Fast Prototyping of a Malay WordNet System

LIM Lian Tze and Nur HUSSEIN
Computer Aided Translation Unit
School of Computer Sciences
Universiti Sains Malaysia
Penang, Malaysia
{liantze,hussein}@cs.usm.my

ABSTRACT

This paper outlines an approach to produce a prototype WordNet system for Malay semi-automatically, by using bilingual dictionary data and resources provided by the original English WordNet system. Senses from an English-Malay bilingual dictionary were first aligned to English WordNet senses, and a set of Malay synsets were then derived. Semantic relations between the English WordNet synsets were extracted and re-applied to the Malay synsets, using the aligned synsets as a guide. A small Malay WordNet prototype with 12429 noun synsets and 5805 verb synsets was thus produced. This prototype is a first step towards building a full-fledged Malay WordNet.

KEYWORDS

WordNet, Malay, lexical knowledge base, fast prototyping

1 INTRODUCTION

Traditional dictionaries compile lexical information about word meanings by listing them alphabetically by the headwords. While this arrangement is convenient for a human reader who wants to look up the meaning of a word, it does not provide much information about explicit semantic relations between words, besides the usual synonyms and antonyms.

WordNet [6, 8] is a lexical database system for English words, designed based on psycholinguistic principles. It organises word meanings (senses) on a semantic basis, rather than by the surface morphological forms of the words. This is done by grouping synonyms into sets, and then defining various relations between the synonym sets (synsets). Some examples of the semantic relations defined include hypernymy (the `is-a` relation) and meronymy (the `part-of` relation).

Armed with such semantic relations, WordNet became an invaluable resource for natural language processing (NLP) researchers in tackling problems like information retrieval, word sense disambiguation, and question answering. As the original WordNet contains only English words, there have been efforts to create WordNet-like systems for other languages. See the Global WordNet Association's website [4] for a list of such projects.

Currently, no WordNet-like lexical database system exist for the Malay language. Such a resource will be useful indeed for NLP research involving Malay texts. While the construction of a complete WordNet-like system is a daunting undertaking which requires lexicographic expertise, it is possible to build a prototype system semi-automatically using resources accessible at our site. The prototype Malay WordNet system and data can then be further scrutinised, fine-tuned and improved by human lexicographers.

The main aim of developing this prototype was to explore the design and tools available in a WordNet system, rather than a full attempt to develop high quality Malay WordNet data. Therefore, the methods we adopted are not as extensive as other efforts in constructing non-English WordNets, such as the work reported in [1, 2].

2 METHODOLOGY

We describe how a prototype Malay WordNet can be constructed semi-automatically using a English-Malay bilingual dictionary, the original English WordNet, and alignments between the two resources.

The developers of the English WordNet, the Cognitive Science Laboratory at Princeton University, have made available some useful tools that allow the custom development of WordNet-like systems [7]. They include:

- English WordNet database files,
- WordNet Browser, a GUI front-end for searching and viewing WordNet data,
- WordNet database search functions (as C library functions),
- `GRIND`, a utility tool for converting lexicographer input files into WordNet database files.

If lexicographer input files for Malay words can be created following the required syntax, `GRIND` can be used to process them to produce Malay WordNet database files, to be viewed using the WordNet browser. This can be done

by first establishing a set of Malay word synsets and the semantic relations between them, and then generating the lexicographer files.

2.1 Malay Synsets

Kamus Inggeris Melayu Dewan (KIMD) [5] is an English-Malay bilingual dictionary and provides Malay equivalent words or phrases for each English word sense. Linguists at our research group had previously aligned word senses from KIMD and WordNet 1.6. Not all KIMD and WordNet 1.6 senses were included; only the more common ones were processed.

Here are some example alignments for some senses of *dot*, *consolidation* and *integration*:

Listing 1: Aligned senses of *dot*

```
kimd (dot, n, 1, 0, [small round spot, small circular shape], <titik,
     bintik> ).
wordnet (110025218, 'dot', n, 1, 0, [a very small circular shape] ).
```

Listing 2: Aligned senses of *consolidation*

```
kimd (consolidation, n, 1, 0, [act of combining, amalgamating], <
     penggabungan, penyatuan>).
wordnet (105491124, 'consolidation', n, 1, 0, [combining into a
     solid mass]).
wordnet (100803600, 'consolidation', n, 2, 0, [the act of combining
     into an integral whole]).
```

Listing 3: Aligned senses of *integration*

```
kimd (integration, n, 1, c, [act of c. (combining into a whole)], <
     penyepaduan, pengintegrasian>).
wordnet (100803600, 2, 'integration', n, 2, 0, [the act of combining
     into an integral whole]).
```

(The 9-digit number in each English WordNet sense above is a unique identifier to the synset it belongs to.)

A set of Malay synsets may be approximated based on the KIMD–WordNet alignment using Algorithm 1.

Algorithm 1 Constructing Malay synsets

```
for all English synset es do
  ms-equivs ⇐ empty //list of Malay equivalent words
  ms ⇐ null //Equivalent Malay synset
  for all s ∈ {KIMD senses aligned to es} do
    add Malay equivalent(s) of s to ms-equivs
  end for
  ms ⇐ new synset containing ms-equivs
  Set ms to be equivalent Malay synset to es
end for
```

Following this algorithm, the following Malay synsets are derived from the sense alignments in Listings 1–3. The corresponding English WordNet synsets are also shown:

- (*titik, bintik*)
  (110025218: point, dot; [a very small circular shape])
- (*penggabungan, penyatuan*)
  (105491124: consolidation; [combining into a solid mass])
- (*penggabungan, penyatuan, penyepaduan, pengintegrasian*)
  (100803600: consolidation, integration; [the act of combining into an integral whole])

2.2 Synset Relations

For this fast prototyping exercise, we have decided to create semantic relations between the Malay synsets based on the existing relations between their English equivalents. Algorighm 2 shows how this can be done.

Algorithm 2 Creating relations between Malay synsets

```
Require: lookup_ms(es):
  returns Malay synset equivalent to English synset es
Require: lookup_es(ms):
  returns English synset equivalent to Malay synset ms
Require: get_target(R, es):
  returns target (English) synset of English synset es for
  relation R
  for all Malay synset ms do
    es ⇐ lookup_es(ms)
    for all relation R with a pointer from es do
      ms′ ⇐ null
      es′ ⇐ es
      if R is transitive then
        repeat
          es′ ⇐ get_target(R, es)
          ms′ ⇐ lookup_ms(es′)
        until es′ = null or ms′ ≠ null
      else
        es′ ⇐ get_target(R, es)
        ms′ ⇐ lookup_ms(es′)
      end if
      if ms′ ≠ null then
        add (R, ms′) to list of relations that applies to ms.
      end if
    end for
  end for
```

As an example, the hypernymy relation holds between the English synsets (*point, dot*) and (*disk, disc, saucer*). Therefore, a hypernymy relation is established between the corresponding Malay synsets (*bintik, titik*) and (*ceper, piring*).

However, while searching for target synsets for a relation R, it is always possible that there is no Malay equivalent for an English synset. If R is transitive, as are hypernymy and meronymy, we continue to search for the next target synset in the transitive relation chain, until we reach the last English synset in the chain.

To illustrate, consider the English and Malay synsets in Figure 1. The English synset (*disk, disc, saucer*) has the hypernym (*round shape*), which in turn has the hypernym (*shape, form*). While (*round shape*) does not have a corresponding Malay synset in our data, (*shape, form*) does have one as (*bentuk, corak*). Therefore, a hypernymy relation is established between (*ceper, piring*) and (*bentuk, corak*).

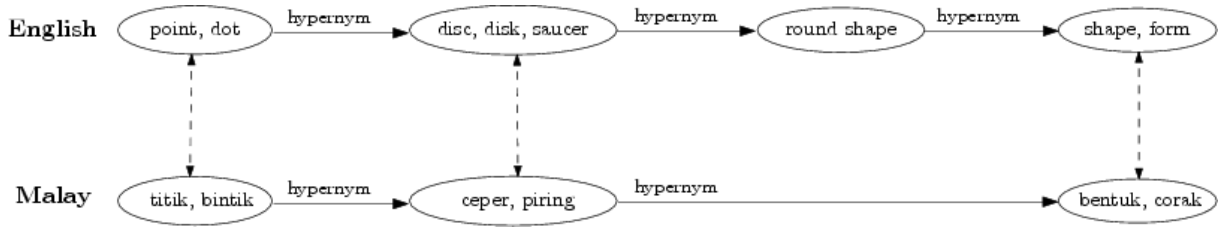

Figure 1: English and Malay synsets forming a hypernymy chain

2.3 Lexicographer Files

WordNet systems organise synsets of different syntactic categories, i.e. nouns, verbs, adjectives and adverbs, separately. In addition, the English WordNet also assign semantic fields to the synsets, such as `noun.location`, `noun.animal` and `verb.emotion`. Synsets of different categories are to be stored in separate lexicographer files, the names of which correspond to their semantic fields.

For each Malay synset identified in section 2.2, we look up f, the semantic field of its equivalent English synset. The Malay synset, together with its relations and target synsets, is then appended to the lexicographer file f.

3 IMPLEMENTATION

The procedures described in sections 2.2 and 2.3 were implemented as a suite of tools called `LEXGEN` in C and Java. As a first step, only noun and verb synsets were processed with `LEXGEN`. Since KIMD does not provide Malay glosses, `LEXGEN` reuses glosses from English WordNet. The resulting lexicographer files were then put through `GRIND`, producing a small Malay WordNet system.

4 RESULTS

The prototype Malay WordNet system currently contains 12429 noun synsets and 5805 verb synsets. Its small coverage of the English WordNet (81426 noun synsets and 13650 verb synsets) is understandable as only a subset of KIMD and WordNet senses was used in the earlier alignment work. The prototype also includes the hypernymy, hyponymy, troponymy, meronymy, holonymy, entailment and causation relations. Figure 4 shows the Malay synset (*bintik, titik*) and its hypernyms as viewed in the WordNet Browser.

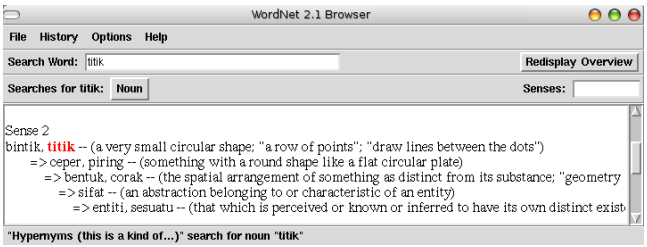

Figure 2: Malay WordNet as viewed in Browser

5 DISCUSSION

The Malay WordNet prototype is adequate for demonstrating what a WordNet system has to offer for Malay. This is especially helpful to give a quick preview to users who are not yet familiar with the WordNet or lexical sense organisation paradigm. However, as acknowledged at the very beginning, its current quality is far from satisfactory.

Part of the problem is in the dictionary used. The KIMD–WordNet alignment work was part of a project to collect glosses for English word senses from different dictionaries. As such, the suitability of Malay equivalents to be lemmas were not the main concern: all Malay equivalents were simply retained in the alignment files.

This leads to unsuitable Malay WordNet synset members in some cases: since KIMD is a unidirectional English to Malay dictionary, not all Malay equivalents it provides can stand as valid lemmas. For example, KIMD provides *orang, anggota, dan lain-lain yang tidak hadir* (literally 'person, member, etc. who are not present') as the Malay equivalent for English *absentee*. While this is valid as a Malay gloss or description for the synset, it is unsuitable to be a member lemma of a synset.

In addition, we also lack Malay gloss information for the Malay synsets as these were not provided in KIMD. The prototype Malay WordNet, therefore, is forced to have English text as glosses, intead of Malay glosses.

We also noted that the English WordNet provide verb frames, e.g. Somebody —s something for a sense of the verb *run*. The first problem is that we have yet to establish a list of verb frames for Malay. Secondly, even if there were, there is not necessarily a one-to-one mapping between the English and Malay verb frames. Thirdly, as the English verb frames are hard-coded into `GRIND` and WordNet, extensive re-programming would be required to use these utilities on different languages. Therefore, we have not attempted to handle Malay verb frames for this prototype.

`GRIND` imposes a maximum of sixteen senses per word form in each lexicographer file. This might be a problem if there are Malay words that are very polysemous. Possible alternatives are:

- further split the senses into different lexicographer files so that each file would not contain more than sixteen senses of the same word,
- aim for coarser sense distinctions, or
- re-program `GRIND`.

Finally, the derivation of Malay synsets from the KIMD–WordNet alignments may be flawed. This is because multi-

ple KIMD senses may be aligned to a WordNet sense, and vice versa. Referring back to Listing 2 and the list of Malay synsets at the end of Section 2.1, we see that the Malay words *penggabungan* and *penyatuan* from *one* KIMD sense now appear in *two* synsets. To non-lexicographers, such as the authors of this paper, it is unclear how this situation should be handled. Are there now two senses of *penyatuan* and *penggabungan*, or should the Malay synsets (*penggabungan, penyatuan*) and (*penggabungan, penyatuan, penyepaduan, pengintegrasian*) be merged? Since there are opinions that the English WordNet is too fine-grained, the synsets can perhaps be merged to avoid the problem for Malay WordNet. Nevertheless, we think a lexicographer would be more qualified to make a decision.

6 FUTURE WORK

The aim of work on the prototype Malay WordNet is but to explore the architecture and software tools required in a WordNet system. Future work will focus more on systematically compiling lexical data for a Malay WordNet system by lexicographers and linguistic experts. We highlight some issues of interest here.

- A Malay monolingual lexicon or dictionary should be used to determine the Malay synsets, the gloss text for each synset, as well as the synset's semantic field.
- The semantic fields are hard-coded into `GRIND` and WordNet. Therefore, if we are to have localised semantic fields in Malay, e.g. `noun.orang` (`noun.person`) and `noun.haiwan` (`noun.animal`), or to add new fields, `GRIND` and WordNet will need to be modified.
- Semantic relations need to be defined between the Malay synsets. This may be aided by machine learning strategies, such as those used in [1], besides human efforts.
- A list of Malay verb frames need to be drawn up and assigned to each verb sense.
- Currently, the Malay word senses are ordered at random. Ideally, the senses should be numbered to reflect their usage frequency in natural texts. A sense-tagged Malay corpus will help in this, as was done in the English WordNet [7, p.112].
- It would also be interesting to align the Malay WordNet to EuroWordNet [3], which contains wordnets for several European languages. As EuroWordNet is aligned to English WordNet 1.5, some re-mapping would have to be performed if we wish to re-use the KIMD–WordNet alignment, or the prototype, as a rough guide.

7 CONCLUSION

Creating a new set of Wordnet lexicographer files from scratch for a target language is a daunting task. A lot of work needs to be done in compiling the lexicographer input files and identifying relations between synsets in the language. However, we have been successful in rapidly constructing a prototype Malay Wordnet by bootstrapping the synset relations off the English Wordnet. Hopefully, this will lay the foundation for the creation of a more complete Malay Wordnet system.

REFERENCES

[1] J. Atserias, S. Climent, X. Farreres, G. Rigau, and H. Rodríguez. Combining multiple methods for the automatic construction of multilingual wordnets. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'97)*, Tzigov Chark, Bulgaria, 1997.

[2] I. Azarova, O. Mitrofanova, A. Sinopalnikova, M. Yavorskaya, and I. Oparin. RussNet: Building a lexical database for the Russian language. In *Proceedings of Workshop on WordNet Structures and Standardisation and How this affect Wordnet Applications and Evaluation*, pages 60–64, 2002.

[3] EuroWordNet. Eurowordnet: Building a multilingual database with wordnets for several European languages, 2006. URL `http://www.illc.uva.nl/EuroWordNet/`. Last accessed September 15, 2006.

[4] Global WordNet Assoc. Wordnets in the world, 2006. URL `http://www.globalwordnet.org/gwa/wordnet_table.htm`. Last accessed September 15, 2006.

[5] A. H. Johns, editor. *Kamus Inggeris Melayu Dewan*. Dewan Bahasa dan Pustaka, Kuala Lumpur, Malaysia, 2000.

[6] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography (special issue)*, 3(4):235–312, 1990.

[7] R. I. Tengi. Design and implementation of the wordnet lexical database and searching software. In C. Fellbaum, editor, *WordNet: An Electronic Lexical Database*, chapter 4, pages 105–127. MIT Press, Cambridge, Massachusetts, 1998.

[8] WordNet. WordNet: a lexical database for the English language, 2006. URL `http://wordnet.princeton.edu/`. Last accessed September 15, 2006.