Acquiring Ontologies Using Deep and Shallow Processing

Eric Nichols† Francis Bond‡
†Nara Institute of Science and Technology
eric-n@is.aist-nara.ac.jp
‡Nippon Telegraph and Telephone Corporation
bond@cslab.kecl.ntt.co.jp

Abstract

In this paper, we outline the development of a system that automatically constructs ontologies for Japanese by extracting knowledge from dictionary definition sentences using Robust Minimal Recursion Semantics (RMRS), a semantic formalism that permits underspecification. We will show that by combining deep and shallow parsing resources through the common formalism of RMRS, we can extract ontological relations in greater quality and quantity. Our approach also has the advantages of requiring a small amount of rules and being easily adaptable to any language with RMRS resources. We give evaluation of the ontologies extracted by comparing them to WordNet and GoiTaikei.

1 Introduction

Ontologies are an important resource in natural language processing. They have been shown to be useful in tasks such as machine translation, question answering, and word-sense disambiguation, among others where information about the relationship and similarity of words can be exploited. While there are large, hand-crafted ontologies available for several languages, such as WordNet for English [6] and GoiTaikei for Japanese [8], these resources are difficult to construct and maintain entirely by hand.

There has been a great deal of work on the creation of ontologies from machine readable dictionaries (a good summary is [15]), mainly for English. Recently, there has also been interest in Japanese as well [14, 13, 1]. Most approaches use either a specialized parser or a set of regular expressions tuned to a particular dictionary, often with hundreds of rules.

In this paper, we take advantage of recent advances in both deep parsing and semantic representation to combine general purpose deep and shallow parsing technologies with a simple special relation extractor. Our basic approach is to parse dictionary definition sentences with multiple shallow and deep processors, generating semantic representations of varying specificity. The semantic representation is robust minimal recursion semantics (RMRS:Section 2.2.1). We then extract ontological relations using the most informative semantic representation for each definition sentence.

In this paper we discuss the construction of an ontology for Japanese using the the Japanese Semantic Database Lexeed [10]. The deep parser uses the Japanese Grammar JACY [12] and the shallow parser is based on the morphological analyzer ChaSen.

We carried out two evaluations. The first gives an automatically obtainable measure by comparing the extracted ontological relations by verifying the existence of the relations in exisiting WordNet [6] and GoiTaikei [8] ontologies. The second is a small scale human evaluation of the results.

2 Resources

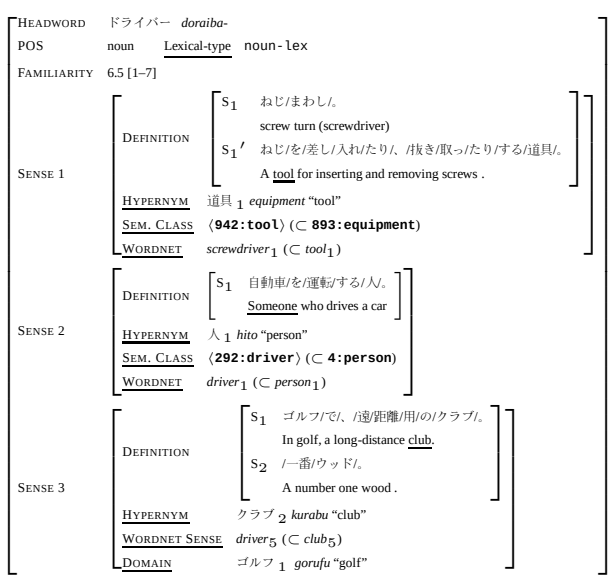

Figure 1: Entry for the Word *doraiba-* "driver" from Lexeed

2.1 Lexeed

The Lexeed Semantic Database of Japanese is a machine readable dictionary that covers the most common words in Japanese [10]. An example entry for the word ドライバー *doraibā* "driver" is given in Figure 1, with English glosses

added. The underlined material was not in Lexeed originally, we add it in this paper. Lexeed has 28,000 words divided into 46,000 senses and defined with 75,000 definition sentences.

2.2 Parsing Resources

We used the robust minimal recursion semantics (RMRS) designed in the Deep Thought project [4], along with tools from the Deep Linguistic Processing with HPSG Initiative (DELPH-IN: `http://www.delph-in.net/`).

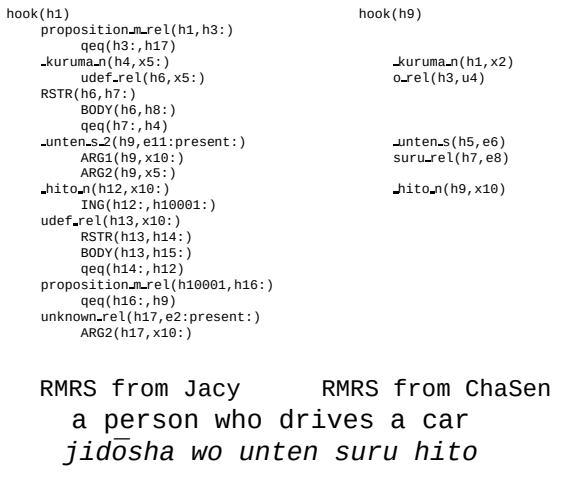

Figure 2: Deep and Shallow RMRS structures for *doraibā*$_2$

2.2.1 Robust Minimal Recursion Semantics

Robust Minimal Recursion Semantics is a form of flat semantics which is designed to allow deep and shallow processing to use a compatible semantic representation, while being rich enough to support generalized quantifiers [7]. The full representation is basically the same as minimal recursion semantics [5]: a bag of labeled elementary predicates and their arguments, a list of scoping constraints, and a handle that provides a hook into the representation. The main difference is that handles must be unique, and there is an explicit distinction between grammatical and real predicates.

Examples of deep and shallow results for the same sentence 自動車を運転する人 *jidōsha wo unten suru hito* "a person who drives a car (lit: car-ACC drive do person)" are given in Figure 2 (omitting the indexing). Real predicates are prefixed by an underbar (_). The deep parse gives information about the scope, message types and argument structure, while the shallow parse gives little more than a list of real and grammatical predicates with a hook.

2.2.2 Deep Parser (JACY and PET)

The Japanese grammar JACY [12] was run with the PET System for the high-efficiency processing of typed feature structures [3].

2.2.3 Shallow Parser (based on ChaSen)

ChaSen [11] was used for shallow processing of Japanese. Predicate names were produced by transliterating the pronunciation field and mapping the part-of-speech codes to the RMRS super types. The part-of-speech codes were also used to judge whether predicates were real or grammatical. Since Japanese is a head-final language, the hook value was set to be the handle of the rightmost real predicate.

3 Ontology Construction

Our approach to ontology construction is to process a definition sentence with shallow and deep parsers and extract ontological relations from the most informative RMRS output. Here, we will describe the algorithm we use to extract ontological relations from an RMRS structure:

1. Count the total number of non-grammatical predicates in the sentence
   - IF the total number of real predicates is one, return that predicate in the form: ⟨**synonym**: headword, predicate⟩
2. Initialize a stack of semantic relations to be processed with the semantic relation from the HOOK of most the informative RMRS structure for a given defi nition sentence
3. Pop a semantic relation from the stack and check it against special predicates that require additional processing
   - When a relation indicating coordination or conjunction is found, locate all of its arguments and push them onto the stack for processing
   - IF a special predicate is found, extract its relations and add them to the stack
   - ELSE IF the current semantic relation is a real predicate, add it to list of extracted semantic heads

   Repeat until the stack is empty
4. Return the ontological relations in the form: ⟨**relation**: headword, semantic_head⟩

Figure 3: Semantic Head Extraction algorithm

This processing is following in the long tradition of parsing such special relationships (also called "empty heads", "function nouns" or "relators") [15, Chapter 9]. The main innovation is to extract them from the semantic representations produced by a combination of deep and shallow parsing, rather than using either regular expressions or parsers designed specially to parse definition sentences.

4 Results and Evaluation

We summarize the relationships acquired in Table 2. The first two lines show thesaurus type relations: implicit hyper-

Results for Jacy

Relation	Nouns	Other	Verbs	Verbal Nouns	All
synonym	5185 / 6656 (77.90)	1120 / 2302 (48.65)	1212 / 2064 (58.72)	704 / 952 (73.95)	8221 / 11974 (68.66)
hypernym	10210 / 17073 (59.80)	1272 / 3829 (33.22)	1331 / 4224 (31.51)	524 / 1859 (28.19)	13337 / 26985 (49.42)
name	40 / 88 (45.45)	4 / 6 (66.67)	–/– (–.–)	–/– (–.–)	44 / 94 (46.81)
abbreviation	34 / 102 (33.33)	7 / 14 (50.00)	–/– (–.–)	4 / 11 (36.36)	45 / 127 (35.43)
meronym	102 / 219 (46.58)	6 / 17 (35.29)	–/– (–.–)	–/– (–.–)	108 / 237 (45.57)
Total	15571 / 24138 (64.51)	2409 / 6168 (39.06)	2543 / 6288 (40.44)	1232 / 2823 (43.64)	21755 / 39417 (55.19)

Results for ChaSen

Relation	Nouns	Other	Verbs	Verbal Nouns	All
synonym	4761 / 6015 (79.15)	1039 / 2037 (51.01)	1152 / 1881 (61.24)	676 / 904 (74.78)	7628 / 10837 (70.39)
hypernym	11260 / 23853 (47.21)	1650 / 5994 (27.53)	2195 / 6793 (32.31)	2795 / 6186 (45.18)	17900 / 42826 (41.80)
Total	16021 / 29868 (53.64)	2689 / 8031 (33.48)	3347 / 8674 (38.59)	3471 / 7090 (48.96)	25528 / 53663 (47.57)

Results for Deepest

Relation	Nouns	Other	Verbs	Verbal Nouns	All
synonym	5351 / 6913 (77.40)	1222 / 2579 (47.38)	1220 / 2084 (58.54)	720 / 977 (73.69)	8513 / 12553 (67.82)
hypernym	13988 / 26206 (53.38)	1949 / 6429 (30.32)	2375 / 7488 (31.72)	2947 / 6385 (46.16)	21259 / 46508 (45.71)
name	40 / 88 (45.45)	4 / 6 (66.67)	–/– (–)	–/– (–)	44 / 94 (46.81)
abbreviation	34 / 102 (33.33)	7 / 14 (50.00)	–/– (–)	4 / 11 (36.36)	45 / 127 (35.43)
meronym	102 / 219 (46.58)	6 / 17 (35.29)	–/– (–)	–/– (–)	108 / 237 (45.57)
Total	19515 / 33528 (58.21)	3188 / 9045 (35.25)	3595 / 9572 (37.56)	3671 / 7374 (49.78)	29969 / 59519 (50.35)

Table 2: Results confirmed for Lexeed (for 46,000 senses)

Special predicate	Ontological relation
isshu_n_1	hypernym
hitotsu_n_2	hypernym
soushou_n_1	hyponym
ryakushou_s_1	abbreviation
ryaku_s_1	abbreviation
keishou_n_1	name:honorifi c
zokushou_n_1	name:slang
meishou_n_1	name
bubun_n_1	meronym
ichibu_n_1	meronym

Table 1: Special predicates and associated relations

nyms and synonyms. The second three names show other relations: names, abbreviations, and meronyms. Implicit hypernyms and synonyms are by far the most common relations: fewer than 10% of entries are marked with an explicit relationship.

Results are shown for Lexeed, using only the RMRS produced by ChaSen, using the results for JACY, and using the deepest possible result (JACY if it exists, then backing off to ChaSen).

As one would expect, the word based analysis using ChaSen finds more actual relationships, but does not provide enough information to find anything beyond implicit hypernyms and synonyms. The grammar based analyses have lower coverage, but allow us extract some of the knowledge given explicitly in the lexicon.

We carried out two evaluations. The first was an automatic evaluation, comparing our extracted triples $\langle$`relation`: word1, word2$\rangle$ with existing resources. The second was a small scale hand evaluation of a sample of the relations.

4.1 Verification with Hand-crafted Ontologies

Because we are interested in comparing lexical semantics across languages, we compared the extracted ontology with resources in both the same and different languages.

We verified our results by comparing the hypernym links to the manually constructed Japanese ontology **GT**. It is a hierarchy of 2,710 semantic classes, defined for over 264,312 nouns [8]. The semantic classes are principally defined for nouns (including verbal nouns), although there is some information for verbs and adjectives. Senses are linked to **GT** semantic classes by the following heuristic: look up the semantic classes C for both the headword (w_i) and the genus term(s) (w_g). If at least one of the index word's semantic classes is subsumed by at least one of the genus' semantic classes, then we consider their relationship confirmed (1).

$$\exists(c_h, c_g) : \{c_h \subset c_g; c_h \in C(w_h); c_g \in C(w_g)\} \quad (1)$$

In the event of an explicit hyponym relationship indicated between the headword and the genus, the test is reversed: we look for an instance of the genus' class being subsumed by the headword's class ($c_g \subset c_h$).

To test cross-linguistically, we looked up the headwords in a translation lexicon (**ALT-J/E** [9] and EDICT [2]) and

then did the confirmation on the set of translations $c_i \subset C(T(w_i))$. Although looking up the translation adds noise, the additional information provided by the relationship triple effectively filters it out again.

For example, for ドライバー $_3$ *doraiba-*$_3$ "driver $_3$", **GT** does not find any relationship, as it does not have the golf club semantic class label for ドライバー *doraiba-*. However, looking up $T($ ドライバー $)$ gives {*driver, screwdriver*} and the extracted hypernym is クラブ $_2$ *kurabu* "club". WordNet recognizes that *driver* $_5$ is a kind of *wood*$_8$ which is a kind of *club*$_5$ (using senses and relations from WordNet 2.0 [6]). We thus simultaneously confirm the link is good; find an appropriate translation for this sense of ドライバー (and its hypernym); and link these to the appropriate WordNet synsems.

The results of the evaluation for lexeed are shown in Table 2. Relations verified in either **GT** or WordNet are classed as verified. Using the deepest RMRS results, we report a confirmation rate of 58.21% for nouns, 37.65% for verbs, and 50.35% for verbal nouns. This is comparable to [13], who reports 61.4% for nouns alone.

4.2 Human Evaluation

We conducted a hand-verification of a selection of our automatically acquired relations. 1,471 relations were selected using a stratified method over the entirety of our results (every 35th relationship, ordered by link-type and then headword). In this evaluation we only consider synonyms and any relationships extracted from the first sentence: the second and subsequent definition sentences tend to contain other information not relevant to hypernym relations. The results were then evaluated by native speakers of Japanese were given the definition word, the semantic head we retrieved, the posited relation type, and the original lexical entry and asked to evaluate if the relation was accurate.

The human judges found the relations presented to them to be accurate 88.99% of the time. In the 162 relations that were judged unacceptable, it was also determined that a relation did exist in 95 cases, but it was incorrect (i.e. a **`synonym`** in place of a **`hypernym`** and so on). These errors had three sources: the most common was a lack of identified explicit relationships; the next was lack of information from the shallow parse and the last was errors in the argument structure of the deep parse. [13] report slightly higher results for extracting noun relationships only (91.8%).

5 Discussion

We were able to successfully combine deep and shallow processing to extract ontological information from lexical resources. We showed that, by using a well defined semantic representation, the extraction can be generalized so much that it can be used on very different dictionaries from different languages. This is an improvement on the common approach to using more and more detailed regular expressions (e.g. [13]). Although this provides a quick start, the results are not generally reusable. In comparison, the ChaSen-RMRS engine is immediately useful for a variety of tasks.

The other innovation of our approach is the cross-lingual evaluation. As a by-product of the evaluation we enhance the existing resources (such as the GCIDE or WordNet) by linking them, so that information can be shared between them. Further, we hope to use the cross-lingual links to fill in gaps in the monolingual resources. Finally, we can trivially extract links from the **GT** ontology to WordNet, thus combining two useful resources and allowing us to compare them in detail.

6 Conclusion

We have demonstrated how deep and shallow processing techniques can be used together to enrich the acquisition of ontological information by constructing ontologies for English and Japanese. Our approach requires few rules and is thus easy to maintain and expand, and it can be easily extended to cover any language that has RMRS resources. In future research, we plan to extend our processing to retrieve more ontological relations and to investigate means of improving the accuracy of output of both deep and shallow processors.

References

[1] F. Bond, E. Nichols, S. Fujita, and T. Tanaka. Acquiring an ontology for a fundamental vocabulary. In *20th International Conference on Computational Linguistics: COLING-2004*, pages 1319–1325, Geneva, 2004.

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

[3] U. Callmeier. Preprocessing and encoding techniques in PET. In S. Oepen, D. Flickinger, J. Tsujii, and H. Uszkoreit, editors, *Collabarative Language Engineering*, chapter 6, pages 127–143. CSLI Publications, Stanford, 2002.

[4] U. Callmeier, A. Eisele, U. Schäfer, and M. Siegel. The deepthought core architecture framework. In *Proceedings of LREC-2004*, volume IV, Lisbon, 2004.

[5] A. Copestake, D. P. Flickinger, C. Pollard, and I. A. Sag. Minimal Recursion Semantics. An introduction. 2003.

[6] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[7] A. Frank. Constraint-based RMRS construction from shallow grammars. In *20th International Conference on Computational Linguistics: COLING-2004*, pages 1269–1272, Geneva, 2004.

[8] S. Ikehara, M. Miyazaki, S. Shirai, A. Yokoo, H. Nakaiwa, K. Ogura, Y. Ooyama, and Y. Hayashi. *Goi-Taikei — A Japanese Lexicon*. Iwanami Shoten, Tokyo, 1997. 5 volumes/CDROM.

[9] S. Ikehara, S. Shirai, A. Yokoo, and H. Nakaiwa. 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, 1991. (`http://xxx.lanl.gov/abs/cmp-lg/9510008`).

[10] K. Kasahara, H. Sato, F. Bond, T. Tanaka, S. Fujita, T. Kanasugi, and S. Amano. Construction of a Japanese semantic lexicon: Lexeed. SIG NLC-159, IPSJ, Tokyo, 2004. (in Japanese).

[11] Y. Matsumoto, Kitauchi, Yamashita, Hirano, Matsuda, and Asahara. *Nihongo Keitaiso Kaiseki System: Chasen*, version 2.2.1 manual edition, 2000. url=http://chasen.aist-nara.ac.jp.

[12] M. Siegel and E. M. Bender. Effi cient deep processing of Japanese. In *Procedings of the 3rd Workshop on Asian Language Resources and International Standardization at the 19th International Conference on Computational Linguistics*, Taipei, 2002.

[13] T. Tokunaga, Y. Syotu, H. Tanaka, and K. Shirai. Integration of heterogeneous language resources: A monolingual dictionary and a thesaurus. In *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium, NLPRS2001*, pages 135–142, Tokyo, 2001.

[14] H. Tsurumaru, K. Takesita, I. Katsuki, T. Yanagawa, and S. Yoshida. An approach to thesaurus construction from Japanese language dictionary. In *IPSJ SIGNotes Natural Language*, volume 83-16, pages 121–128, 1991. (in Japanese).

[15] Y. A. Wilkes, B. M. Slator, and L. M. Guthrie. *Electric Words*. MIT Press, 1996.