Development of the Japanese WordNet

Hitoshi ISAHARA, Francis BOND, Kiyotaka UCHIMOTO,

Masao UTIYAMA, Kyoko KANZAKI

National Institute of Information and Communications Technology
3-5 Hikari-dai, Seika-cho, Souraku-gun, Kyoto 619-0289 Japan
E-mail: {isahara, bond, uchimoto, mutiyama, kanzaki}@nict.go.jp

Abstract

After a long history of compilation of our own lexical resources, EDR Japanese/English Electronic Dictionary, and discussions with major players on development of various WordNets, Japanese National Institute of Information and Communications Technology started developing the Japanese WordNet in 2006 and will publicly release the first version, which includes both the synset in Japanese and the annotated Japanese corpus of SemCor, in June 2008. As the first step in compiling the Japanese WordNet, we added Japanese equivalents to synsets of the Princeton WordNet. Of course, we must also add some synsets which do not exist in the Princeton WordNet, and must modify synsets in the Princeton WordNet, in order to make the hierarchical structure of Princeton synsets represent thesaurus-like information found in the Japanese language, however, we will address these tasks in a future study. We then translated English sentences which are used in the SemCor annotation into Japanese and annotated them using our Japanese WordNet. This article describes the overview of our project to compile Japanese WordNet and other resources which relate to our Japanese WordNet.

1. Introduction

After a long history of compilation of our own lexical resources, EDR Japanese/English Electronic Dictionary, and discussions with major players on development of various WordNets, we, Japanese National Institute of Information and Communications Technology (NICT), started developing the Japanese WordNet in 2006 and will publicly release the first version, which includes both the synset in Japanese and the annotated Japanese corpus of SemCor, in June 2008 just after the LREC2008 conference.

As the first step in compiling the Japanese WordNet, we added Japanese equivalents to synsets of the Princeton WordNet. Of course, we must also add some synsets which do not exist in the Princeton WordNet, and must modify synsets in the Princeton WordNet, in order to make the hierarchical structure of Princeton synsets represent thesaurus-like information found in the Japanese language, however, we will address these tasks in a future study.

We then translated English sentences which are used in the SemCor annotation into Japanese and annotated them using our Japanese WordNet[1].

This article describes the overview of our project to compile Japanese WordNet and other resources which relate to our Japanese WordNet.

2. Procedure

To develop and publicly release the Japanese WordNet, we made a draft plan to compile the first version of our Japanese WordNet in two years. We finished adding Japanese word to ten thousand WordNet synsets as a foundation of Japanese WordNet in 2006 Japanese fiscal year, and we are adding Japanese words to more synsets and compiling other resources related to our Japanese WordNet in 2007 Japanese fiscal year.

Stage 1 (2006)

As the first step of compiling the Japanese WordNet, we chose ten thousand synsets from the Princeton WordNet (WordNet 3.0) and add Japanese words to the synsets. To choose the ten thousand synsets, we firstly used the core synsets (4,959) of the WordNet. Then, we chose those synsets for which the total frequency of the words in the synset is high.

In order to make the development efficient, we automatically translate terms in existing WordNets, e.g., the Princeton English WordNet, French WordNet and Spanish WordNet into Japanese, by using electronic dictionaries, such as EDR English-Japanese dictionary. Our lexicographers are asked to choose proper Japanese words for a synset from a set of Japanese equivalents of terms in the synset, each of which has simple confidence score based on the number of languages and/or dictionaries which output the equivalents.

We tried to translate all English terms in the synset (117,659) into Japanese and could find Japanese equivalents for 62,832 synsets, among which we could find equivalents for 3,391 core synsets. This means that we could find candidate translations of terms in synsets by a simple dictionary lookup as shown in Table 1.

This was possible because terms in core synsets tend to be general terms and can be found in ordinary dictionaries.

[1] We are thinking to annotate newly introduced gloss annotation of WordNet with our Japanese WordNet.

	The number of synsets, for which at least one Japanese equivalent was found.	The total number of synsets.
All synsets in WN3.0	62,832	117,659
Core synsets	3,391	3,434
CBC	1,525	1,526
Core + CBC	4,916	4,960

Table 1. Dictionary lookup for Japanese equivalent

Another statistic we derived here is the semantic-concept relation (Vossen, 2004) between original synsets and our synsets. We checked the relations between terms in the original synsets and our additional Japanese terms. Among 4,916 synsets, Japanese terms completely matched in 1,521 synsets and partially matched in 2,997 synsets. Those synsets that did not completely match are the candidates of modification to make the hierarchical structure of existing synsets represent thesaurus-like information found in the Japanese language. We will consider adding some new Japanese synsets to the original synsets for those synsets that did not completely match.

As for the number of candidates of Japanese equivalents for all synsets, we obtained 215,630 Japanese terms as equivalents of terms in original synsets. We classified them into three types: equivalents obtained by more than two authorities (type 1) such as dictionaries, by two authorities (type 2), and by only one authority (type 3). As for type 1, among 16,620 terms, 9,042 terms (54%) are correct and used as Japanese terms in synsets. For type 2, among 26,268 terms, 13,986 terms (38%) are correct, and for type 3, among 162,641 terms, 33,908 terms (20%) are correct. This shows that if we use as many authorities as possible, the ratio of correct translations increases and therefore the efficiency of compilation of Japanese WordNet will be improved.

The actual procedure of assignment of equivalents is as follows:

1. We examine the remarks, examples and synonyms in the Princeton WordNet.
2. We consult an English-Japanese dictionary on the meanings of synonymous words.
3. We choose the Japanese equivalents which match the remarks in WordNet, by consulting a Japanese word dictionary.
4. We decide semantic conceptual relations among words.

The form we are using during the process described above is shown in Figure 1.

During the process of compiling the Japanese WordNet, we encountered several interesting problems:

- Representation of different transcription in Japanese. There are many orthographic variants in Japanese.
- Differences in part-of-speech system between Japanese and English, e.g., Japanese adjectives act differently from English adjectives.
- Differences of concept between Japanese and English, e.g., some of the meanings of the Japanese word “kioku (contents of memory, instrument of memory, action of memorization and so on)” match the meanings of “memory” in English, but others do not.
- Differences between concept structures in Japanese and English, for which we should add Japanese-peculiar synsets to the original wordnet.

We have finished developing the Japanese WordNet for ten thousand synsets and moved to Stage 2.

Stage 2 (2007)

As the second step, we are enlarging the Japanese WordNet and also are developing Japanese SemCor. As for enlarging the Japanese WordNet, we will cover ten thousand frequent words in the Juman dictionary and 25,060 synsets (224,260 words) which appear in the MultiSemCor corpus.

We will translate all text in MultiSemCor into Japanese and annotate them using our Japanese WordNet. We will check all lexical gaps that we will encounter during annotation of the Japanese SemCor.

We will complete all preparations for releasing the Japanese WordNet and Japanese annotated SemCor by June 2008.

3. Related Works

3.1 Previous researches

There are much work on building a Japanese WordNet.

- Noun part – synsets and glosses – translated into Japanese (Hayashi, 1999)
- Some entries translated using context (Kaji & Watanabe, 2006)
- Translation of (English) WordNet and EDR into RDF (Koide et al,. 2006)

But still no large-scale freely available Japanese WordNet

3.2 Ongoing related works

At NICT, we are conducting several researches related/using Japanese WordNet (Bond et al., 2008; Kanzaki et al., 2008; Charoenporn et al., 2008).

We are one of the members of Kyoto Project (Vossen et al., 2008) at FP7 in EU.

4. Future plan

We have four future plans for the development of Japanese WordNet.

One plan is adding synsets which do not exist in the Princeton WordNet, in order to make the hierarchical structure of Princeton synsets represent thesaurus-like information peculiar to Japanese language.

Second plan is combining WordNet with other Japanese lexical resources via our Japanese WordNet. NICT holds all copyrights of the EDR electronic dictionary which has 200 thousand words in Japanese and English and 400 thousand concepts constructed in a concept hierarchy. We have already linked 981 synsets of WordNet 1.6 to EDR concept entries. We will make this information available to the public with our Japanese WordNet. We are also planning to combine information in EDR with WordNet to create one big lexical resource with various kinds of information such as Japanese words, their part-of-speech, collocation information, and valence patterns. We are currently developing Japanese-Chinese dictionary as an extension of EDR Japanese-English dictionary. Once it is finished, we can also combine resources in Chinese with WordNet via EDR dictionary.

The third plan is to develop WordNet for south-east Asian languages. NICT has Thai Computational Linguistics Laboratory (TCL) in Thailand which develops NLP technology and resources in south-east Asia, such as Thai, Indonesia and Malay. We have developed a tool for collaboration named Knowledge Unifying Initiator (KUI) (Sornlertlamvanich, 2007) and started collaborative development of WordNet for the languages in the region using KUI as its platform.

The output of these activities will be publicly released in the future.

The fourth plan is much more ambitious. NICT has research groups on text processing, speech processing and image processing at NICT's Knowledge Creating Communication Research Center in Kyoto. We are thinking to develop a huge real-world knowledge base and our Japanese WordNet will be a core of this database.

5. References

Bond, Francis et al., Bootstrapping a WordNet using multiple existing WordNets., n *Proceedings of the 6th International Conference on Language Resources and Evaluation* (LREC 2008), 2008.

Charoenporn, Thatsanee et al. Semi-automatic Compilation of Asian WordNet, In *Proceedings of Annual Meeting of Japanese Society of NLP*, 2008.

Hayashi Yoshihiko, Translating WordNet noun part into Japanese for cross-language natural language appreciations. In *Technical Reports of SIG NLP NL130-10*, 1999 (in Japanese)

Koide Seiji et al, OWL expressions on WordNet and EDR, In *AI society Semantic Web Ontology SIG13, SIG-SWO-A601-03*, 2006

Kaji Hiroyuki & Watanabe Mariko, Automatic construction of Japanese WordNet. In *Proceedings of the 5th International Conference on Language Resources and Evaluation* (LREC 2006), 2006.

Kanzaki Kyoko et al., Extraction of Attribute Concepts from Japanese Adjectives, In *Proceedings of the 6th International Conference on Language Resources and Evaluation* (LREC 2008), 2008.

Sornlertlamvanich, Virach et al., Collaborative Platform for Multilingual Resource Development and Intercultural Communication, In: *Intercultural Collaboration*, Lecture Notes in Computer Science, Volume 4568/2007, Springer, 2007.

Vossen, Piek. EuroWordNet: a multilingual database of autonomous and language-specific wordnets connected via an Inter-Lingual-Index, In: *Special issue on multilingual databases. International Journal of Linguistics 17/2*, 2004.

Vossen, Piek et al., KYOTO: A System for Mining, Structuring, and Distributing Knowledge Across Languages and Cultures, In *Proceedings of the 6th International Conference on Language Resources and Evaluation* (LREC 2008), 2008.

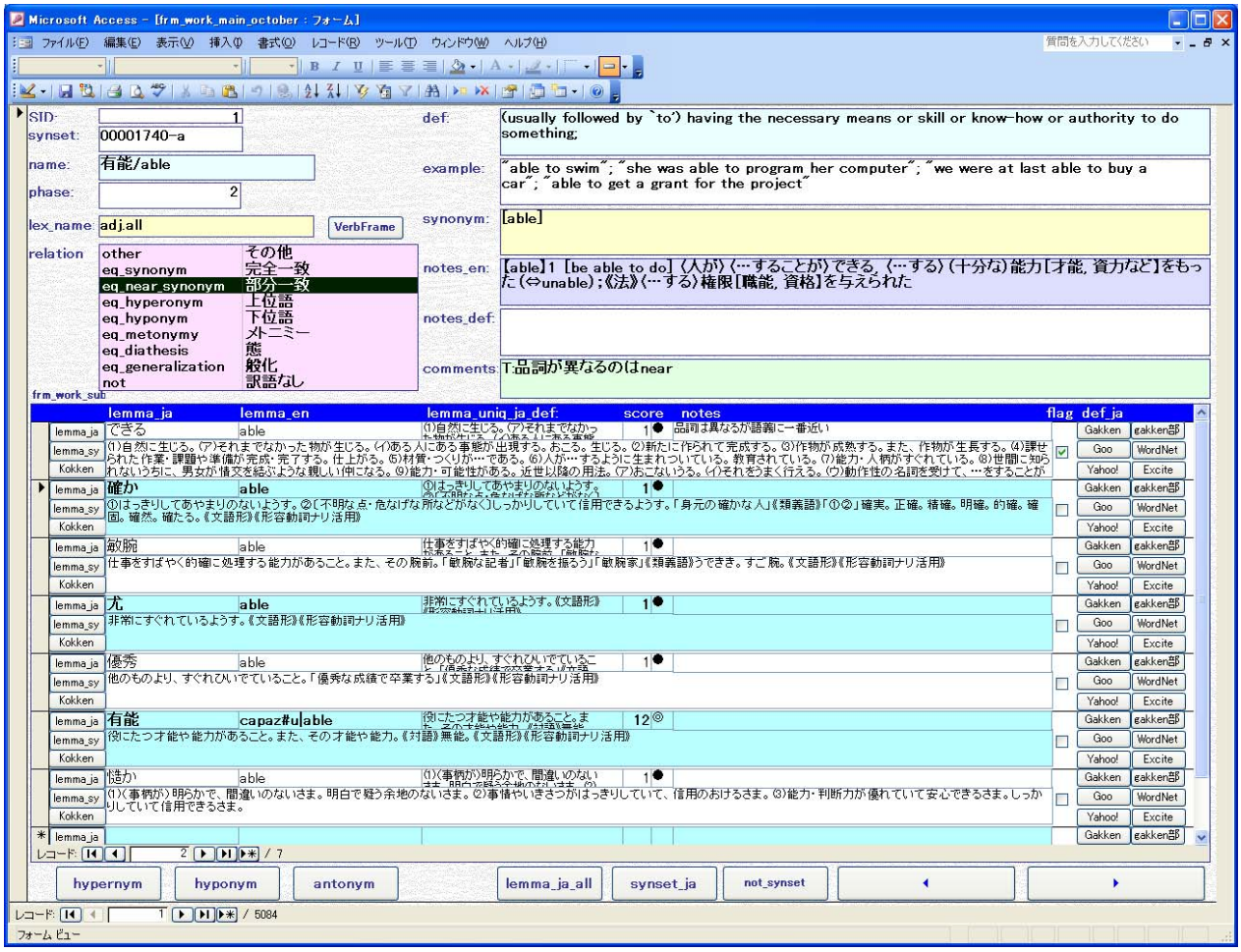

Figure 1. Entry Form for Compilation of Japanese WordNet