Extending the Japanese WordNet

Francis Bond,[†] **Hitoshi Isahara,**[‡] **Kiyotaka Uchimoto,**[†]
[†] **Takayuki Kuribayashi**[†] and **Kyoko Kanzaki**[‡]
[†] NICT Language Infrastructure Group, [‡] NICT Language Translation Group
MASTAR Project
National Institute of Information and Communications Technology
<{bond,isahara,uchimoto,kuribayashi,kanzaki}@nict.go.jp>

1 Introduction

Our goal is to make a semantic lexicon of Japanese that is both **accesible** and **usable**. To this end we are constructing and releasing the Japanese WordNet (`WN-Ja`) (Bond et al., 2008b,a).

We have almost completed the first stage, where we automatically translated the English and Euro WordNets, and are hand correcting it. We introduce this in Section 2. Currently, we are extending it in three main areas: the first is to add more concepts to the Japanese WordNet, either by adding Japanese to existing English synsets or by creating new synsets (§ 3). The second is to link the synsets to text examples (§ 4). Finally, we are linking it to other resources: the semantic lexicon GoiTaikei (Ikehara et al., 1997) and a collection of illustrations taken from the Open ClipArt Library (Phillips, 2005) (§ 5).

2 Current State

Currently, the `WN-Ja` consists of 157,646 senses (word-synset pairs) 36,922 concepts (synsets) and 73,113 unique Japanese words. The relational structure (hypernym, meronym, domain, ...) is based entirely on the English WordNet 3.0 (Fellbaum, 1998). Of these entries, 81% have been checked by hand, 11% were automatically created by linking through multiple languages and 8% were automatically created by adding non-ambiguous translations, as described in Bond et al. (2008a). For up-to-date information on `WN-Ja` see: `nlpwww.nict.go.jp/wn-ja`.

An example of the entry for the synset 02076196-n is shown in Figure 1. Most fields come from the English WordNet. We have added the underlined fields (Ja Synonyms, Illustration, GoiTaikei) and are currently adding the translated gloss. In the initial automatic construction there were 27 Japanese words associated with the synset,[1] including many inappropriate translations for other senses of *seal* (e.g., 判こ *hanko* "stamp"). These were reduced to three after checking: アザラシ, 海豹 *azarashi* "seal" and シール *shi-ru* "seal".

The main focus of this year's work has been this trimming of badly translated words. The result is a WordNet with a reasonable coverage of common Japanese words. The precision per sense to be just over 90%. We have aimed at high coverage at the cost of precision for two reasons: (i) we think that the WordNet must have a reasonable coverage to be useful for NLP tasks and (ii) we expect to continue refining the accuracy over the following years.

3 Increasing Coverage

We are increasing the coverage in two ways. The first is to continue to manually correct the automatically translated synsets: there are still some 27,000 unchecked synsets. More interestingly, we wish to add synsets for Japanese concepts that may not be expressed in the English WordNet. To decide which new concepts to add, we will be guided by the other tasks we are are doing: annotation and linking. We intend to create new synsets for words found in the corpora we annotate that are not currently covered, as well for concepts that we want to link to. An example for the first is the concept 御飯 *gohan* "cooked rice", as opposed to the grain 米 *kome* "rice". An example of the second is シングル *shinguru* "single: a song usually extracted from a current or upcoming album to promote the album". This is a very common hypernym in Wikipedia but missing from the English WordNet.

[1] アザラシ, シール, スタンプ, 上封, 判, 判こ, 判子, 刻印, 加判, 印, 印判, 印形, 印章, 印鑑, 印鑑, 印鑰, 印顆, 墨引, 墨引き, 封, 封じ目, 封印, 封目, 封着, 封緘, 押し 手, 押印, 押手, 押捺, 捺印, 極印, 海豹, 版行, 符節, 緘, 証印, 調印

Synset	02076196-n
Synonyms	ja 海豹, アザラシ, シール en seal fr phoque
	 animal/seal.png
Def (en)	"any of numerous marine mammals that come on shore to breed; chiefly of cold regions"
Def (ja)	「繁殖のために岸に上がる海洋性哺乳動物の各種；主に寒帯地域に」
Hypernyms	アシカ亜目/pinniped
Hyponyms	？/crabeater_seal ？/eared_seal 海驢/earless_seal
GoiTaikei	⟨⟨537:beast⟩⟩

Figure 1: Example Entry for Seal/海豹

Name	Sentences	Words	Content Words
Semcor	12,842	224,260	120,000
Glosses	165,977	1,468,347	459,000
Kyoto	38,383	969,558	527,000

Table 1: Corpora to be Sense Tagged

As far as possible, we want to coordinate the creation of new synsets with other projects: for example KorLex: the Korean WordNet already makes the cooked rice/grain distinction, and the Princeton WordNet should also have a synset for this sense of *single*.

4 Text Annotation

We are in the process of annotating three texts (Table 1). The first two are translations of WordNet annotated English Texts (SemCor and the WordNet glosses), the third is the Japanese newspaper text that forms the Kyoto Corpus. We expect to finish translating and annotate all of SemCor, translate the WordNet glosses and start annotation on the Kyoto Corpus in 2009.

This annotation is essential for finding missing senses in the Japanese WordNet, as well as getting the sense distributions that are needed for supervised word sense disambiguation.

4.1 SemCor

Semcor is a textual corpus in which words have been both syntactically and semantically tagged. The texts included in Semcor were extracted from the Brown corpus (Francis and Kucera, 1979) and then linked to senses in the Princeton WordNet. The frequencies in this corpus were used to give the sense frequencies in WordNet (Fellbaum, 1998). A subset of this corpus (MultiSemCor) was translated into Italian and used as a corpus for the Italian WordNet (Bentivogli et al., 2004). We are translating this subset into Japanese.

In the same way as Bentivogli et al. (2004), we are exploiting Cross-Language Annotation Transfer to seed the Japanese annotation. For example, consider (1)[2]. The content words *answer, was, simple, honest* are tagged in Semcor. They can be aligned with their translations 答え *kotae* "answer" 簡単 *kantan* "simple", 率直 *socchoku* "honest" and だった *datta* "was". This allows us to tag the Japanese translation with the same synsets as the English

(1) His answer$_i$ was$_j$ simple$_k$ but honest$_l$.
答え$_i$ は 簡単$_k$ ながらも 率直$_l$ な ものだった$_j$ 。

We chose a translated Semcor as the basis of annotation for two main reasons: (i) the corpus can be freely redistributed — we expect the glosses to be useful as an aligned corpus — and (ii) it has several other annotations associated with it: Brown corpus POS annotation, Penn Treebank syntactic annotation, and the Italian Translations from the MultiSemCor corpus.

4.2 WordNet glosses

Our second translated corpus is formed from the WordNet glosses (and example sentences) themselves (e.g., the **def** field shown in Figure 1). The English glosses have also been annotated with word senses as the *Princeton WordNet Gloss Corpus*. In the same way that we do for SemCor, we are translating the glosses and seeding the annotations.

Using the glosses as the base for a sense annotated corpus is attractive for the following rea-

[2]Sentence 96 in b13.

sons: (i) the translated corpus can be freely redistributed — we expect the glosses to be useful as an aligned corpus and also to be useful for many other open lexicons; (ii) the glosses are useful for Japanese native speakers using the WordNet, (iii) the glosses are useful for unsupervised sense disambiguation techniques such as LESK (Baldwin et al., 2008) and (iv) other projects have also translated synset glosses (e.g. Spanish and Korean), so we can hope to create a multilingual corpus here as well.

4.3 Kyoto Text Corpus

The Kyoto Text corpus consists of newspaper text from the Mainichi Newspaper 1995, segmented and annotated with Japanese POS tags and dependency trees (Kurohashi and Nagao, 2003). We hope to annotate at least parts of it during 2009.

Even though the Kyoto Text Corpus is not freely redistributable, we have chosen to annotate it due to the wealth of annotation associated with it: dependency trees, predicate-argument relations and co-reference (Iida et al., 2007), translations into English and Chinese (Uchimoto et al., 2004) and sense annotations from the Hinoki project (Bond et al., 2006). We also felt it was important to tag some native Japanese text, not only translated text.

5 Linking to other resources

In our initial release, we link WordNet to two other resources: Nihongo GoiTaikei (Ikehara et al., 1997) and a collection of pictures from the Open Clip Art Library (OCAL: Phillips (2005)).

The basic approach is to find confident matches automatically and then generalize from them. We find matches in three ways:

MM Monosemous monolingual matches
e.g. *cricket bat* or 海豹

MB Monosemous bilingual matches
e.g. ⟨海豹↔*seal*⟩

HH Hypernym/Hyponym pairs
e.g. ⟨seal ⊂ mammal⟩

Similarly, we will also link the concepts from the EDR lexicon (EDR, 1990) and the hypernym-hyponym links from Torishiki-kai (Kuroda et al., 2009).

5.1 GoiTaikei

Linking Goi-Taikei, we used not only the Japanese dictionary published in Ikehara et al. (1997), but also the Japanese-English dictionary used in the machine translation system ALT-J/E (Ikehara et al., 1991). We attempted to match synsets to semantic categories by matching the Japanese, English and English-Japanese pairs to unambiguous entries in Goi-Taikei. For example, the synset shown in Figure 1 was automatically assigned the semantic category ⟨⟨537:beast⟩⟩, as 海豹 appears only once in `WN-Ja`, with the synset shown, and once in the Japanese dictionary for ALT-J/E with a single semantic category.

We are currently evaluating our results against an earlier attempt to link WordNet and GoiTaikei that also matched synset entries to words in Goi-Taikei (Asanoma, 2001), but did not add an extra constraint (that they must be either monosemous or match as a hypernym-hyponym pair).

5.2 Open ClipArt Library

In order to make the sense distinctions more visible we also semi-automatically link synsets to illustrations from the Open Clip Art Library (OCAL: Phillips (2005)). This adds a new modality to the knowledge linked in the semantic net. Illustrations of concepts are useful for a variety of tasks. One is pedagogical — it is useful to have pictures in learners' dictionaries. Another is in cross-cultural communication - for example in Pangea, where children use pictons (small concept representing pictures) to write messages (Takasaki and Mori, 2007).

We use the OCAL collection distributed as SVG (scalable vector graphic) images in the Ubuntu Fiesty distribution based on the release of October 2005 (v 0.18). It contains 6,826 unique images, organized in a shallow file hierarchy.

Each image is associated with a collection of explicit metadata, including a title, description and a set of tags, all of which are recommended rather than obligatory. We consider the title and simplified basename to be the entry for an illustration, and its tags the hypernyms. For example, for `seal.svg`, its title is *Etiquette Icons* and it is tagged as *animal* and *mammal*. We look in wordnet for hypernym synsets of *seal* that include *mammal* and find the following: `seal#n#9`

$\subset$ placental#n#1 $\subset$ mammal#n#1. Therefore, this picture illustrates seal#n#9 rather than the other synsets associated with the word *seal*

There are several sources of metadata. We first use explicit metadata such as TITLE for the root word and TAGS for the hypernyms. If there is no explicit metadata (which is true for around a third of the images) or we couldn't find a match, then we look for implicit metadata. We take the basename seal and delete any numbers from the end. We also add directory names to the tag list (animals/seal.svg): in this case "animal".

We have only linked a small subset of illustrations (936 out of 6,826 images) and an even smaller proportion of wordnet (758 out of 82,115 noun synsets). However, because any illustrated synset also (in theory) illustrates its hypernyms, we have indirectly illustrated far more than 758 synsets: these figures are better than they seem.

6 Discussion

We we presented the current state of the Japanese WordNet (157,646 senses, 36,922 concepts and 73,113 unique Japanese words, with links to Goi-Taikei and OCAL) and outlined our plans for further work (more words, links to corpora and other resources). We hope that WN-Ja will become a useful resource not only for natural language processing, but also for teaching and linguistic research.

Acknowledgments We would like to thank NTT, in particular Sanae Fujita, for their cooperation in linking to the Goi-Taikei.

References

Naoki Asanoma. 2001. Alignment of ontologies:wordnet and goi-taikei. In *NAACL Wokshop on WordNet & Other Lexical Resources*, pages 89–94. Pittsburgh, USA.

Timothy Baldwin, Su Nam Kim, Francis Bond, Sanae Fujita, David Martinez, and Takaaki Tanaka. 2008. MRD-based word sense disambiguation: Further extending Lesk. In *Proc. of the 3rd International Joint Conference on Natural Language Processing (IJCNLP-08)*, pages 775–780. Hyderabad, India.

Luisa Bentivogli, Pamela Forner, and Emanuele Pianta. 2004. Evaluating cross-language annotation transfer in the MultiSemCor corpus. In *20th International Conference on Computational Linguistics: COLING-2004*, pages 364–370. Geneva.

Francis Bond, Sanae Fujita, and Takaaki Tanaka. 2006. The Hinoki syntactic and semantic treebank of Japanese. *Language Resources and Evaluation*, 40(3–4):253–261. URL http://www.springerlink.com/content/h4t478u428054058/fulltext%.pdf, (Special issue on Asian language technology).

Francis Bond, Hitoshi Isahara, Kyoko Kanzaki, and Kiyotaka Uchimoto. 2008a. Boot-strapping a WordNet using multiple existing WordNets. In *Sixth International conference on Language Resources and Evaluation (LREC 2008)*. Marrakech.

Francis Bond, Hitoshi Isahara, Kyoko Kanzaki, and Kiyotaka Uchimoto. 2008b. Using multilingual WordNets to compile a Japanese WordNet. In *14th Annual Meeting of the Association for Natural Language Processing*, pages 853–856. Tokyo.

EDR. 1990. Concept dictionary. Technical report, Japan Electronic Dictionary Research Institute, Ltd.

Christine Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

W. N. Francis and H. Kucera. 1979. *BROWN CORPUS MANUAL*. Brown University, Rhode Island, third edition. (http://khnt.aksis.uib.no/icame/manuals/brown/).

Ryu Iida, Mamoru Komachi, Kentaro Inui, and Yuji Matsumoto. 2007. Annotating a Japanese text corpus with predicate-argument and coreference relations. In *ACL Workshop: Linguistic Annotation Workshop*, pages 132–139. Prague.

Satoru Ikehara, Masahiro Miyazaki, Satoshi Shirai, Akio Yokoo, Hiromi Nakaiwa, Kentaro Ogura, Yoshifumi Ooyama, and Yoshihiko Hayashi. 1997. *Goi-Taikei — A Japanese Lexicon*. Iwanami Shoten, Tokyo. 5 volumes/CDROM.

Satoru Ikehara, Satoshi Shirai, Akio Yokoo, and Hiromi Nakaiwa. 1991. Toward an MT system without pre-editing — effects of new methods in **ALT-J/E** — . In *Third Machine Translation Summit: MT Summit III*, pages 101–106. Washington DC. URL http://xxx.lanl.gov/abs/cmp-lg/9510008.

Kow Kuroda, Jae-Ho Lee, Hajime Nozawa, Masaki Murata, and Kentaro Torisawa. 2009. Manual cleaning of hypernyms in Torishiki-Kai. In *15th Annual Meeting of The Association for Natural Language Processing*, pages C1–3. Tottori. (in Japanese).

Sadao Kurohashi and Makoto Nagao. 2003. Building a Japanese parsed corpus — while improving the parsing system. In Anne Abeillé, editor, *Treebanks: Building and Using Parsed Corpora*, chapter 14, pages 249–260. Kluwer Academic Publishers.

Jonathan Phillips. 2005. Introduction to the open clip art library. http://rejon.org/media/writings/ocalintro/ocal_intro_phillips.html. (accessed 2007-11-01).

Toshiyuki Takasaki and Yumiko Mori. 2007. Design and development of a pictogram communication system for children around the world. In *First International Workshop on Intercultural Collaboration (IWIC-2007)*, pages 144–157. Kyoto.

Kiyotaka Uchimoto, Yujie Zhang, Kiyoshi Sudo, Masaki Murata, Satoshi Sekine, and Hitoshi Isahara. 2004. Multilingual aligned parallel treebank corpus reflecting contextual information and its applications. In Gilles Sérasset, editor, *COLING 2004 Multilingual Linguistic Resources*, pages 57–64. COLING, Geneva, Switzerland. URL http://acl.ldc.upenn.edu/W/W04/W04-2208.bib.