Fast Prototyping of a Malay WordNet System

LIM Lian Tze Nur HUSSEIN

Unit Terjemahan Melalui Komputer (UTMK)
Universiti Sains Malaysia
Penang, Malaysia

Language, Artificial Intelligence and Computer Science for Natural Language Processing Applications Workshop

Outline

Outline

NLP for Malay Language

- WANTED: Malay lexical resource
- Sense level information
- Relation between senses
- WordNet systems for different languages
- No Malay!

Let's Build Malay WordNet!

Let's Build Malay WordNet!

But...

This may (will) take years!

Let's Build Malay WordNet!

But...

This may (will) take years!

So...

A quick prototype first perhaps?
Also, to explore architecture and functions of a WordNet system

Outline

前人种树，后人乘凉。

"Forebears plant the trees, descendants enjoy the shade."

"Bamboo" artwork used by permission of Keng Lye.

Existing Data at UTMK

Kamus Inggeris-Melayu Dewan (KIMD): a bilingual English-Malay dictionary

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

Existing Data at UTMK

Kamus Inggeris-Melayu Dewan (KIMD): a bilingual English-Malay dictionary

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]).

Resources from English WordNet (Princeton)

Data

- lexicographer files (human editable)
- database files (generated)

Tools

- WordNet Browser (Windows, GNU/Linux)
- WordNet lookup API (C libraries)
- GRIND (lexicographer $\rightarrow$ database files)

Constructing Malay WordNet (Overview)

- Define synsets
- Generating lexicographer files
  - Copying relations from English WordNet
  - Group synsets in files by semantic fields
  - Format lexicographer files as required
- Run GRIND on lexicographer files
- Browse Malay WordNet database files in WordNet Browser!

Creating Synsets

- Use KIMD–WordNet alignment
- No Malay gloss/definition text, so re-use English gloss for now

KIMD–WordNet alignment

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]).

Creating Synsets

- Use KIMD–WordNet alignment
- No Malay gloss/definition text, so re-use English gloss for now

KIMD–WordNet alignment

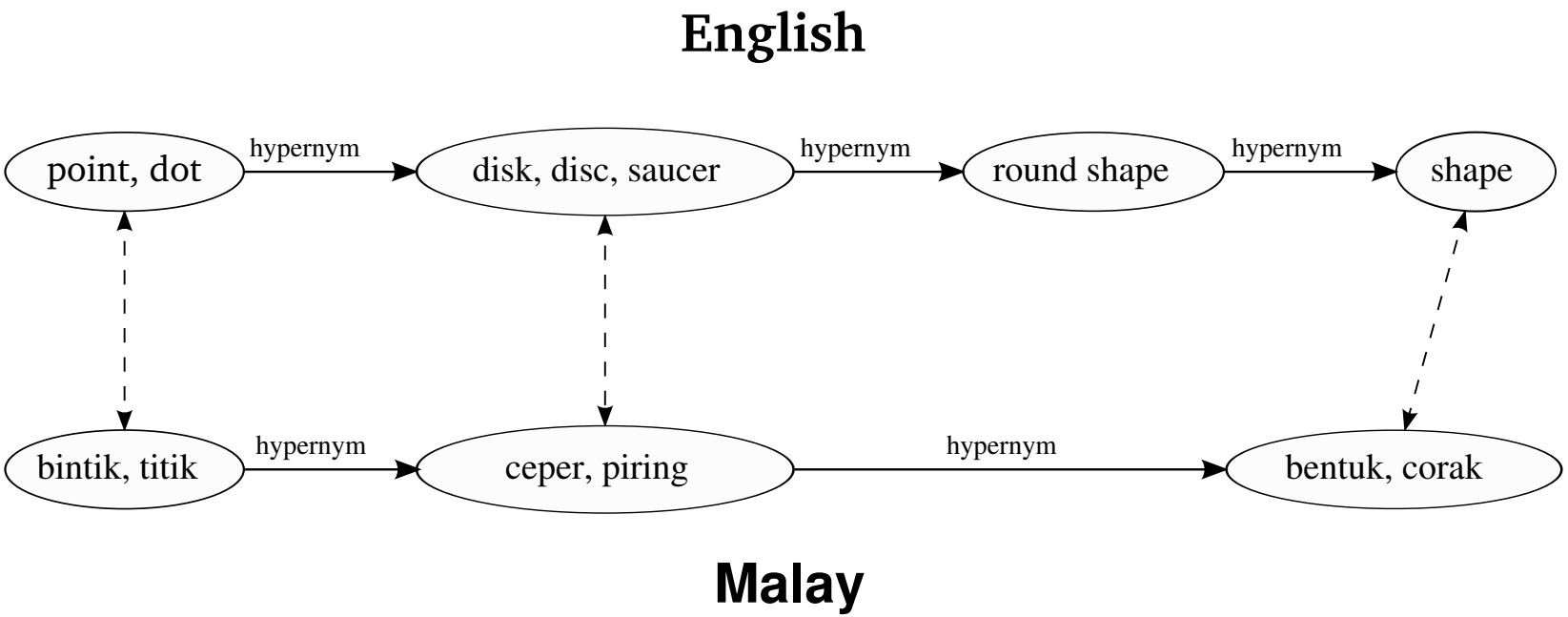
kimd (dot, n, 1, [small round spot, (appearing in large numbers on dress, leaf, etc)], ⟨*titik, bintik*⟩).
wordnet (110025218, 'dot', n, 1, [a very small circular shape]).

Malay WordNet synset

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

Relations between Synsets

- Refer to relations between English WordNet synsets
- Copy selected relations over to Malay WordNet where possible



Semantic Fields

- Synsets are placed in separate lexicographer files according to semantic fields
- Refer to aligned English WordNet synsets

noun.shape

wordnet (110025218, 'dot', n, 1, [a very small circular shape]).

Semantic Fields

- Synsets are placed in separate lexicographer files according to semantic fields
- Refer to aligned English WordNet synsets

noun.shape

wordnet (110025218, 'dot', n, 1, [a very small circular shape]).

noun.shape

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

Lexicographer File Snippet

. . .

{ titik_pertemuan, noun.Tops:bentuk,@ (a connecting point at which several lines come together) }
{ tempat_sambung, noun.Tops:bentuk,@ (the shape or manner in which things come together and a connection is made) }
{ bintik, titik, ceper,@ (a very small circular shape) }
{ pori, ruang,@ (any tiny hole admitting passage of a liquid (fluid or gas)) }
{ pohon, rajah,@ (a figure that branches from a single root) }
{ susuh, taji, juluran,@ (any sharply pointed projection) }

. . .

Outline

Malay WordNet Prototype

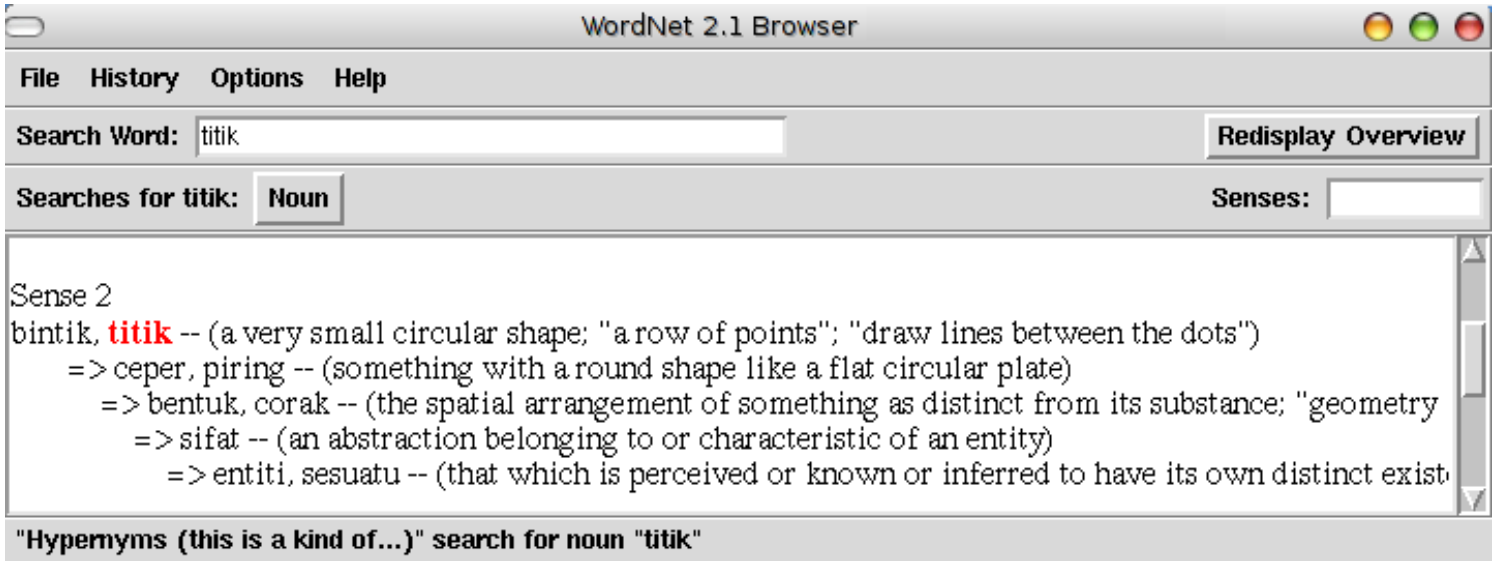
Nouns

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

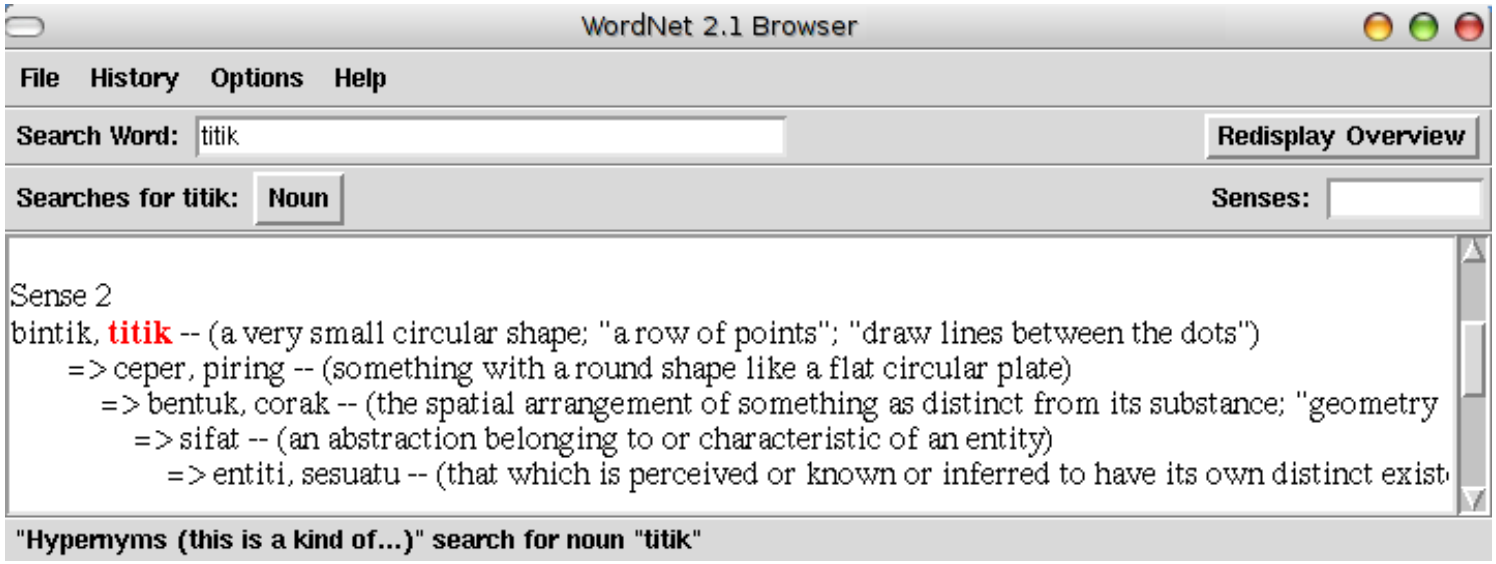
Verbs

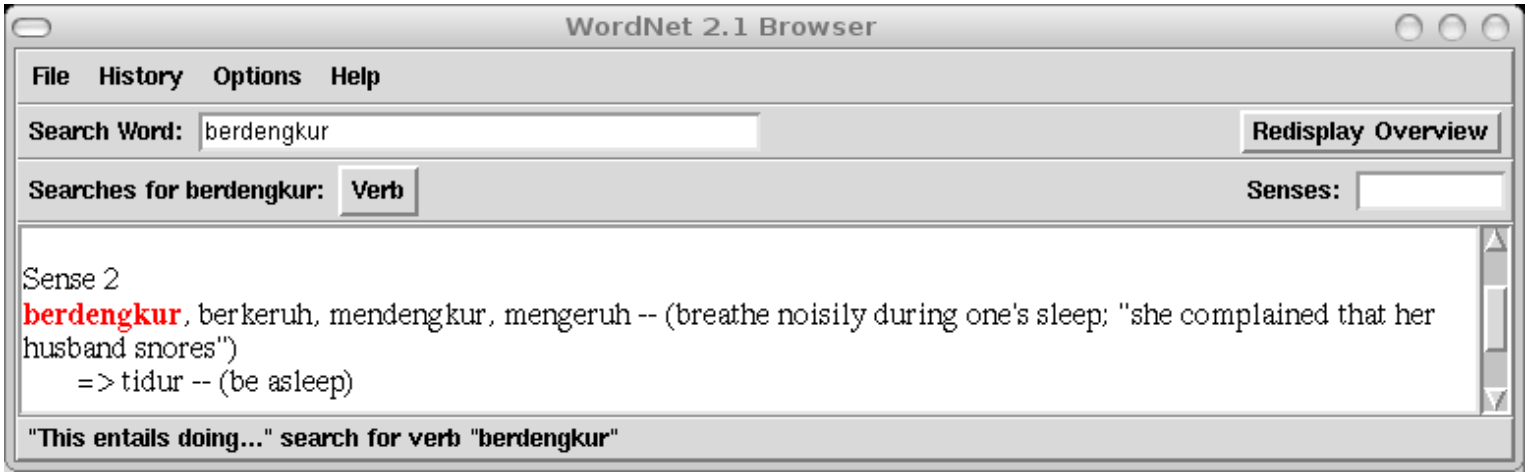
- 5805 synsets
- hypernymy/troponymy
- cause
- entailment

Screenshots



Screenshots





Outline

Problems and Issues
(It's a hacky prototype after all!)

- Unsuitable Malay entries
  - KIMD is uni-directional, English-Malay dictionary
  - e.g. "absentee" "*orang, anggota, dan lain-lain yang tidak hadir*"
- No Malay gloss (yet)
- No Malay verb frames (yet)
- Derivation of Malay synsets
  - merging/splitting synsets
  - words specific to Malay language/culture e.g. "*baju kurung*"
- Not all relations for English synsets make sense in Malay

Future Work

- Use Malay monolingual lexicon
- Verb frames for Malay
- Proper "learning" of relations e.g. machine learning techniques
- Full localisation for Malay users (GUI, terms and glossary etc)
- Alignment to EuroWordNet

Summary

- Creating a WordNet system is **HARD WORK**.
- Lexicographic expertise is sorely needed
- We achieved some understanding about the software tools and process to help lexicographers compile the necessary data

Thank You for Your Attention!