Building an Ontology-Based Multilingual Lexicon for Word Sense Disambiguation in Machine Translation

Lian-Tze Lim and Tang Enya Kong

Unit Terjemahan Melalui Komputer,
Universiti Sains Malaysia,
11800 Minden,
Penang, Malaysia.
{liantze, enyakong}@cs.usm.my

Abstract. Word sense disambiguation (WSD) requires the establishment of a list of the different meanings of words. WSD efforts in machine translation require, in addition, the equivalent translation words in target languages. To facilitate WSD in machine translation systems, we propose the construction of an ontology-based multilingual lexicon, from various existing language resources, as an alternative to existing hierarchical lexicons such as *WordNet* and *Roget's Thesaurus*. The information in the lexicon to be constructed can also be used for other natural language processing tasks.

1 Introduction

In natural language, words having different meanings in different contexts are said to be ambiguous. While it comes naturally to humans, deciding what an ambiguous word means in a particular discourse can be very problematic for machines. As an example, consider the English word *log*. A computer might wrongly translate the English sentence:

*The computer **logs** have been deleted.*

into the Malay sentence

**__Balak__ komputer telah dipotong.*

or literally, **the computer **wood** has been cut.*!

Word sense disambiguation (WSD) refers to the task of determining the correct meaning or sense of an ambiguous word in context [1]. This requires first establishing a list of all different meanings (senses) for all the words under consideration. Disambiguation is then performed by evaluating the context of an occurrence of an ambiguous word and the sense entries in the said list, in order to assign the correct sense to the word occurrence under consideration [2]. The selection of equivalent words in a target language, to translate ambiguous words in a source language, as in the example above, is often termed word selection.

Many researchers turn to *WordNet* [3] as a linguistic resource for WSD, due to its broad coverage, rich lexical information, and free availability. However, the sense distinctions in *WordNet* are often deemed too fine-grained for practical natural language processing tasks [2]. Some researchers, such as in [1], attempted to alleviate this by "lumping" together English *WordNet* senses that are translated to the same Chinese words. Others (e.g. the authors of [4]) constructed their own lexical knowledge-bases suited to their needs.

To better facilitate WSD for machine translation, we propose to use an ontologybased multilingual lexicon, which will contain various linguistic information, as part of our language resources.

2 Building an Ontology-based Multilingual Lexicon

We first give a brief overview of ontologies and the use of taxonomic structures in lexical resources, before outlining how an ontology-based multilingual lexicon can be constructed.

2.1 Taxonomies and Ontologies

An ontology is an "explicit formal specifications of the terms in the domain and relations among them" [5]. It defines concepts, terms and vocabularies in a domain, and also the relationship among these concepts. Concepts are organised in a taxonomic structure, with subclasses inheriting properties and specialising from superclasses. Current semantic web technologies also have the added capability of inferring new facts from old facts already captured in the ontology. An ontology, together with a set of instances of the classes or concepts defined, constitute a knowledge base about the domain being described [6].

Using taxonomies in lexical resources is not a new idea. *Roget's Thesaurus* [7] groups words with similar meanings in hierarchies (with few number of levels) of classes and sections, while *WordNet* is well-known for its "is-a" relations (amongst other types of relations) between "synsets", or groups of synonymous words. However, *Roget's Thesaurus* does not include the definition of words. In fact, words in a group are merely related, not synonymous. In addition, words under a common heading can be of different syntactic categories. On the other hand, while *WordNet* uses different approaches in categorising words of different syntactic categories, Kilgariff and Yallop [8] argued that *WordNet*'s hierarchical structure cannot be used if one wishes to move from a fine-grained approach to a coarse-grained one.

The main aim of *Roget's Thesaurus* is to help writers choose the appropriate word [8], whereas *WordNet* was constructed based on psycholinguistic principles [3]. Neither are traditional dictionaries (in book- or electronic-form) perfect resources for WSD work [2]. Therefore, we propose the construction of an alternative ontology-based multilingual lexicon.

2.2 Construction of the Lexicon

The construction of an ontology-based multilingual lexicon involves four tasks:

- building the taxonomic structure of the ontology,
- preparing lexical entries and the information they contain,
- categorising the lexical entries under the appropriate semantic classes in the ontology, and
- specifying suitable relations among the lexical entries.

The Taxonomy. We construct our taxonomic structure of the lexicon based on *GoiTaikei* [9], an electronic Japanese lexicon. *GoiTaikei* contains around 300,000 Japanese words categorised under 3000 semantic classes in three hierarchies: general nouns, proper nouns, and "phenomenons" (verbs, adjectives and adverbs). The Japanese words are marked with part-of-speech information and the semantic classes they belong to, while words in the "phenomenon" hierarchy are organised as a valency dictionary with selectional restrictions. The ontology hierarchies in *GoiTaikei* are particularly desirable, since *GoiTaikei* was developed for use with a machine translation system.

The label of each semantic class in *GoiTaikei* was translated at Unit Terjemahan Melalui Komputer (UTMK) to English, and the hierarchical structure recreated as an Ontology Web Language (OWL) [10] file[1]. We used *Protégé 2000* [11], an ontology editor, for this purpose.

The Lexical Entries. Each lexical entry represents a distinct sense of a word, and contains the following information:

- a Malay word form,
- its definition text, in Malay,
- equivalent word(s) in other languages,
- part-of-speech,
- etymology,
- morphological structure,
- syllable segmentation and IPA,
- example sentence(s).

Common Malay words of everyday-use are selected and the relevant information extracted from *Kamus Dewan* [12], a Malay monolingual dictionary, and *Kamus Inggeris Melayu Dewan* [13], an English-Malay bilingual dictionary, respectively. If need arises, the lexicon can be enriched further by incorporating extra information for each lexical entry, such as equivalent words in other languages, or linkages to senses in other dictionaries (e.g. *WordNet* – the linkages between word senses in *Kamus Dewan* and *WordNet 1.5* senses have been performed and is available at UTMK).

[1] or as a database, if need be

Categorisation of the Lexical Entries. The instances of the semantic classes are the lexical entries, which needs to be classified under their respective classes. The Malay word in each lexical entry will be translated to a Japanese word having the equivalent sense to that entry, and the Japanese word will be looked-up in *GoiTaikei* to identify the semantic class in which it belongs. The lexical entry (with the original Malay word) will then be inserted into the semantic class in our ontology-based lexicon.

The Relations. As mentioned earlier, relations can be specified to link various concepts and instances (in this case, lexical entries). Firstly, the taxonomic structure implies that superclasssubclass relations already exist in our lexicon. *GoiTaikei* itself includes valency and selectional restriction information for verbs, adjectives and adverbs, which can be incorporated into our lexicon. We can take a further leaf out of *WordNet*: if a relation exists between two synsets in *WordNet*, we can create a link between the corresponding two lexical entries in our lexicon. However, of the myriad types of relations in *WordNet*, we are still considering the suitable ones to be included and used in our WSD algorithm.

3 Using the Ontology-based Multilingual Lexicon for Word Selection

Part-of-speech (POS) information has been shown to solve 87% of all word ambiguities [14]. This is useful, since our lexical ontology has separate hierarchies for words of different POS, and contemporary POS-taggers are of high accuracy [15]. Elsewhere, the dependency structure of a sentence also gives clues to resolve disambiguities to a certain level, and there is current work in extracting "structural templates" from bilingual knowledge-bases at UTMK [16] to serve this purpose.

As part of his MSc work, Lim [17] developed an unsupervised, knowledge-based sense-tagger using the definition texts in dictionaries. After annotating definition entries in *WordNet* with "semantic primitives", Lim used the derived "lexical conceptual distance" data (LCDD) to measure the relatedness between words to determine the sense of an ambiguous word. While he did not make use of the many lexical relations in *WordNet*, he suggested that taking these – or some hierarchical net of a computer-tractable lexicon – into account would improve the algorithm's accuracy.

We plan to extend Lim's Lexical Conceptual Data Distance (LCDD) algorithm by incorporating the hierarchical structure of the ontology, and any relationships that will be defined. Different heuristics may be used when calculating the LCDD of words of different syntactic categories, as they seem to "behave" differently [2][3].

Resolving the correct sense of the occurrence of an ambiguous word will first involve identifying the semantic class to which the correct sense belongs. Since the lexical entries contain words from different languages with equivalent senses,

the multilingual lexicon can be used for performing WSD on input sentences in any language that is included in the ontology. As an example, consider the English word *hand*, which we decide (for the sake of illustration) to list the following four senses for it (as a noun) in the lexicon:

- part of arm below wrist (*tangan* in Malay),
- manual worker (*pekerja* in Malay),
- handwriting (*tulisan* in Malay), and
- help (*bantuan* in Malay).

Figure 1, which is a (much simplified) subset of the taxonomic structure of the lexicon, shows how the four senses of the word hand are categorised under different semantic classes.

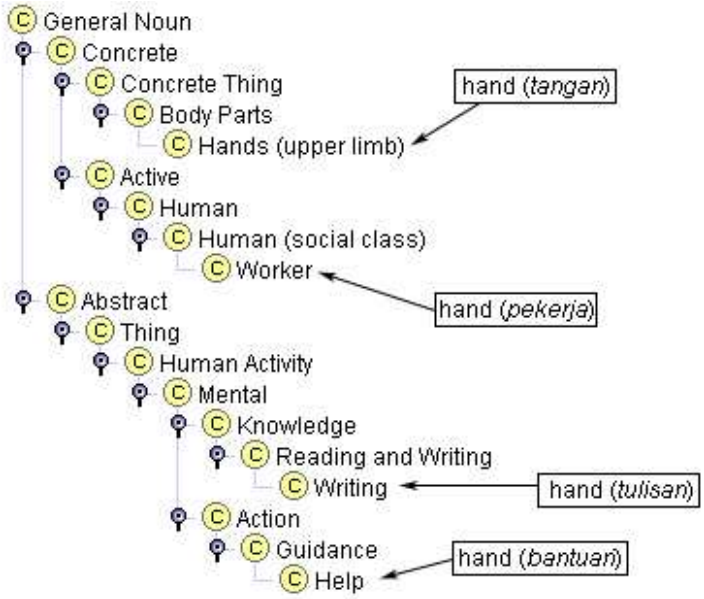

Fig. 1. Subset of the ontology-based multilingual lexicon, showing the semantic classes which the different senses of the English word *hand* belong to

Given the input English sentence,

*The farm **hands** are going on a strike.*

we can calculate the LCDD between the ambiguous word *hands* and the remaining words in the sentence, based on the definition texts and the structural information in the ontology. Once the word is disambiguated, i.e. the lexical entry corresponding to the sense of this particular occurrence is found (in this case,

the entry under *Worker*), the equivalent translation word can then be extracted. If multiple equivalent translation words are found for the selected sense, statistical information for word co-occurrence extracted from parallel corpora can be utilised to select a translation word that gives a more "natural", grammatical output sentence, as proposed in [18].

In addition, the information contained in the ontology-based lexicon can also be reused for other natural language processing tasks, such as speech synthesis. Homonyms are words having distinct meanings but the same lexical form, and are often pronounced differently when used to mean different things. For example, the Malay word *semak* (a bush), is pronounced differently from *semak* (to check or inspect). The phonological information in the multilingual lexicon can then be used to synthesise correct pronunciations of homonyms.

4 Future Work

The multilingual lexical ontology is still in the early stages of being constructed, and there is much work to be done. The hierarchy of nouns will be constructed as a start. We summarise some future concerns here, some of which have been mentioned earlier:

- identifying suitable relations to be included in the ontology-based lexicon,
- identifying other lexical or semantic information than may be needed, in future, for each lexical entry,
- extending Lim's LCDD algorithm with information from the ontology and other heuristics,
- determining if and how adjectives and adverbs can be re-categorised in the ontology-based lexicon. (*GoiTaikei* classifies verbs, adjectives and adverbs in the same hierarchy.)

One shortcoming of our work is that since the lexical entries in the lexicon are prepared by hand, it will be a time- and labour-consuming task. Another possible future work would be to automatically acquire lexical information from various sources, and to automatically insert new lexical entries into the lexicon, based on existing entries and the definition text of the new entry.

5 Conclusion

We propose the construction of an ontology-based multilingual lexicon, from existing language resources, as part of an approach to WSD in machine translation. The lexical ontology will also include a variety of information, including definition texts, equivalent translation words in other languages, phonological and morphological information, such that it can be used for NLP tasks other than machine translation, including information search and retrieval, speech processing, text categorisation and language identification.

References

1. Ng, H. T., Wang, B., Chan, Y. S.: Exploiting Parallel Texts for Word Sense Disambiguation: An Empirical Study. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, Sapporo Convention Center, Sapporo, Japan (1990) 455–462.
2. Ide, N., Véronis, J.: Word Sense Disambiguation: The State of the Art. Computational Linguistics, 24(1) (1998) 1–41.
3. Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.: Introduction to WordNet: An On-line Lexical Database. International Journal of Lexicography (special issue), 3(4) University of Chicago Press, Chicago, Illinois (1990) 235–312.
4. Kang, S. J., Lee, J. H.: Ontology-Based Word Sense Disambiguation by Using Semi-Automatically Constructed Ontology. In Proceedings of MT Summit VIII, Santialgo de Compostela, Galicia, Spain (2001)
5. Gruber, T.: A Translation Approach to Portable Ontology Specification. Knowledge Acquisition, 5(2) (1993) 199–220
6. Noy, N. F., McGuiness, D. L.: Ontology Development 101: A Guide to Creating Your First Ontology. In: SMI Technical Report SMI-2001-0880 (2001)
7. ROGET: Roget's Thesaurus of English Words and Phrases. Lloyd, S. (ed) Penguin Books (1984)
8. Kilgariff, A., Yallop, C.: What's in a Thesaurus? In Proceedings of the 2nd International Conference on Language Resources and Evaluation, Athens, Greece (2000) 1371-1379.
9. Ikehara, S., Miyazaki, M., Shirai, S., Yokoo, A., Nakaiwa, H., Ogura, K., Ooyama, Y., Hayashi, Y.: GoiTaikei – A Japanese Lexicon CDROM. Iwanami Shoten, Tokyo, Japan. (1999) (website: `http://www.kecl.ntt.co.jp/mtg/resources/GoiTaikei/index-en.html`)
10. McGuiness, D. L., van Harmelen, F.: OWL Web Ontology Language Overview: W3C Recommendation 10 February 2004. Word Wide Web Consortium (2004) (website: `http://www.w3.org/TR/owl-features`)
11. PROTÉGÉ: Protégé 2000. Stanford Medical Informatics, Stanford University School of Medicine (2000) (website: `http://protege.stanford.edu/index.html`)
12. KD: Kamus Dewan Edisi Baru. Sheikh Othman bin Sheikh Salim et al (eds). Dewan Pustaka dan Bahasa, Kuala Lumpur, Malaysia (1993)
13. KIMD: Kamus Inggeris Melayu Dewan: An English-Malay Dictionary. Johns A. H. et al (eds). Dewan Pustaka dan Bahasa, Kuala Lumpur, Malaysia (2000)
14. Wilks, Y., Stevenson, M.: Sense Tagging: Semantic Tagging with a Lexicon. In Proceedings of the SIGLEX Workshop on Tagging Text with Lexical Semantics, Washington, D.C. (1997) 74–78.
15. Brill, E.: Transformation-based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging. Computational Linguistics, 21(4) (1995) 543-566.
16. Ye, H. H., Tang, E. K.: Learning Translation Templates from Bilingual Knowledge Bank. In Compilation of Computer Science Postgraduate Colloquium, Universiti Sains Malaysia, Penang, Malaysia (2004) 83–84
17. Lim, B. T.: Semantic-Primitive-Based Lexical Consultation System. MSc Thesis, Universiti Sains Malaysia, Penang, Malaysia (2003)
18. Lee, H. A., Kim, G. C.: Translation Selection through Source Word Sense Disambiguation and Target Word Selection. In Proceedings of COLING 2002, Taipei, Taiwan (2002)