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*. Apart from providing equivalent words from different languages, the lexicon will be used to extend a WSD algorithm that calculates lexical conceptual distance data. 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 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 (from a bilingual dictionary or lexicon), to translate ambiguous words in a source language, as in the example above, is often termed target word selection.

Many researchers turn to *WordNet* [3] as a 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]. Hence, merely linking non-English words to *WordNet* is insufficient: further processing (e.g. combining and/or dropping) of *WordNet* senses is often required. 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 ontology-based 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 hierarchical 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 and hierarchical structures in lexical resources is not a new idea. *Roget's Thesaurus* 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 [7] 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 [7], whereas *WordNet* was constructed based on psycholinguistic principles

[3]. Neither are traditional dictionaries (in either book or electronic forms) 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* [8], an electronic Japanese lexicon. *GoiTaikei* contains around 300,000 Japanese words categorised under 3,000 classes in three hierarchies: general nouns, proper nouns, and "phenomenons" (verbs, adjectives and adverbs). The hierarchies here are desirable, since *GoiTaikei* was developed for use with a machine translation system. There are no class definitions in *GoiTaikei* [9]; instead, the classes are used to semantically specify word senses. The Japanese words are marked with part-of-speech information and the classes they are associated with, while words in the "phenomenon" hierarchy are organised as a valency dictionary with selectional restrictions.

It may be noted that the classes set out in *GoiTaikei* are not semantically universal. Therefore, they may not necessarily tally with classes defined in other ontologies of any type and funciton. For our purposes, *GoiTaikei*'s classes serve as specifications of word senses for natural language processing tasks.

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 word form in English,
- part-of-speech,
- definition keywords for a particular sense of the word,
- equivalent word(s) in other languages,
- definition entries from various dictionaries associated with this sense.

Figure 1 shows a sample lexical entry for the word *impartial* which contains the equivalent Chinese and Malay word sense entries.

[1] or as a database, if need be

```
impartial
wordnet(300280426, 'impartial', a, [free from undue bias or
   preconceived opinions]).
dict_modern_chinese(公平, 1, [处理事情合情合理，不偏袒哪一方面]).
dict_modern_chinese(无私, 1, [不自私]).
kamus_dewan(adil, [yg atau dgn berdasarkan pertimbangan(peraturan,
   ketentuan, dll) yg wajar atau berpatutan (bkn orang, tindakan,
   hukuman, keputusan, undang-undang, dll), tidak memihak ke mana-
   mana, (apabila memutuskan sesuatu dsb), tidak sewenang-wenang]).
kamus_dewan(saksama, [tidak berat sebelah, tidak menaruh prasangka,
   adil]).
```

Fig. 1. A sample lexical entry for the word *impartial*. The definition entries were extracted from *WordNet*, the *Dictionary of Modern Chinese Words* and *Kamus Dewan*

For each word sense, the corresponding entries from dictionaries of different languages are matched, and the equivalent words in target languages extracted. The linking of lexical entries to *WordNet 1.5* have also been performed and is available at UTMK.

Categorisation of the Lexical Entries. The instances of the semantic classes are the lexical entries, which needs to be associated with the relevant classes. The English word in each lexical entry will be translated to a Japanese word having the equivalent sense to that entry. This Japanese word is then looked up in *GoiTaikei* to identify the semantic classes in which it appears. The lexical entry (with the original English word) will then be added to the corresponding 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 edges in *GoiTaikei*'s noun tree represents hyponymy and meronymy [9]. Therefore, our lexicon already contains "is-a" and "part-of" relations for nouns. In addition, *GoiTaikei* includes valency and selectional restriction information for verbs, adjectives and adverbs, which can be incorporated into our lexicon. Morphological relations among word forms can also be extracted from dictionaries. 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* (as well as other facets and properties), we are still considering the suitable ones, besides hyponymy and meronymy, to be included and used in our WSD algorithm.

3 Using the Ontology-based Multilingual Lexicon for Target Word Selection in Machine Translation

Part-of-speech (POS) information often gives helpful clues as to the correct sense of an ambiguous word [12]. This is useful, since our lexical ontology has separate hierarchies for words of different POS, and contemporary POS-taggers are of high accuracy [13]. Therefore, current WSD efforts are mostly geared towards solving ambiguaties in the same syntactic category. Elsewhere, the dependency structure of a sentence also gives clues to resolve ambiguities to a certain level, and there is current work in extracting structural templates [14] and identifying multi-word verbs [15] from bilingual knowledge-bases at UTMK to serve this purpose.

As part of his MSc work, Lim [16][17] developed an unsupervised, knowledge-based sense-tagger using the definition texts in dictionaries. First of all, using Guo's method [18], a set of descriptive semantic primitives were extracted from a dictionary. After annotating the definition entries in *WordNet* with semantic primitives, Lexical Conceptual Distance Data (LCDD) between word senses was derived to measure the relatedness between them, in order to determine the sense of an ambiguous word. While he did not make use of the many lexical relations in *WordNet*, Lim 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 LCDD algorithm by incorporating the hierarchical structure of our ontology, and any relationships that will be defined. For example, the LCDD among *classes* of word senses can be computed as a function of the LCDD among word senses in those classes. Also, different heuristics may be used when calculating the LCDD of words of different syntactic categories, as they seem to "behave" differently [2][3].

To resolve the correct sense of an ambiguous word in an input sentence, we compute the LCDD between its possible senses and those of the words in context, as well as the classes involved. 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 noun *hand*, for which we decide (for the sake of illustration) to list the following four senses 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 2, 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.

Given the input English sentence,

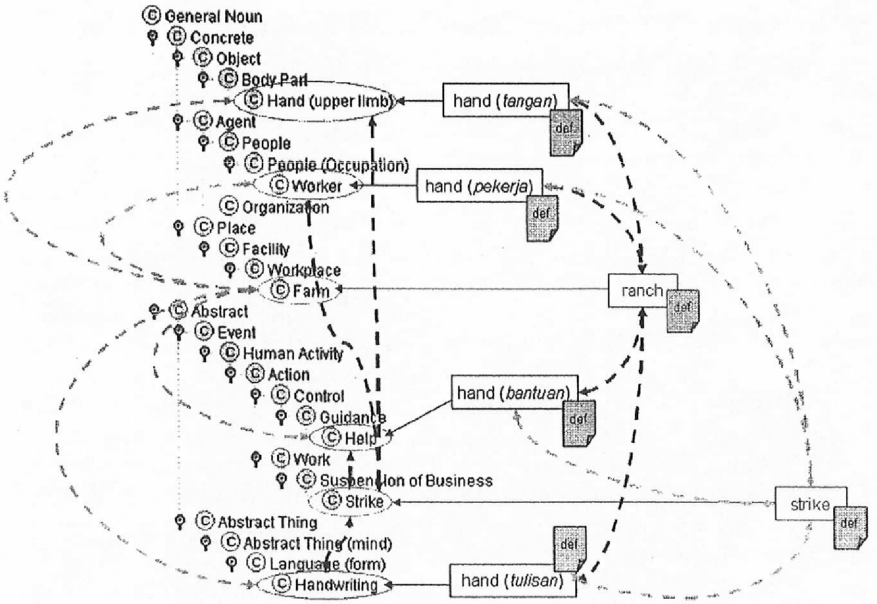


Fig. 2. Subset of the multilingual lexicon, showing the semantic classes in which the different senses of the English word *hand* appear. Arrows indicate pairs of words or classes where LCDDs will be computed for the example input sentence *The ranch hands are going on a strike*

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

and we wish to disambiguate the meaning of *hand*, we calculate the LCDD between the four senses of *hands* and the other content words (e.g. *farm* and *strike*) in the sentence. We also compute the LCDD between the classes to which the word senses belong. This will be done based on the definition texts and the structural information in the ontology. Once the sense of the word *hands* 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 [19].

This lexicon can be reused for other natural language processing tasks, such as speech synthesis, if we enrich the ontology-based lexicon with other information (e.g. syllable segmentations and IPA notations). Homonyms are words having distinct meanings but the same lexical form, and are often pronounced differently when used to mean different things. For instance, 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. (They reside in the same hierarchy in *GoiTaikei*.)

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.

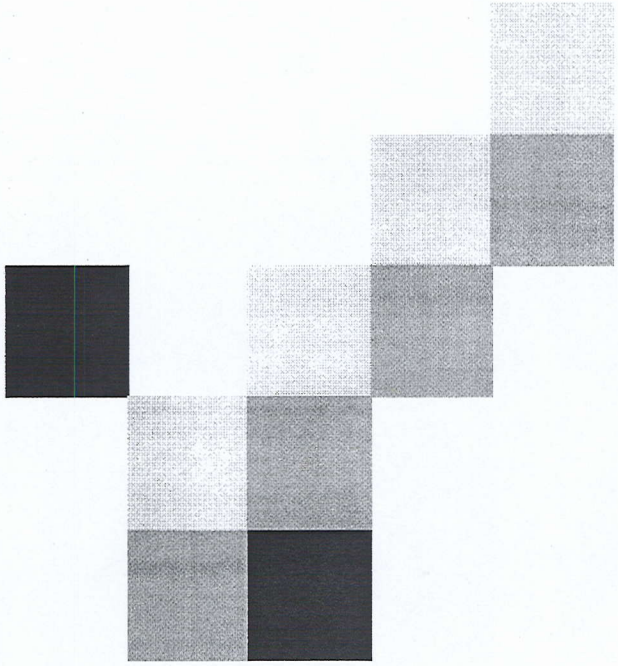
5 Conclusion

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 lexicon 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. 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.
8. 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`)
9. Asanoma, N.: Alignment of Ontologies: WordNet and Goi-Taikei. NAACL WordNet and Other Lexical Resources, Pittsburgh, USA (2001) 89–94
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. 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.
13. 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.
14. Ye, H. H., Tang, E. K.: Learning Translation Templates from Bilingual Knowledge Bank. In Compilation of Extended Abstracts of Computer Science Postgraduate Colloquium, Universiti Sains Malaysia, Penang, Malaysia (2004) 83–84
15. Kee, T. H., Tang, E. K., Chuah, C. K.: Identifying Multi-Word Verbs (MWV) from Bilingual Knowledge Bank. In Compilation of Extended Abstracts of Computer Science Postgraduate Colloquium, Universiti Sains Malaysia, Penang, Malaysia (2004)
16. Lim, B. T., Tang, E. K., Guo, C. M.: Building a Semantic Primitive Based Lexical Consultation System. Pre-Coling 2002 Seminar On Linguistic Meaning Representation And Their Applications Over The World Wide Web, Penang, Malaysia (2002)
17. Lim, B. T.: Semantic-Primitive-Based Lexical Consultation System. MSc Thesis, Universiti Sains Malaysia, Penang, Malaysia (2003)
18. Guo, C. M.: Deriving a Natural Set of Semantic Primitives from Longman Dictionary of Contemporary English. Proceedings of the Second Irish Conference on Artificial Intelligence and Cognitive Science. (1989) 218–227
19. 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)



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

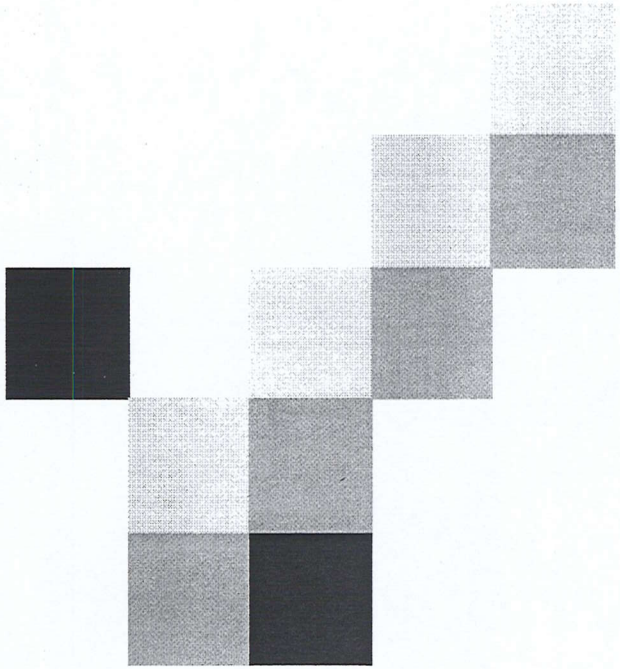
Lim Lian Tze

Unit Terjemahan Melalui Komputer
Pusat Pengajian Sains Komputer
Universiti Sains Malaysia
Penang, Malaysia

liantze@cs.usm.my

Presentation Overview

- Introduction
- Building an Ontology-based Multilingual Lexicon
- Using the Lexicon for Target Word Selection in Machine Translation
- Future Work
- Conclusion



Introduction

Word Sense Disambiguation

- Ambiguous words: words with multiple meanings
- WSD: determine correct meaning (sense) of ambiguous word in particular discourse
- Need of WSD in machine translation (word selection)
  - Input: *The computer **logs** were deleted.*
  - Output: ***Balak** komputer telah dipotong.*

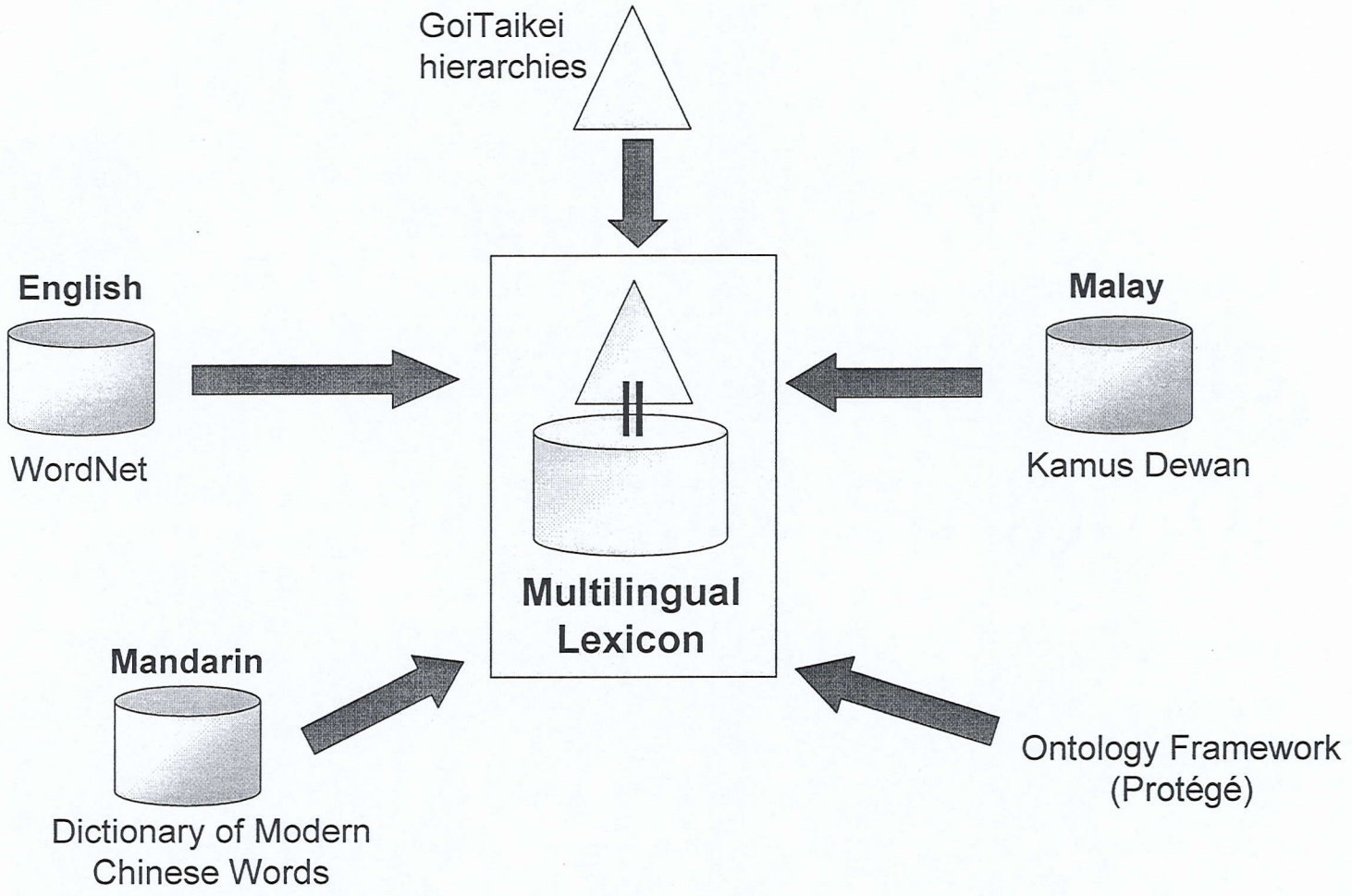
⇨ Based on the list of meanings of words as defined in a bilingual dictionary

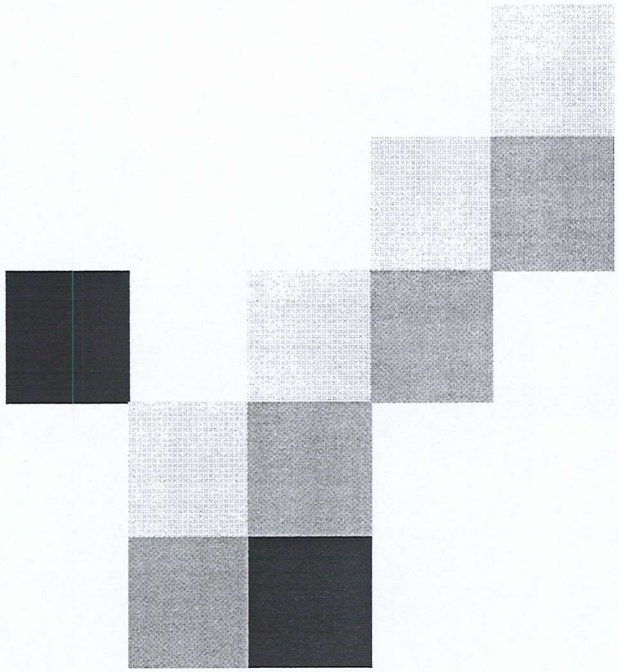
Language Resource for WSD

- (Bilingual) list of words and senses
- WordNet
  - broad coverage, rich lexical information, freely available
  - too fine-grained for practical NLP tasks
  - Linking of words in target languages to WordNet senses is insufficient

⇨ Propose to construct multilingual lexicon based on ontology framework

Combining Lexical Resources





Building an Ontology-based Multilingual Lexicon

Existing Lexical Resources using Hierarchical Structures

- Roget's Thesaurus, WordNet
- Shortcomings – not perfect resources for WSD

⇨Build our own

Construction of the Lexicon

- Building the hierarchical structures
- Preparing the lexical entries
- Classifying or categorising the lexical entries
- Specifying suitable relations among the lexical entries

The Hierarchies

- Based on GoiTaikei – A Japanese Lexicon
- 3,000 semantic classes in 3 hierarchies
  - General nouns
  - Proper nouns
  - "Phenomenons" (verbs, adjectives, adverbs)
- Each Japanese word tagged with
  - POS
  - semantic class(es)
  - "phenomenons": phrasal patterns with selectional restrictions
- Japanese label of classes translated to English
- Structure re-created in ontology web language (OWL) file/database

GoiTaikei–A Japanese Lexicon, Ikehara et al (1999)
http://www.kecl.ntt.co.jp/mtg/resources/GoiTaikei/index-en.html

The Hierarchies (cont.)

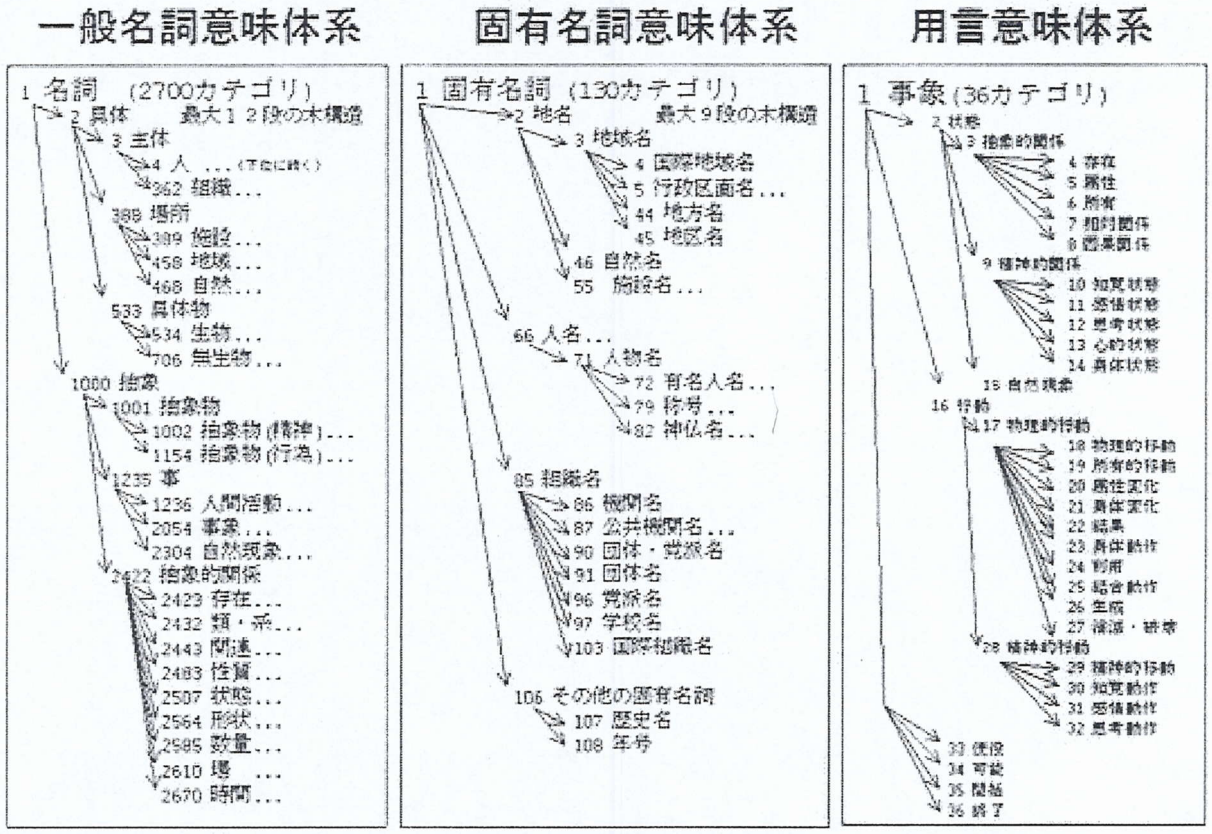
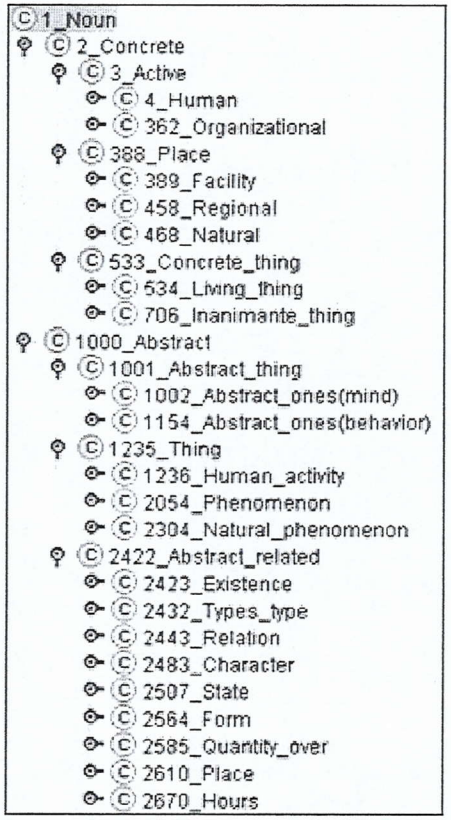


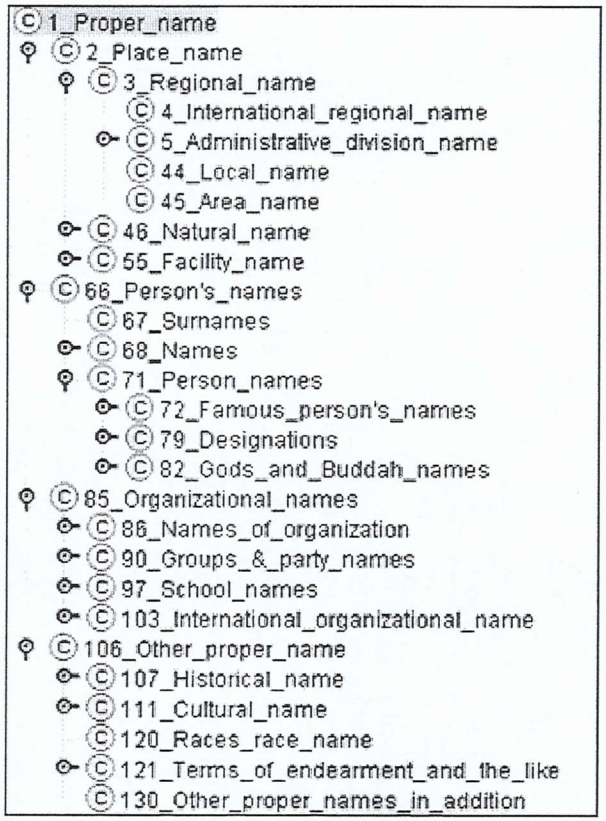
図1 意味の体系

Source: GoiTaikei–A Japanese Lexicon, Ikehara et al (1999)
http://www.kecl.ntt.co.jp/mtg/resources/GoiTaikei/index-en.html

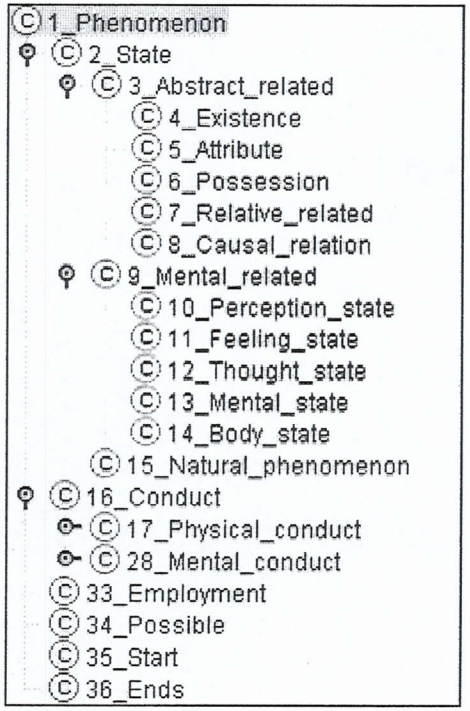
The Hierarchies (cont.)



General Noun Hierarchy



Proper Noun Hierarchy

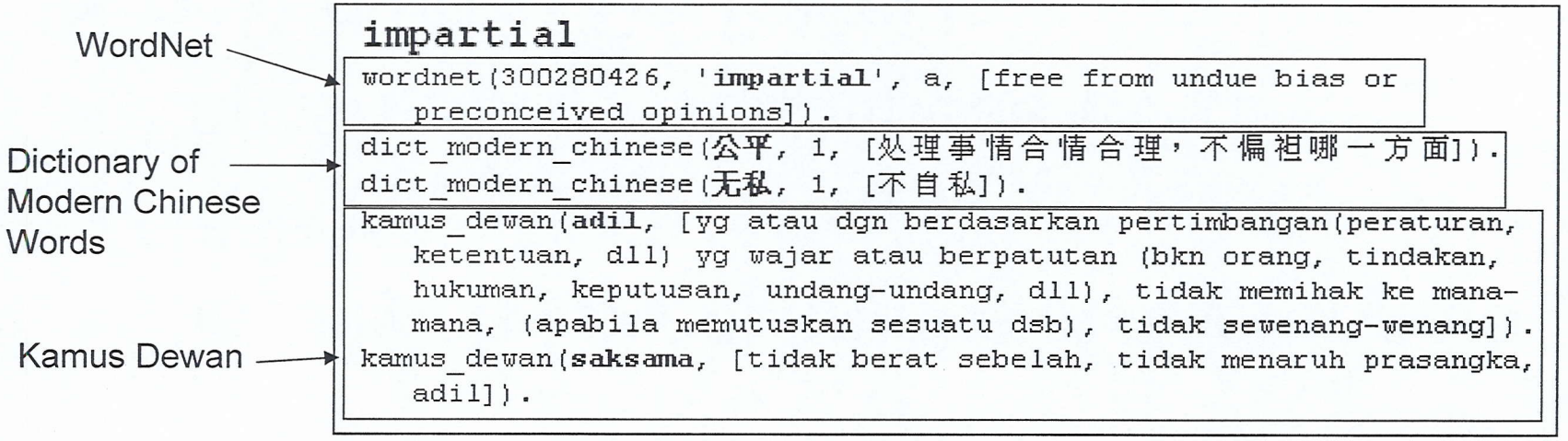


Phenomenon Hierarchy

The Lexical Entries

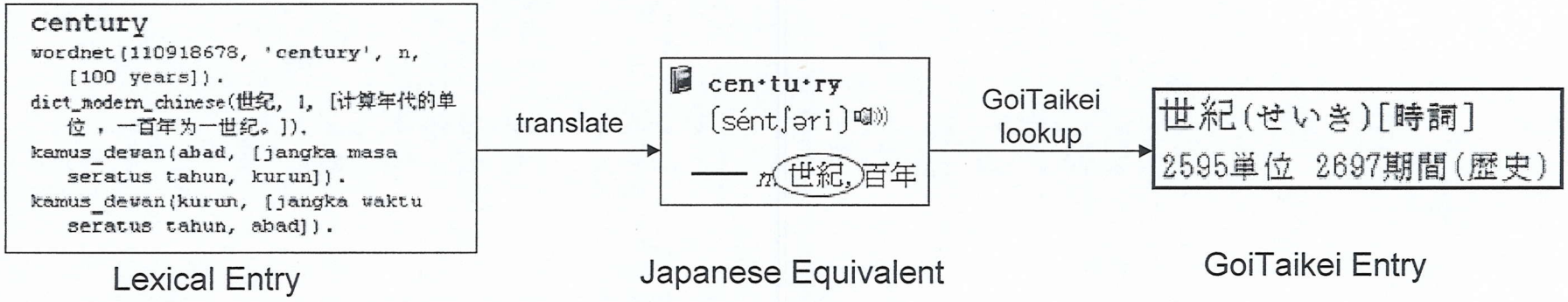
- Each lexical entry represents a sense of a word
- Information included:
  - English word-form
  - POS
  - definition keywords
  - equivalent word(s) in other languages
  - definition entries from dictionaries

The Lexical Entries (cont.)



Classifying the Lexical Entries

- Classifying lexical entries in appropriate classes
- English word → Japanese word
- looked up in GoiTaikei to determine semantic class



The Relations

- GoiTaikei noun hierarchy: hyponymy ("is-a") and meronymy ("part-of")
- GoiTaikei: phrasal patterns and selectional restrictions for verbs, adjectives

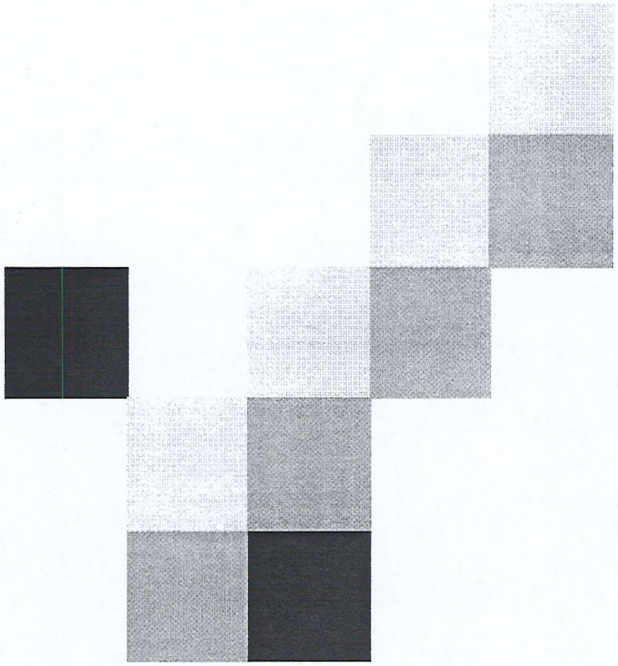
```
取る(とる)
(2)  19所有的移動  (動作)
  N1が N2を N3から/より 取る  N1 charge N3 N2
  [N1(3主体) N2(1199料金 1190金銭) N3(3主体)]
```

```
取る(とる)
(15)  32思考動作  (状態)
  N1が N2を N3に 取る  N1 reserve N2 at/in N3
  [N1(3主体) N2(2612席 868部屋 437宿泊施設 932券
  986乗り物) N3(388場所 2610場)]
```

Source: GoiTaikei–A Japanese Lexicon, Ikehara et al (1999)
http://www.kecl.ntt.co.jp/mtg/resources/GoiTaikei/index-en.html

The Relations (cont.)

- Morphological relations between words
- WordNet: various types of semantic relations
  - Hyponymy and meronymy already present in GoiTaikei noun hierarchies
  - (still considering types of relations suitable to be included)



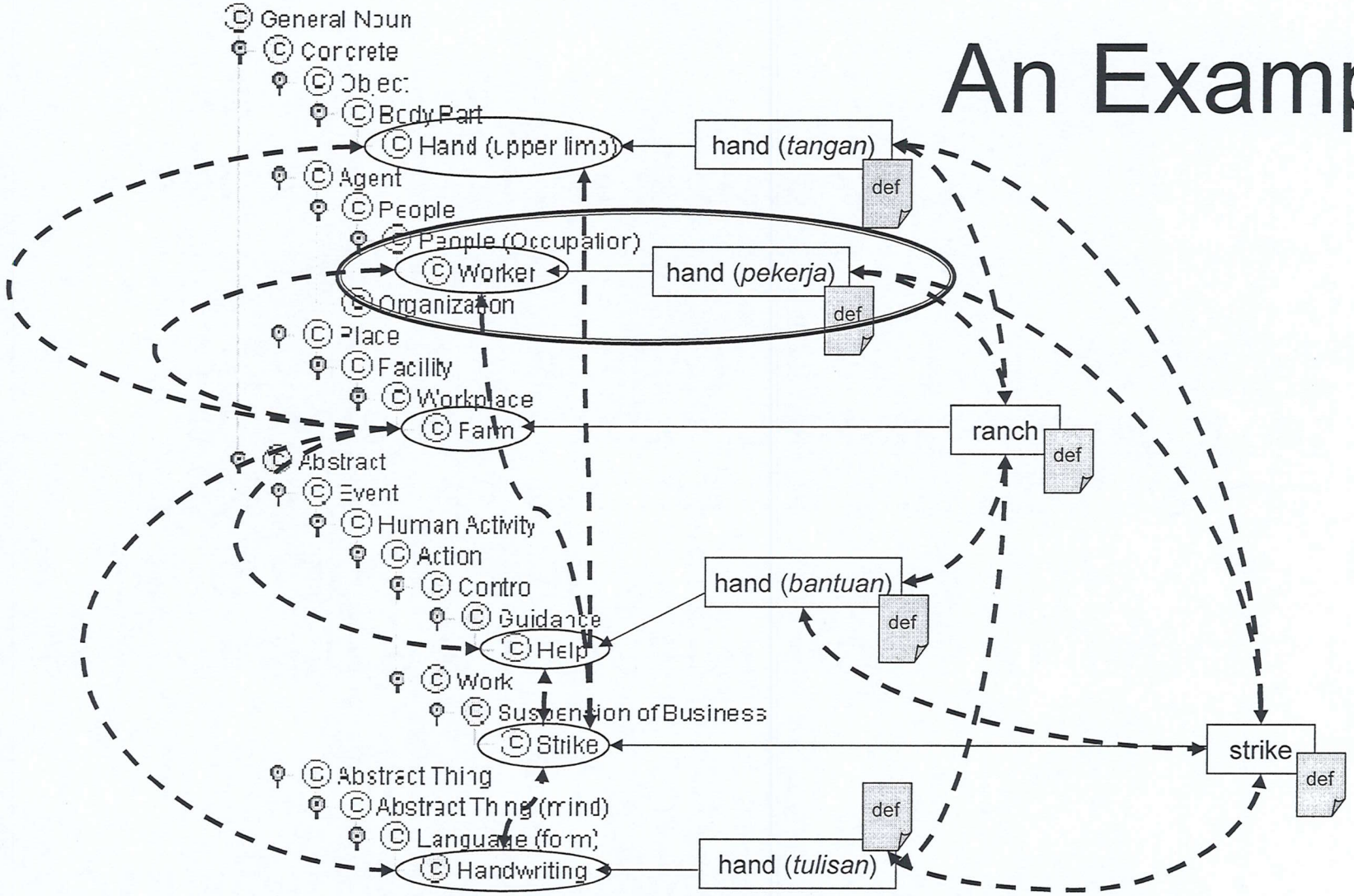
Using the Ontology-based Multilingual Lexicon for Word Selection

Using the Ontology-based Multilingual Lexicon for Word Selection

- Lim et al (2002) calculates Lexical Conceptual Distance Data (LCDD) as measure of relatedness between word senses, using definition texts
- Extension: compute LCDD between *classes* of words too
- Apply different heuristics and weights – words of different POS "behave" differently (Miller et al 1990, Ide and Véronis 1998)

Lim, B.T, Guo, C. M., Tang, E. K.: Building a Semantic-Primitive-Based Lexical Consultation System (2002);
Miller, G. et al: Introduction to WordNet: An On-line Lexical Database (1990);
Ide, N., Véronis, J.: Word Sense Disambiguation (1998)

An Example

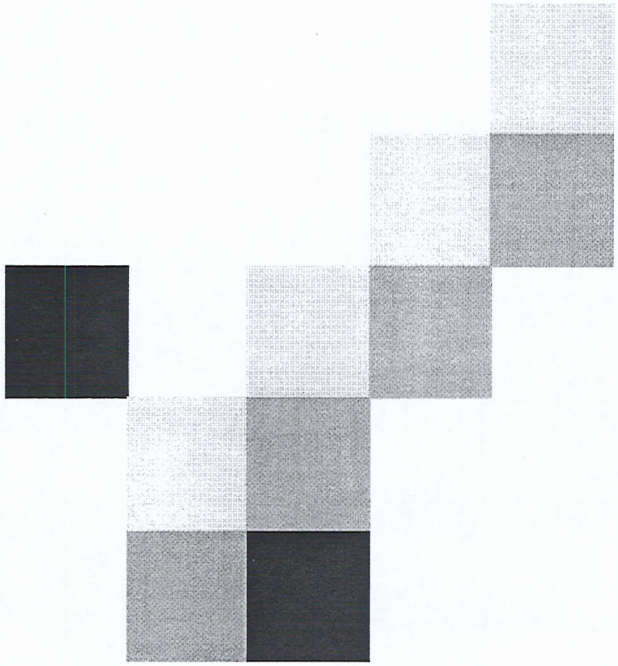


Input: The ranch **hands** are going on a strike.

Using the Ontology-based Multilingual Lexicon (cont.)

- If multiple equivalent words in target language found?
  - Can use co-occurrence data from parallel corpora for a more "natural", grammatical output, as done by Lee and Kim (2002)
- Miscellaneous
  - speech synthesis: homonyms
  - eg. "*semak*"

Lee, H. A., Kim, G. C.: Translation Selection through Source Word Sense Disambiguation and Target Word Selection



Future Work and Conclusion

Future Work

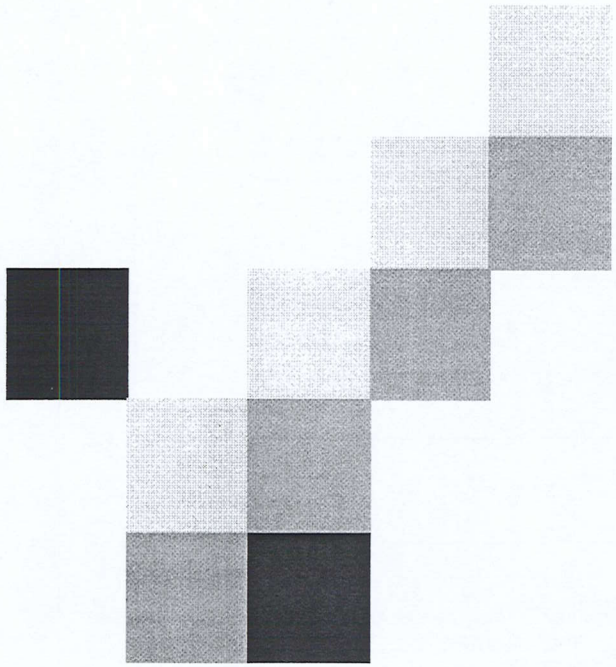
- Early stages – still much to be done!
- Some concerns:
  - identifying suitable relations
  - identifying other information for lexical entries
  - extending LCDD algorithm with structural or relational information
  - determining if and how adjectives and adverbs can be re-categorised

Future Work (cont.)

- Manual preparation → time and labour consuming
- Investigate automation of:
  - acquiring lexical information from various sources
  - inserting new lexical entries into the lexicon, given existing entries in lexicon and definition texts of new entries (bootstrapping)

Conclusion

- Proposed construction of a multilingual lexicon, using ontology framework, for WSD in machine translation
- Includes definition texts, equivalent translations in other languages
- Using existing language resources (GoiTaikei, WordNet, etc)
- Reusable for other NLP tasks



Thank You