy
- cause
- entailment

Screenshots





Topics

Wikipedia Article Dumps

```
<page>
  <title>Marikh</title>
  <text>...
    {{Infobox Planet
    ....
    }}
    ...
    [[en:Mars]]
    [[es:Marte (planeta)]]
    ....
  </text>
</page>
<page>
  <title>Laut Kaspia</title>
  <text>...
    [[Kategori:Tasik di Eropah|Kaspia]]
    [[Kategori:Tasik di Rusia|Kaspia]]
    [[Kategori:Tasik di Asia|Kaspia]]
    ...
    [[en:Caspian Sea]]
    [[es:Mar Caspio]]
    ....
  </text>
```

Categories and Multilingual Translations

- `[[Kategori:Tasik di Eropah|Kaspia]]`
  (Category: European Lakes → Caspia)
- `[[es:Mar Caspio]]`
  Spanish (es) translation = Mar Caspio
  (Multilingual dictionary!)
- Spanish Wikipedia article about the Caspian Sea can be accessed at
  `http://es.wikipedia.org/wiki/MarCaspio`
  (Bilingual/multilingual comparable corpora!)

Adding Entries to WordNet Bahasa

1. For title for each Indonesian and Malaysian Wikipedia article, look up its corresponding English title in Princeton WordNet.
2. If only one synset is found, map the Indonesian/Malaysian title to it.
3. If multiple synsets are found, compare the hypernyms chain of each synset to the semantic type and categories of the Wikipedia article. The first synset whose hypernym chain contains the semantic type or one of the categories, is chosen as the synset to be mapped to.

Adding Entries to WordNet Bahasa (cont'd)

- 8480 new mappings
- 3725 new synsets
- 732 new Malay entries (i.e. used in both Malaysian and Indonesian)
- 2109 new Malaysian entries
- 5473 new Indonesian entries

Topics

The Wikidata Project

- Central storage for the structured data of Wikimedia projects including Wikipedia, Wikivoyage, Wikisource, and others
- All interlingual links will be moved to Wikidata eventually
- Hence some articles Wikipedia dumps are 'missing' these links!
- Wikidata HTTP API: `http://www.wikidata.org/w/api.php`

Example: Retrieving Multilingual Translations

- http://www.wikidata.org/w/api.php?action=wbgetentities&sites=mswiki&titles=serangan%20jantung&normalize&format=json&props=datatype|labels|descriptions|aliases&languages=ms|id|en
- XML response: (Other formats are possible)

```
<?xml version="1.0"?>
<api success="1">
  <normalized>
    <n from="serangan jantung" to="Penginfarkan miokardium" />
  </normalized>
  <entities>
    <entity id="Q12152" type="item">
      <labels>
        <label language="ms" value="Penginfarkan miokardium" />
        <label language="id" value="Serangan jantung" />
        <label language="en" value="heart attack" />
      </labels>
      <descriptions>
```

Example: Retrieving Multilingual Translations (cont'd)

```
        <description language="en" value="interruption of blood
            supply to a part of the heart" />
      </descriptions>
      <aliases>
        <alias language="en" value="acute myocardial infarction" />
        <alias language="en" value="AMI" />
        <alias language="en" value="MI" />
        <alias language="en" value="myocardial infarction" />
        <alias language="ms" value="Serangan jantung" />
      </aliases>
    </entity>
  </entities>
</api>
```

- Mappings to WordNet synsets done, but not yet checked and/or officially added
- Cursory glance: lots of identical lexicalisation to English

Topics

Extending Specific Hierarchies

- Cultural-specific concepts e.g. clothing items, food and dishes, etc…
- E.g. Malay Wikipedia category ‘Masakan Malaysia’ (Malaysian dishes)
  - ayam golek, buah keras, acar timun, ikan bakar, mi rebus, apam…
- Some items (e.g. ‘Ayam penyet’) aren’t in Indonesian nor Malay Wikipedia, but are listed in *English* Wikipedia!
- How can we tell if the title of an Wikipedia article is a foreign language word?

Kamus Dewan

- The main Malay monolingual dictionary in Malaysia
- Published by Dewan Bahasa dan Pustaka
- Currently annotating contents with TEI to give structure (as part of another project)
- Allows for easier, more targeted searches
- (Still in progress – expected completion Nov 2014)
- Not open-source…but can be used for research by arrangement e.g. can be used for some quick additions to WordNet Bahasa

Lexical Items

- Derived words as subentries of root word – very rich!
- Lots of MWEs including peribahasa (idioms), *usually* fixed word order and little morphosyntactic process
- No POS!
  - Do something based on definition text?

Extending Specific Hierarchies

Search by definitions (Start with the simple ones)

- Definition = ‘sj (masakan | makanan) …’ (A dish…)
  - bamiyah, besengek, caca, dalca, gudeg…
  - Other posibilities: dances, articles of clothing, musical instruments…
- (KD definition texts themselves cannot be used in WordNet Bahasa – copyright issues)
- (Mine from Wikipedia? CC-BY-SA/GFPL)
- ‘Flat’ hierarchy for starters

Names of Flora & Fauna

- KD contains a huge number of binomial nomenclature (Latin names) for flora & fauna
- Match up Malay names with English translations via Latin names
- Some might not yet be in Princeton WordNet
- Many may not even have English equivalent names!

Penjodoh Bilangan (Classifiers)

- KD indicates classifiers
- batang: penjodoh bilangan bagi benda yang panjang-panjang (classifiers for longish thigs)
- But how to create/project relations between ‘batang’ synset and ‘long things’?
- And exceptions?
  - biji: penjodoh bilangan (bagi benda kecil dll) (for small things)
  - . . . but ‘sebiji meriam’ (a cannon)!

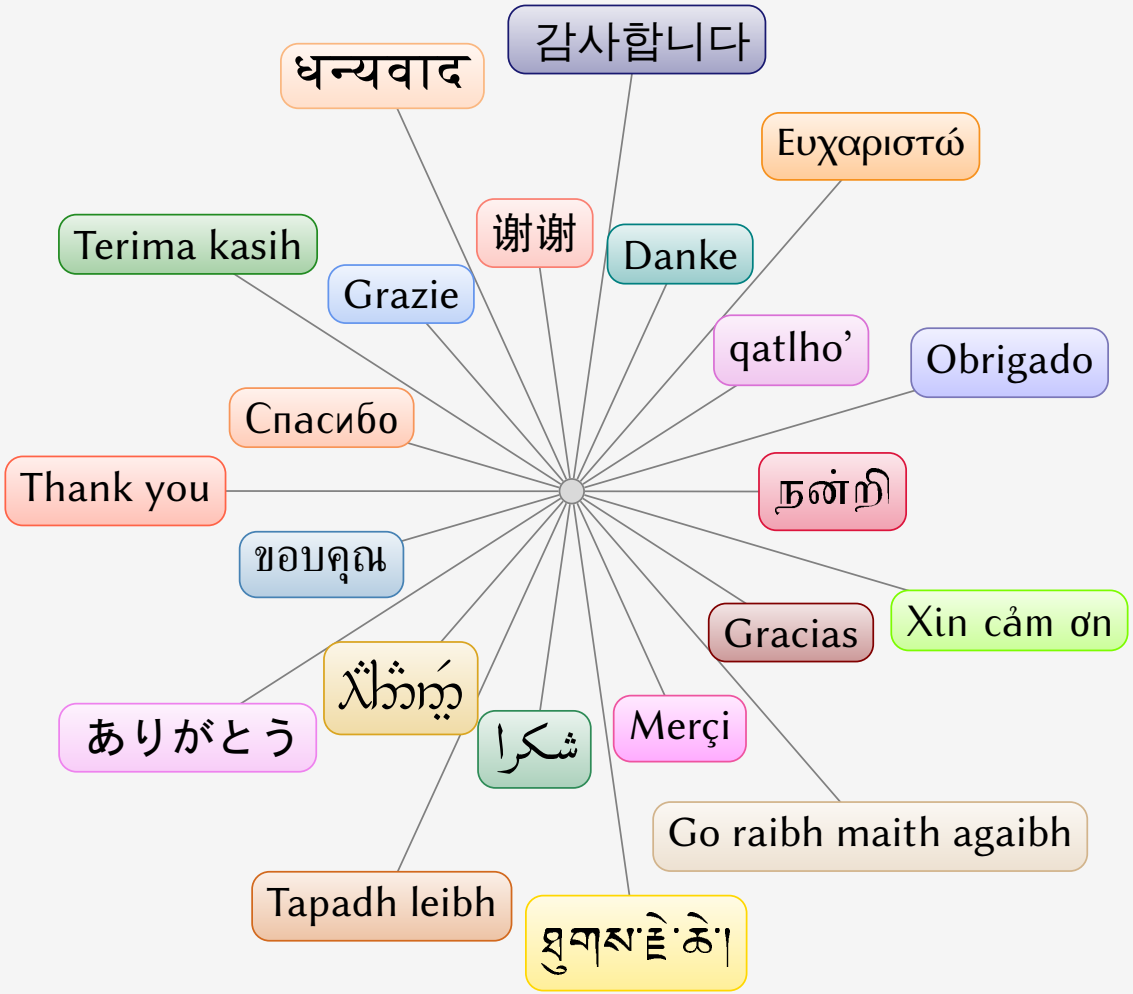
Named Entities

- Collect/extract gazetteer lists of named entities in Malay
  - Existing gazetteer lists: location names, Wikipedia categories…
  - Rule-based: list of prefix/suffix
  - Corpus-based: news articles?

Further Processing

- More advanced processing (including discovering relations) may (should) be possible in future
  - Study the definition text patterns in Kamus Dewan
  - Greater availability of of Malay corpus

Thank You!


धन्यवाद
감사합니다
Ευχαριστώ
Terima kasih
谢谢
Danke
Grazie
qatlho'
Obrigado
Спасибо
Thank you
நன்றி
ขอบคุณ
Gracias
Xin cảm ơn
ありがとう
شكرا
Merçi
Go raibh maith agaibh
Tapadh leibh
ཐུགས་རྗེ་ཆེ།