Multilingual Ontology Acquisition from Multiple MRDs

Eric Nichols[♭], **Francis Bond**[♮], **Takaaki Tanaka**[♮], **Sanae Fujita**[♮], **Dan Flickinger** [♯]

♭ Nara Inst. of Science and Technology
Grad. School of Information Science
Nara, Japan
eric-n@is.naist.jp

♮ NTT Communication Science Labs
Natural Language Research Group
Keihanna, Japan
{bond,takaaki,sanae}@cslab.kecl.ntt.co.jp

♯ Stanford University
CSLI
Stanford, CA
danf@csli.stanford.edu

Abstract

In this paper, we outline the development of a system that automatically constructs ontologies by extracting knowledge from dictionary definition sentences using Robust Minimal Recursion Semantics (RMRS). Combining deep and shallow parsing resource through the common formalism of RMRS allows us to extract ontological relations in greater quantity and quality than possible with any of the methods independently. Using this method, we construct ontologies from two different Japanese lexicons and one English lexicon. We then link them to existing, hand-crafted ontologies, aligning them at the word-sense level. This alignment provides a representative evaluation of the quality of the relations being extracted. We present the results of this ontology construction and discuss how our system was designed to handle multiple lexicons and languages.

1 Introduction

Automatic methods of ontology acquisition have a long history in the field of natural language processing. The information contained in ontologies is important for a number of tasks, for example word sense disambiguation, question answering and machine translation. In this paper, we present the results of experiments conducted in automatic ontological acquisition over two languages, English and Japanese, and from three different machine-readable dictionaries.

Useful semantic relations can be extracted from large corpora using relatively simple patterns (e.g., (Pantel et al., 2004)). While large corpora often contain information not found in lexicons, even a very large corpus may not include all the familiar words of a language, let alone those words occurring in useful patterns (Amano and Kondo, 1999). Therefore it makes sense to also extract data from machine readable dictionaries (MRDs).

There is a great deal of work on the creation of ontologies from machine readable dictionaries (a good summary is (Wilkes et al., 1996)), mainly for English. Recently, there has also been interest in Japanese (Tokunaga et al., 2001; Nichols et al., 2005). Most approaches use either a specialized parser or a set of regular expressions tuned to a particular dictionary, often with hundreds of rules. Agirre et al. (2000) extracted taxonomic relations from a Basque dictionary with high accuracy using Constraint Grammar together with hand-crafted rules. However, such a system is limited to one language, and it has yet to be seen how the rules will scale when deeper semantic relations are extracted. In comparison, as we will demonstrate, our system produces comparable results while the framework is immediately applicable to any language with the resources to produce RMRS. Advances in the state-of-the-art in parsing have made it practical to use deep processing systems that produce rich syntactic and semantic analyses to parse lexicons. This high level of semantic information makes it easy to identify the relations between words that make up an ontology. Such an approach was taken by the MindNet project (Richardson et al., 1998). However, deep parsing systems often suffer from small lexicons and large amounts of parse ambiguity, making it difficult to apply this knowledge broadly.

Our ontology extraction system uses Robust Minimal Recursion Semantics (RMRS), a formalism that provides a high level of detail while, at the same time, allowing for the flexibility of underspecification. RMRS encodes syntactic information in a general enough manner to make processing of and extraction from syntactic phenomena including coordination, relative clause analy-

Proceedings of the 2nd Workshop on Ontology Learning and Population, pages 10–17,
Sydney, July 2006. ©2006 Association for Computational Linguistics

sis and the treatment of argument structure from verbs and verbal nouns. It provides a common format for naming semantic relations, allowing them to be generalized over languages. Because of this, we are able to extend our system to cover new languages that have RMRS resourses available with a minimal amount of effort. The underspecification mechanism in RMRS makes it possible for us to produce input that is compatible with our system from a variety of different parsers. By selecting parsers of various different levels of robustness and informativeness, we avoid the coverage problem that is classically associated with approaches using deep-processing; using heterogeneous parsing resources maximizes the quality and quantity of ontological relations extracted. Currently, our system uses input from parsers from three levels: with morphological analyzers the shallowest, parsers using Head-driven Phrase Structure Grammars (HPSG) the deepest and dependency parsers providing a middle ground.

Our system was initially developed for one Japanese dictionary (Lexeed). The use of the abstract formalism, RMRS, made it easy to extend to a different Japanese lexicon (Iwanami) and even a lexicon in a different language (GCIDE).

Section 2 provides a description of RMRS and the tools used by our system. The ontological acquisition system is presented in Section 3. The results of evaluating our ontologies by comparison with existing resources are given in Section 4. We discuss our findings in Section 5.

2 Resources

2.1 The Lexeed Semantic Database of Japanese

The Lexeed Semantic Database of Japanese is a machine readable dictionary that covers the most familiar open class words in Japanese as measured by a series of psycholinguistic experiments (Kasahara et al., 2004). Lexeed consists of all open class words with a familiarity greater than or equal to five on a scale of one to seven. This gives 28,000 words divided into 46,000 senses and defined with 75,000 definition sentences. All definition sentences and example sentences have been rewritten to use only the 28,000 familiar open class words. The definition and example sentences have been treebanked with the JACY grammar (§ 2.4.2).

2.2 The Iwanami Dictionary of Japanese

The Iwanami Kokugo Jiten (Iwanami) (Nishio et al., 1994) is a concise Japanese dictionary. A machine tractable version was made available by the Real World Computing Project for the SENSEVAL-2 Japanese lexical task (Shirai, 2003). Iwanami has 60,321 headwords and 85,870 word senses. Each sense in the dictionary consists of a sense ID and morphological information (word segmentation, POS tag, base form and reading, all manually post-edited).

2.3 The Gnu Contemporary International Dictionary of English

The GNU Collaborative International Dictionary of English (GCIDE) is a freely available dictionary of English based on Webster's Revised Unabridged Dictionary (published in 1913), and supplemented with entries from WordNet and additional submissions from users. It currently contains over 148,000 definitions. The version used in this research is formatted in XML and is available for download from `www.ibiblio.org/webster/`.

We arranged the headwords by frequency and segmented their definition sentences into sub-sentences by tokenizing on semicolons (;). This produced a total of 397,460 pairs of headwords and sub-sentences, for an average of slightly less than four sub-sentences per definition sentence. For corpus data, we selected the first 100,000 definition sub-sentences of the headwords with the highest frequency. This subset of definition sentences contains 12,440 headwords with 36,313 senses, covering approximately 25% of the definition sentences in the GCIDE. The GCIDE has the most polysemy of the lexicons used in this research. It averages over 3 senses per word defined in comparison to Lexeed and Iwanami which both have less than 2.

2.4 Parsing Resources

We used Robust Minimal Recursion Semantics (RMRS) designed as part of the Deep Thought project (Callmeier et al., 2004) as the formalism for our ontological relation extraction engine. We used deep-processing tools from the Deep Linguistic Processing with HPSG Initiative (DELPH-IN: `http://www.delph-in.net/`) as well as medium- and shallow-processing tools for Japanese processing (the morphological analyzer

ChaSen and the dependency parser CaboCha) from the Matsumoto Laboratory.

2.4.1 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, with fine-grained atomic components of semantic content so shallow methods can contribute just what they know, yet with enough expressive power for rich semantic content including generalized quantifiers (Frank, 2004). The architecture of the representation is based on Minimal Recursion Semantics (Copestake et al., 2005), including a bag of labeled elementary predicates (EPs) and their arguments, a list of scoping constraints which enable scope underspecification, and a handle that provides a hook into the representation.

The representation can be underspecified in three ways: relationships can be omitted (such as quantifiers, messages, conjunctions and so on); predicate-argument relations can be omitted; and predicate names can be simplified. Predicate names are defined in such a way as to be as compatible (predictable) as possible among different analysis engines, using a lemma_pos_subsense naming convention, where the subsense is optional and the part-of-speech (pos) for coarse-grained sense distinctions is drawn from a small set of general types (**n**oun, **v**erb, **s**ahen (verbal noun), . . .). The predicate `unten_s` (運転 *unten* “drive”), for example, is less specific than `unten_s_2` and thus subsumes it. In order to simplify the combination of different analyses, the EPs are indexed to the corresponding character positions in the original input sentence.

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 Figures 1 and 2 (omitting the indexing). Real predicates are prefixed by an under-bar (_). 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.

2.4.2 Deep Parsers (JACY, ERG and PET)

For both Japanese and English, we used the PET System for the high-efficiency processing of typed feature structures (Callmeier, 2000). For Japanese, we used JACY (Siegel, 2000), for English we used the English Resource Grammar (ERG: Flickinger 2000).[1]

JACY The JACY grammar is an HPSG-based grammar of Japanese which originates from work done in the Verbmobil project (Siegel, 2000) on machine translation of spoken dialogues in the domain of travel planning. It has since been extended to accommodate written Japanese and new domains (such as electronic commerce customer email and machine readable dictionaries).

The grammar implementation is based on a system of types. There are around 900 lexical types that define the syntactic, semantic and pragmatic properties of the Japanese words, and 188 types that define the properties of phrases and lexical rules. The grammar includes 50 lexical rules for inflectional and derivational morphology and 47 phrase structure rules. The lexicon contains around 36,000 lexemes.

The English Resource Grammar (ERG) The English Resource Grammar (ERG: (Flickinger, 2000)) is a broad-coverage, linguistically precise grammar of English, developed within the Head-driven Phrase Structure Grammar (HPSG) framework, and designed for both parsing and generation. It was also originally launched within the Verbmobil (Wahlster, 2000) spoken language machine translation project for the particular domains of meeting scheduling and travel planning. The ERG has since been substantially extended in both grammatical and lexical coverage, reaching 80-90% coverage of sizeable corpora in two additional domains: electronic commerce customer email and tourism brochures.

The grammar includes a hand-built lexicon of 23,000 lemmas instantiating 850 lexical types, a highly schematic set of 150 grammar rules, and a set of 40 lexical rules, all organized in a rich multiple inheritance hierarchy of some 3000 typed feature structures. Like other DELPH-IN grammars, the ERG can be processed by several parsers and generators, including the LKB (Copestake, 2002) and PET (Callmeier, 2000). Each successful ERG analysis of a sentence or fragment includes a fine-grained semantic representation in MRS.

For the task of parsing the dictionary definitions in GCIDE (the GNU Collaborative Interna-

[1] Both grammars, the LKB and PET are available at `<http://www.delph-in.net/>`.

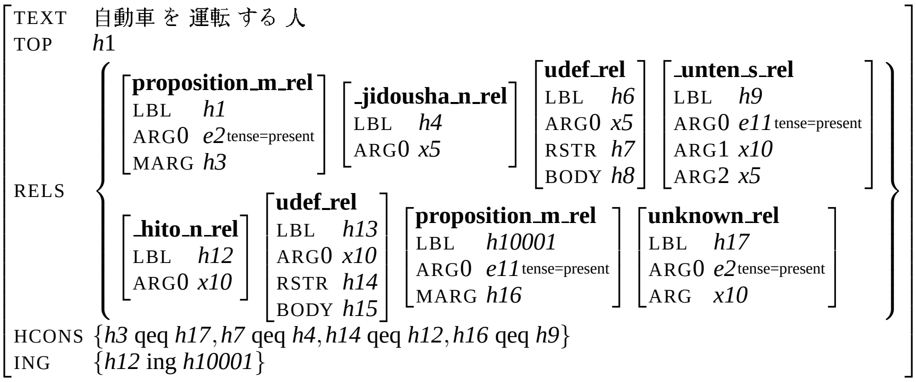
TEXT 自動車 を 運転 する 人
TOP h1
RELS
proposition_m_rel LBL h1 ARG0 e2 tense=present MARG h3
_jidousha_n_rel LBL h4 ARG0 x5
udef_rel LBL h6 ARG0 x5 RSTR h7 BODY h8
_unten_s_rel LBL h9 ARG0 e11 tense=present ARG1 x10 ARG2 x5
_hito_n_rel LBL h12 ARG0 x10
udef_rel LBL h13 ARG0 x10 RSTR h14 BODY h15
proposition_m_rel LBL h10001 ARG0 e11 tense=present MARG h16
unknown_rel LBL h17 ARG0 e2 tense=present ARG x10
HCONS {h3 qeq h17, h7 qeq h4, h14 qeq h12, h16 qeq h9}
ING {h12 ing h10001}

Figure 1: RMRS for the Sense 2 of *doraiba-* "driver" (Cabocha/JACY)

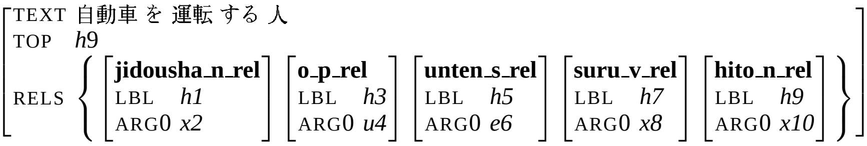
TEXT 自動車 を 運転 する 人
TOP h9
RELS
jidousha_n_rel LBL h1 ARG0 x2
o_p_rel LBL h3 ARG0 u4
unten_s_rel LBL h5 ARG0 e6
suru_v_rel LBL h7 ARG0 x8
hito_n_rel LBL h9 ARG0 x10

Figure 2: RMRS for the Sense 2 of *doraiba-* "driver" (ChaSen)

tional Dictionary of English; see below), the ERG was minimally extended to include two additional fragment rules, for gap-containing VPs and PPs (idiosyncratic to this domain), and additional lexical entries were manually added for all missing words in the alphabetically first 10,000 definition sentences.

These first 10,000 sentences were parsed and then manually tree-banked to provide the training material for constructing the stochastic model used for best-only parsing of the rest of the definition sentences. Using POS-based unknown-word guessing for missing lexical entries, MRSes were obtained for about 75% of the first 100,000 definition sentences.

2.4.3 Medium Parser (CaboCha-RMRS)

For Japanese, we produce RMRS from the dependency parser Cabocha (Kudo and Matsumoto, 2002). The method is similar to that of Spreyer and Frank (2005), who produce RMRS from detailed German dependencies. CaboCha provides fairly minimal dependencies: there are three links (dependent, parallel, apposition) and they link base phrases (Japanese *bunsetsu*), marked with the syntactic and semantic head. The CaboCha-RMRS parser uses this information, along with heuristics based on the parts-of-speech, to produce underspecified RMRSs. CaboCha-RMRS is capable of making use of HPSG resources, including verbal case frames, to further enrich its output. This allows it to produce RMRS that approaches the granularity of the analyses given by HPSG parsers. Indeed, CaboCha-RMRS and JACY give identical parses for the example sentence in Figure 1. One of our motivations in including a medium parser in our system is to extract more relations that require special processing; the flexibility of CaboCha-RMRS and the RMRS formalism make this possible.

2.4.4 Shallow Parser (ChaSen-RMRS)

The part-of-speech tagger, ChaSen (Matsumoto et al., 2000) 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 right-most real predicate. This is easy to do for Japanese, but difficult for English.

3 Ontology Construction

We adopt the ontological relation extraction algorithm used by Nichols et al. (2005). Its goal is to identify the semantic head(s) of a dictionary definition sentence – the relation(s) that best summarize it. The algorithm does this by traversing the RMRS structure of a given definition sentence starting at the HOOK (the highest-scoping semantic relationship) and following its argument structure. When the algorithm can proceed no further, it returns the a tuple consisting of the definition word and the word identified by the semantic relation where the algorithm halted. Our extended algorithm has the following characteristics: sentences with only one content-bearing relation are assumed to identify a synonym; special relation processing (§ 3.1) is used to gather meta-information and identify ontological relations; processing of coordination allows for extraction of multiple ontological relations; filtering by part-of-speech screens out unlikely relations (§ 3.2).

3.1 Special Relations

Occasionally, relations which provide ontological meta-information, such as the specification of domain or temporal expressions, or which help identify the type of ontological relation present are encountered. Nichols et al. (2005) identified these as **special relations**. We use a small number of rules to determine where the semantic head is and what ontological relation should be extracted. A sample of the special relations are listed in Table 1. This technique follows in a long tradition of special treatment of certain words that have been shown to be particularly relevant to the task of ontology construction or which are semantically content-free. These words or relations have also be referred to as “empty heads”, “function nouns”, or “relators” in the literature (Wilkes et al., 1996). Our approach generalizes the treatment of these special relations to rules that are portable for any RMRS (modulo the language specific predicate names) giving it portability that cannot be found in approaches that use regular expressions or specialized parsers.

Special Predicate (s)		Ontological Relation
Japanese	English	
isshu, hitotsu	form, kind, one	`hypernym`
ryaku(shou)	abbreviation	`abbreviation`
bubun, ichibu	part, peice	`meronym`
meishou	name	`name`
keishou	’polite name for’	`name:honorific`
zokushou	’slang for’	`name:slang`

Table 1: Special predicates and their associated ontological relations

Augmenting the system to work on English definition sentence simply entailed writing rules to handle special relations that occur in English. Our system currently has 26 rules for Japanese and 50 rules for English. These rules provide processing of relations like those found in Table 1, and they also handle processing of coordinate structures, such as noun phrases joined together with conjunctions such as *and*, *or*, and punctuation.

3.2 Filtering by Part-of-Speech

One of the problems encountered in expanding the approach in Nichols et al. (2005) to handle English dictionaries is that many of the definition sentences have a semantic head with a part-of-speech different than that of the definition word. We found that differing parts-of-speech often indicated an undesirable ontological relation. One reason such relations can be extracted is when a sentence with a non-defining role, for example indicating usage, is encountered. Definition sentence for non-content-bearing words such as *of* or *the* also pose problems for extraction.

We avoid these problems by filtering by parts-of-speech twice in the extraction process. First, we select candidate sentences for extraction by verifying that the definition word has a content word POS (i.e. adjective, adverb, noun, or verb). Finally, before we extract any ontological relation, we make sure that the definition word and the semantic head are in compatible POS classes.

While adopting this strategy does reduce the number of total ontological relations that we acquire, it increases their reliability. The addition of a medium parser gives us more RMRS structures to extract from, which helps compensate for any loss in number.

4 Results and Evaluation

We summarize the relationships acquired in Table 2. The columns specify source dictionary and parsing method while the rows show the relation type. These counts represent the total number of relations extracted for each source and method combination. The majority of relations extracted are synonyms and hypernyms; however, some higher-level relations such as meronym and abbreviation are also acquired. It should also be noted that both the medium and deep methods were able to extract a fair number of special relations. In many cases, the medium method even extracted more special relations than the deep method. This is yet another indication of the flexibility of dependency parsing. Altogether, we extracted 105,613 unique relations from Lexeed (for 46,000 senses), 183,927 unique relations from Iwanami (for 85,870 senses), and 65,593 unique relations from GCIDE (for 36,313 senses). As can be expected, a general pattern in our results is that the shallow method extracts the most relations in total followed by the medium method, and finally

Relation	**Lexeed**			**Iwanami**			**GCIDE**
	Shallow	Medium	Deep	Shallow	Medium	Deep	Deep
`hypernym`	47,549	43,006	41,553	113,120	113,433	66,713	40,583
`synonym`	12,692	13,126	9,114	31,682	32,261	18,080	21,643
`abbreviation`		340	429		1,533	739	
`meronym`		235	189		395	202	472
`name`		100	89		271	140	

Table 2: Results of Ontology Extraction

the deep method.

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.

For Japanese we verified our results by comparing the hypernym links to the manually constructed Japanese ontology Goi-Taikei (**GT**). It is a hierarchy of 2,710 semantic classes, defined for over 264,312 nouns Ikehara et al. (1997). The semantic classes are mostly defined for nouns (and verbal nouns), although there is some information for verbs and adjectives. For English, we compared relations to WordNet 2.0 (Fellbaum, 1998). Comparison for hypernyms is done as follows: look up the semantic class or synset C for both the headword (w_i) and genus term(s) (w_g). If at least one of the index word's classes is subsumed by at least one of the genus' classes, then we consider the relationship confirmed (1).

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

To test cross-linguistically, we looked up the headwords in a translation lexicon (**ALT-J/E** (Ikehara et al., 1991) and EDICT (Breen, 2004)) 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 filter of the relationship triple effectively filters it out again.

The total figures given in Table 3 do not match the totals given in Table 2. These totals represent the number of relations where both the definition word and semantic head were found in at least one of the ontologies being used in this comparison. By comparing these numbers to the totals given in Section 4, we can get an idea of the coverage of the ontologies being used in comparison. Lexeed has a coverage of approx. 55.74% ($\frac{58,867}{105,613}$), with Iwanami the lowest at 48.20% ($\frac{88,662}{183,927}$), and GCIDE the highest at 69.85% ($\frac{45,814}{65,593}$). It is clear that there are a lot of relations in each lexicon that are not covered by the hand-crafted ontologies. This demonstrates that machine-readable dictionaries are still a valuable resource for constructing ontologies.

4.1.1 Lexeed

Our results using JACY achieve a confirmation rate of **66.84%** for nouns only and **60.67%** overall (Table 3). This is an improvement over both Tokunaga et al. (2001), who reported 61.4% for nouns only, and Nichols et al. (2005) who reported 63.31% for nouns and 57.74% overall. We also achieve an impressive 33,333 confirmed relations for a rate of 56.62% overall. It is important to note that our total counts include all unique relations regardless of source, unlike Nichols et al. (2005) who take only the relation from the deepest source whenever multiple relations are extracted. It is interesting to note that shallow processing out performs medium with 22,540 verified relations (59.40%) compared to 21,806 (57.76%). This would seem to suggest that for the simplest task of retrieving hyperynms and synonyms, more information than that is not necessary. However, since medium and deep parsing obtain relations not covered by shallow parsing and can extract special relations, a task that cannot be performed without syntactic information, it is beneficial to use them as well.

Agirre et al. (2000) reported an error rate of 2.8% in a hand-evaluation of the semantic relations they automatically extracted from a machine-readable Basque dictionary. In a similar hand-evaluation of a stratified sampling of relations extracted from Lexeed, we achieved an error rate of 9.2%, demonstrating that our method is also highly accurate (Nichols et al., 2005).

4.2 Iwanami

Iwanami's verification results are similar to Lexeed's (Table 3). There are on average around 3% more verifications and a total of almost 20,000 more verified relations extracted. It is particularly interesting to note that deep processing per-

Confirmed Relations in Lexeed

Method / Relation	hypernym	synonym	Total
Shallow	58.55 % (16585 / 28328)	61.93 % (5955 / 9615)	59.40 % (22540 / 37943)
Medium	55.97 % (15431 / 27570)	62.61 % (6375 / 10182)	57.76 % (21806 / 37752)
Deep	54.78 % (4954 / 9043)	67.76 % (5098 / 7524)	60.67 % (10052 / 16567)
All	55.22 % (23802 / 43102)	60.46 % (9531 / 15765)	56.62 % (33333 / 58867)

Confirmed Relations in Iwanami

Method / Relation	hypernym	synonym	Total
Shallow	61.20 % (35208 / 57533)	63.57 % (11362 / 17872)	61.76 % (46570 / 75405)
Medium	60.69 % (35621 / 58698)	62.86 % (11037 / 17557)	61.19 % (46658 / 76255)
Deep	63.59 % (22936 / 36068)	64.44 % (8395 / 13027)	63.82 % (31331 / 49095)
All	59.36 % (40179 / 67689)	61.66 % (12931 / 20973)	59.90 % (53110 / 88662)

Confirmed Relations in GCIDE

POS / Relation	hypernym	synonym	Total
Adjective	2.88 % (37 / 1283)	16.77 % (705 / 4203)	13.53 % (742 / 5486)
Noun	57.60 % (7518 / 13053)	50.71 % (3522 / 6945)	55.21 % (11040 / 19998)
Verb	24.22 % (3006 / 12411)	21.40 % (1695 / 7919)	23.12 % (4701 / 20330)
Total	39.48 % (10561 / 26747)	31.06 % (5922 / 19067)	35.98 % (16483 / 45814)

Table 3: Confirmed Relations, measured against **GT** and WordNet

forms better here than on Lexeed (63.82% vs 60.67%), even though the grammar was developed and tested on Lexeed. There are two reasons for this: The first is that the process of rewriting Lexeed to use only familiar words actually makes the sentences harder to parse. The second is that the less familiar words in Iwanami have fewer senses, and easier to parse definition sentences. In any case, the results support our claims that our ontological relation extraction system is easily adaptable to new lexicons.

4.3 GCIDE

At first glance, it would seem that GCIDE has the most disappointing of the verification results with overall verification of not even 36% and only 16,483 relations confirmed. However, on closer inspection one can see that noun hypernyms are a respectable 57.60% with over 55% for all nouns. These figures are comparable with the results we are obtaining with the other lexicons. One should also bear in mind that the definitions found in GCIDE can be archaic; after all this dictionary was first published in 1913. This could be one cause of parsing errors for ERG. Despite these obstacles, we feel that GCIDE has a lot of potential for ontological acquisition. A dictionary of its size and coverage will most likely contain relations that may not be represented in other sources. One only has to look at the definition of ドライバー "driver"/*driver* to confirm this; **GT** has two senses ("screwdriver" and "vehicle operator") Lexeed and Iwanami have 3 senses each (adding "golf club"), and WordNet has 5 (including "software driver"), but GCIDE has 6, not including "software driver" but including *spanker* "a kind of sail". It should be beneficial to propagate these different senses across ontologies.

5 Discussion and Future Work

We were able to successfully combine deep processing of various levels of depth in order 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. Tokunaga et al. (2001)). Although this provides a quick start, the results are not generally reusable. In comparison, the shallower RMRS engines are immediately useful for a variety of other tasks.

However, because the hook is the only syntactic information returned by the shallow parser, ontological relation extraction is essentially performed by this hook-identifying heuristic. While this is sufficient for a large number of sentences, it is not possible to process special relations with the shallow parser since none of the arguments are linked with the predicates to which they belong. Thus, as Table 2 shows, our shallow parser is only capable of retrieving hypernyms and synonyms. It is important to extract a variety of semantic relations in order to form a useful ontology. This is one of the reasons why we use a combination of parsers of different analytic levels rather than depending on a single resource.

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 **GT** or WordNet) by linking them, so that information can be shared between them. In this way we can use the cross-lingual links to fill gaps in the monolingual resources. **GT** and WordNet both lack complete cover - over half the relations were confirmed with only one resource. This shows that the machine readable dictionary is a useful source of these relations.

6 Conclusion

In this paper, we presented the results of experiments conducted in automatic ontological acquisition over two languages, English and Japanese, and from three different machine-readable dictionaries. Our system is unique in combining parsers of various levels of analysis to generate its input semantic structures. The system is language agnostic and we give results for both Japanese and English MRDs. Finally, we presented evaluation of the ontologies constructed by comparing them with existing hand-crafted English and Japanese ontologies.

References

Eneko Agirre, Olatz Ansa, Xabier Arregi, Xabier Artola, Arantza Diaz de Ilarraza, Mikel Lersundi, David Martinez, Kepa Sarasola, and Ruben Urizar. 2000. Extraction of semantic relations from a Basque monolingual dictionary using Constraint Grammar. In *EURALEX 2000*.

Shigeaki Amano and Tadahisa Kondo. 1999. *Nihongo-no Goi-Tokusei (Lexical properties of Japanese)*. Sanseido.

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

Ulrich Callmeier. 2000. PET - a platform for experimentation with efficient HPSG processing techniques. *Natural Language Engineering*, 6(1):99–108.

Ulrich Callmeier, Andreas Eisele, Ulrich Schäfer, and Melanie Siegel. 2004. The DeepThought core architecture framework. In *Proceedings of LREC-2004*, volume IV. Lisbon.

Ann Copestake. 2002. *Implementing Typed Feature Structure Grammars*. CSLI Publications.

Ann Copestake, Dan Flickinger, Carl Pollard, and Ivan A. Sag. 2005. Minimal Recursion Semantics. An introduction. *Research on Language and Computation*, 3(4):281–332.

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

Dan Flickinger. 2000. On building a more efficient grammar by exploiting types. *Natural Language Engineering*, 6(1):15–28. (Special Issue on Efficient Processing with HPSG).

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

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. (`http://xxx.lanl.gov/abs/cmp-lg/9510008`).

Kaname Kasahara, Hiroshi Sato, Francis Bond, Takaaki Tanaka, Sanae Fujita, Tomoko Kanasugi, and Shigeaki Amano. 2004. Construction of a Japanese semantic lexicon: Lexeed. SIG NLC-159, IPSJ, Tokyo. (in Japanese).

Taku Kudo and Yuji Matsumoto. 2002. Japanese dependency analysis using cascaded chunking. In *CoNLL 2002: Proceedings of the 6th Conference on Natural Language Learning 2002 (COLING 2002 Post-Conference Workshops)*, pages 63–69. Taipei.

Yuji Matsumoto, Kitauchi, Yamashita, Hirano, Matsuda, and Asahara. 2000. *Nihongo Keitaiso Kaiseki System: Chasen*. `http://chasen.naist.jp/hiki/ChaSen/`.

Eric Nichols, Francis Bond, and Daniel Flickinger. 2005. Robust ontology acquisition from machine-readable dictionaries. In *Proceedings of the International Joint Conference on Artificial Intelligence IJCAI-2005*, pages 1111–1116. Edinburgh.

Minoru Nishio, Etsutaro Iwabuchi, and Shizuo Mizutani. 1994. *Iwanami Kokugo Jiten Dai Go Han [Iwanami Japanese Dictionary Edition 5]*. Iwanami Shoten, Tokyo. (in Japanese).

Patrick Pantel, Deepak Ravichandran, and Eduard Hovy. 2004. Towards terascale knowledge acquisition. In *20th International Conference on Computational Linguistics: COLING-2004*, pages 771–777. Geneva.

Stephen D. Richardson, William B. Dolan, and Lucy Vanderwende. 1998. MindNet: acquiring and structuring semantic information from text. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics: COLING/ACL-98*, pages 1098–1102. Montreal.

Kiyoaki Shirai. 2003. SENSEVAL-2 Japanese dictionary task. *Journal of Natural Language Processing*, 10(3):3–24. (in Japanese).

Melanie Siegel. 2000. HPSG analysis of Japanese. In Wahlster (2000), pages 265 – 280.

Kathrin Spreyer and Anette Frank. 2005. The TIGER RMRS 700 bank: RMRS construction from dependencies. In *Proceedings of the 6th International Workshop on Linguistically Interpreted Corpora (LINC 2005)*, pages 1–10. Jeju Island, Korea.

Takenobu Tokunaga, Yasuhiro Syotu, Hozumi Tanaka, and Kiyoaki Shirai. 2001. 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.

Wolfgang Wahlster, editor. 2000. *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer, Berlin, Germany.

Yorick A. Wilkes, Brian M. Slator, and Louise M. Guthrie. 1996. *Electric Words*. MIT Press.