多言語WordNetを利用した日本語WordNetの作成

Francis Bond　井佐原 均　神崎 享子　内元 清貴
情報通信研究機構
bond@ieee.org,{isahara,kanzaki,uchimoto}@nict.go.jp>

1 Introduction

The WordNet project at Princeton has been a resounding success creating a resource that is widely used in research (Fellbaum, 1998) and emulated in many languages (Vossen, 1998). In order for a lexical resource to be widely adopted it must be both **accesible** and **usable**. The Princeton WordNet is accessible due to its being released under a non-restrictive licence; and usable because it has not just precise information but also reasonable coverage, especially of common words.

Because of this success, there have been many projects to build wordnets for other languages. One of the first was the EuroWordNet project, which built wordnets for several European languages (Vossen, 1998). Unfortunately, most of the wordnets are neither as accesible as the Princeton WordNet, due to more restrictive licences, nor as usable due to more limited cover. Recently, the Global WordNet grid has tried to add even more languages, making the data as accesible as possible (Fellbaum and Vossen, 2007).

There have been several initiatives to create a Japanese wordnet, but none of them have yet produced something that is both accessible and usable. Hayashi (1999) created a translation of the entire noun part of the Princeton WordNet, including both synsets and glosses. This produced a very usable resource, but it was unfortunately not at all accessible. Koide et al. (2006) looked at combining EDR (EDR, 1990) with Princeton WordNet, but did not get beyond converting them both to RDF representations. Kaji and Watanabe (2006) presented a method of translating synsets from English to Japanese using corpus based contexts to improve accuracy, but only tested this on a few words. More recently, Cook (2008) produced a Multi-Lingual Semantic Network by translating monosemous parts of the Princeton WordNet into Japanese, Chinese and German. He also made an interface for browsing and amending the network. This data is accessible, as it is released under an open license, but loses a little on usability as most monosemous entries are for less frequent words.

The amount of previous work shows the great interest and value of producing a Japanese WordNet. We therefore decided to construct one as follows. First, automatically translate the Princeton WordNet into Japanese. Second, manually check the most frequent 20,000 synsets. Third, link the synsets to a corpus. Fourth, release the data under an open license. This WordNet is based on the structure of the English wordnet: Japanese near synonyms are added to the existing English synsets. For example, the English synset consisting of `seal#n#9` “any of numerous marine mammals that come on shore to breed; chiefly of cold regions”[1] has the following Japanese words associated with it: アザラシ *azarashi* “seal” and 海豹 *azarashi* “seal”. Adapting it more fully to Japanese is left to future research.

In this paper we deal with the first step and present a method to quickly and efficiently build an automatic first version of Wordnet. The straightforward way to do this is by looking up the English wordnet entries in an English-Japanese dictionary, and using their translations. The problem with this is that bilingual dictionaries are not marked with WordNet senses, if we look up *seal* we get over 30 entries, including 判子 *seal* “stamp” and 海軍特殊部隊 *gaiguntokushubutai* “Navy Seal”. We need to associate these candidates with the appropriate WordNet senses. Our method takes advantage of the existence of wordnets in multiple languages, and uses them to sense disambiguate the translations.

[1] All examples are from WordNet 3.0.

Part of		Number of Synsets		
Speech	English	French	Spanish	German
Noun	82,115	17,826	7,902	9,951
Verb	13,767	4,919	3,775	5,166
Adjective	18,156	0	3,879	15
Adverb	3,621	0	0	0
Total	117,659	22,745	15,556	15,132

Table 1: Sizes of the Wordnets used

2 Lexical Resources

2.1 Wordnets

We use four wordnets, summarized in Table 1. The largest is the English Wordnet v3.0 (Fellbaum, 1998) with 117,659 entries. The EuroWordnets are considerably smaller, ranging from 15,132 for German up to 22,745 for French (Vossen, 1998), consisting mainly of nouns with some verbs. All of them share the same structure — a collection of synsets joined to make a semantic network.

Because Wordnet keeps growing, both in size and complexity synsets can split up or even potentially merge across versions. The data for German was based on 1.5 and French and Spanish on 1.6. We mapped them into 3.0 using the mappings from Daude et al. (2003). When a synset mapped to more than one synset, we simply linked it to the most highly weighted one.

2.2 Lexicons

We use JMDict, the Japanese→Multilingual dictionary created by Jim Breen (Breen, 2004) for Japanese-English/French/German. We did not use its proper name dictionary, as wordnet does not have a lot of names. To supplement this we also used the EDR Japanese-English lexicon (`http://www2.nict.go.jp/r/r312/EDR/index.html`) and the last downloadable version of the Japanese-English Life Science Dictionary Project (v4) (`http://lsd.pharm.kyoto-u.ac.jp/ja/index.html`). For Japanese-Spanish, we used a small dictionary downloaded from `http://aulex.ohui.net/` (Goihata) and licensed under the GPL. The sizes of these lexicons are listed in Table 2.

The lexical resources, are, as always, not evenly distributed amongst the world's languages — Japanese-English has the most resources, followed by German, then French and then Spanish.

3 Creating the Japanese Wordnet

The approach we are taking to build the Japanese Wordnet is the standard **expand approach**: "translate WordNet synsets to another language and take over the structure" (Vossen, 2005). We did this both to keep a compatible structure with WordNet, and because we had access to a variety of resources to make the task easier.

Our main innovation is that we are using WordNets in multiple languages to disambiguate the Japanese translations, thus providing more reliable estimates.

Consider the following two synsets for *bat*, with their translation shown in Figure 1:

- bat#n#1, chiropteran (nocturnal mouselike mammal with forelimbs modified to form membranous wings . . .)
- bat#n#5 (a club used for hitting a ball in various games)

The Japanese-English lexicon has two translations for *bat* 蝙蝠 *koumori* "bat (mammal)" and バット *batto* "bat (club)". However, because there is no way of distinguishing between them we get a mixture of the meanings with 蝙蝠 *koumori* "bat#n#1" and バット *batto* "bat#n#5". *chiropteran* is not in any of the JE lexicons, and `bat#n#5` has no synonyms. Therefore using only the English Wordnet as source and Japanese⇔English lexicons there is no way to disambiguate them.

However, both synsets are also in the French wordnet: `bat#n#1` is *chauve-souris* and `bat#n#5` is *batte, gourdin*. These are not ambiguous in the same way: *chauve-souris* goes only to *koumori* and *batte* only to *batto*. Thus, if we can match through two languages, the mapping is much more likely to be the correct sense.

Similar approaches have been used to make new bilingual dictionaries: for example, linking Japanese-Malay through Japanese-English, English-Malay, Japanese-Chinese and Chinese-Malay (Bond et al., 2001). The difference here is that the original linking is done through the Wordnet synsets: we are effectively trying to translate a super-synset with synonyms in up to four languages (En, De, Fr, Es).

Part of	Number of Word-Pairs					
Speech	ja-en			ja-de	ja-fr	ja-es
	JMDict	EDR	Lifsci	JMDict	JMDict	Goihata
Noun	165,984	504,450	44,567	143,753	24,348	0
Verb	22,209	184,250	4,741	26,502	7,762	133
Adjective	16,861	44,961	11,212	17,121	4,582	70
Adverb	6,180	20,125	1,266	5,915	1,478	0
Unknown	3	0	0	0	0	3,548
Total	225,803	758,568	62,210	199,260	39,447	3,751

Table 2: Size and Distribution of the various Lexicons

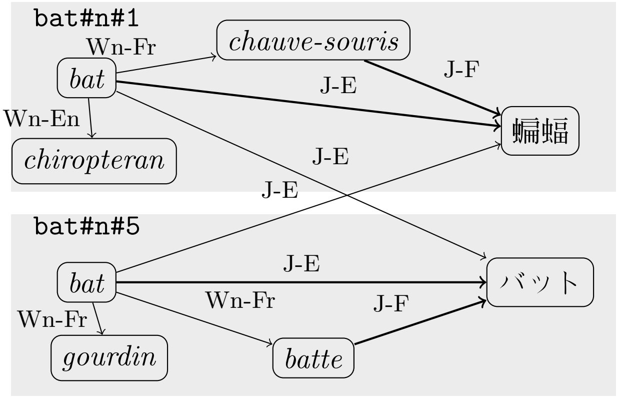

Figure 1: Linking with Multiple Wordnets

Part of	Number of Synsets		
Speech	$s > 10$	$s > 1$	All
Noun	9,243	36,432	42,725
Verb	2,991	9717	10,321
Adjective	629	6,283	8,915
Adverb	9	1,317	1,726
Total	12,872	53,749	63,687

Table 3: Japanese Synsets by score

The actual algorithm we used was as follows:

- For each synset in WordNet 3.0
  - Find equivalents in WN-{Fr,Es,De}
  - Look up translations for all equivalents {J$_e$}, {J$_f$}, {J$_s$}, {J$_d$}
  - Rank Japanese equivalents score $s = |\text{links}| + 10$ for links in two languages

The result is a wordnet with multiple Japanese candidates for most synsets, with a confidence score s equal to the number of bilingual links plus a ten-point bonus for being linked in multiple languages.

For example, for the two entries for *bat* given above, we end up with the following candidates:

- bat#n#1
  - 蝙蝠 (22: JMDict-en, EDR-en, JMDict-fr)
  - バット (3: JMDict-en (x2), EDR-en)
  - 蚊食い鳥 (1: EDR-en), ラケット (1: EDR-en), 打棒 (1: EDR-en), 蚊食鳥 (1: EDR-en), コウモリ (1: EDR-en)
- bat#n#5
  - バット (23: JMDict-en (x2), EDR-en, JMDict-fr)
  - 蝙蝠 (2: JMDict-en, EDR-en)
  - 蚊食い鳥 (1: EDR-en), ラケット (1: EDR-en), 打棒 (1: EDR-en), 蚊食鳥 (1: EDR-en), コウモリ (1: EDR-en)

Japanese words which belong in the synset (good matches) are underlined. The information in brackets is the score and the list of dictionaries used to match. As we expect, the words matching through two languages are correct, the remainder is a mixture of good and bad matches.

4 Results and Evaluation

In this section we report on how many synsets we could translate into Japanese, and with what confidence.

The results are summarized in Table 3. We have found some kind of translation for 63,687 out of the possible 117,007 synsets in Wordnet 3.0 (54.4%). Of these, the EuroWordnet data played

Part of	Number of Synsets		
Speech	$s > 10$	$s > 1$	All
Noun	2,429	3,264	3,279
Verb	656	988	993
Adjective	153	586	653
Adverb	0	0	0
Total	3,238	4,838	4,925

Table 4: Base Japanese Synsets by Score (s) for the Base Concepts

	Appropriate Translation Candidates			
	$s > 10$	$10 > s > 1$	$s = 1$	All
Base	55.30%	39.64%	21.25%	26.56%

Table 5: Base Noun Candidate Precision

a role in over 15,000 synsets. 12,872 synsets had at least one translation candidate confirmed in two or more languages, and 53,749 were confirmed in multiple lexicons.

The results restricted to the 5,000 common base concepts which occupy central positions in the wordnet structures (Fellbaum and Vossen, 2007) are given in Table 4. In this case our cover is almost complete (4,925/5,000 = 98.5%). Most of the entries in Euro WordNet are from these base concepts, and the majority of our translations (64.6%) match in two or more languages. Our coverage is excellent for the base synsets, and good overall: larger than any of the existing non-English WordNets.

To test precision, we evaluate translation candidates by judging the suitability of all of the base synset translation candidates (this is actually part of preparing the WordNet for release). The results are given in Table 5. Translations matching in multiple languages are markedly better than those matching in a single language (55.30% vs 21.25%), matching in multiple dictionaries in a single language is also an indicator of higher accuracy.

For both evaluation methods, translating synsets using multiple languages gives higher precision and lower recall.

5 Discussion and Future Work

In this paper we presented a method of automatically producing a Japanese WordNet of reasonable quantity by cross checking senses across different languages.

In future work we will manually check the most frequent synsets, sense annotate a small corpus and release the data. We hope it will then be used along with other projects such as the Global Wordnet Grid (Fellbaum and Vossen, 2007) and the Multi-lingual Semantic Network (Cook, 2008) to produce an even more useful resource.

References

Francis Bond and James Breen. 2007. Semi-automatic refinement of the JMdict/EDICT Japanese-English dictionary. In *13th Annual Meeting of The Association for Natural Language Processing*, pages 364–367. Kyoto.

Francis Bond, Ruhaida Binti Sulong, Takefumi Yamazaki, and Kentaro Ogura. 2001. Design and construction of a machine-tractable Japanese-Malay dictionary. In *MT Summit VIII*, pages 53–58. Santiago de Compostela, Spain.

J. W. Breen. 2004. JMDict: a Japanese-multilingual dictionary. In *Coling 2004 Workshop on Multilingual Linguistic Resources*, pages 71–78. Geneva.

Darren Cook. 2008. MLSN: A multi-lingual semantic network. In *14th Annual Meeting of the Association for Natural Language Processing*. Tokyo. (this volume).

Jordi Daude, Lluis Padro, and German Rigau. 2003. Validation and tuning of wordnet mapping techniques. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'03)*. Borovets, Bulgaria.

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

Christiane Fellbaum and Piek Vossen. 2007. Connecting the universal to the specific: Towards the global grid. In *First International Workshop on Intercultural Collaboration (IWIC-2007)*, pages 2–16. Kyoto.

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

Yoshihiko Hayashi. 1999. Translating WordNet noun part into Japanese for cross-language natural language applications. In *Technical Reports of SIG on Natural Language Processing NL130-10*, pages 73–80. (in Japanese).

Hiroyuki Kaji and Mariko Watanabe. 2006. Automatic construction of Japanese WordNet. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*. Genoa, Italy. URL http://www.sdjt.si/bib/lrec06/summaries/439.html.

Seiji Koide, Takeshi Morita, Takahira Yamaguchi, Hendry Muljadi, and Hideaki Takeda. 2006. OWL expressions on WordNet and EDR. In *AI society Semantic Web Ontology SIG 13*, SIG-SWO-A601-03. URL http://www.jaist.ac.jp/ks/labs/kbs-lab/sig-swo/fpapers.htm, (in Japanese).

Piek Vossen, editor. 1998. *Euro WordNet*. Kluwer.

Piek Vossen. 2005. Building wordnets. http://www.globalwordnet.org/gwa/BuildingWordnets.ppt.