Extracting Representative Arguments from Dictionaries for Resolving Zero Pronouns

Shigeko Nariyama*♥, Eric Nichols★, Francis Bond*,
Takaaki Tanaka*, Hiromi Nakaiwa*
Communication Science Lab, NTT Kyoto, Japan
*{shigekon, bond, takaaki, hiromi}@cslab.kecl.ntt.co.jp
♥ The University of Melbourne, Australia shigeko@unimelb.edu.au
★Nara Institute of Science and Technology, Nara, Japan eric-n@is.naist.jp

Abstract

We propose a method to alleviate the problem of referential granularity for Japanese zero pronoun resolution. We use dictionary definition sentences to extract 'representative' arguments of predicative definition words; e.g. 'arrest' is likely to take *police* as the subject and *criminal* as its object. These representative arguments are far more informative than 'person' that is provided by other valency dictionaries. They are auto-extracted using both Shallow parsing and Deep parsing for greater quality and quantity. Initial results are highly promising, obtaining more specific information about selectional preferences. An architecture of zero pronoun resolution using these representative arguments is described.

1 Introduction

One of the biggest obstacles for the success of machine translation (MT) is found when the target language requires linguistic information that is implicit in the source language. One well-known example for this is observed in Japanese-to-English translation where Japanese contains an abundance of zero pronouns (ellipses, such as unexpressed subjects and objects); the problem known as 'ellipsis resolution'. The referents of these ellipses must be retrieved and expressed overtly in the English translation to be grammatical. For example, the Japanese sentence given in (1) contains three ellipses indicated by ø (subscripts indicate coreference), which need to be verbalised in the English translation, except perhaps for the first ellipsis.

(1) *Soosain$_j$ ya sinzoku$_k$ ga koishitsu ni hairi,* **ø$_{j+k}$** *Nakagawa yogisha$_i$ o settokushi,* **ø$_j$** **ø$_i$** *taiho shita.*

捜査員$_j$や親族$_k$が 更衣室に入り、
ø$_{j+k}$ ガ 中川容疑者$_i$ を 説得し、**ø$_j$** ガ **ø$_i$** ヲ 逮捕した。

'Detectives$_j$ and relatives$_k$ entered the locker room, **ø$_{j+k}$** persuaded Nakagawa$_i$, the suspect, and **ø$_j$** arrested **ø$_i$**.'

Methods for ellipsis resolution have been proposed using 'verbal semantics', 'grammar based rules', 'stochastic', 'machine learning', and various hybrids of these approaches. Of these, 'verbal semantics' is reported to be the most effective (Nakaiwa and Seki 1999, Isozaki and Hirano 2003). However, the verbal semantics currently available in machine-readable valency dictionaries, such as *Goi-Taikei* - a Japanese lexicon (Ikehara et al. 1997, see §2.2), are often too general and are, thus, insufficient for the purpose of resolving ellipsis (Kawahara and Kurohashi 2004, Iida et al. 2004, inter alia).

We propose a method to alleviate this problem of referential granularity. It uses Japanese dictionary definition sentences to extract referential information that are the representative arguments of the predicative words being defined (i.e. 'definition' words). For example, the Lexeed dictionary (Bond et al. 2004; see §2) provides the following definition about the word *taiho* 逮捕 'arrest', whereby we extract the referential information *police officer* and *criminal*.

(2) *Taiho: keisatsu ga hannin o toraeru koto.*
逮捕: 警察が、犯人 を 捕える こと 。
'Arrest: A police officer captures a criminal.'

These extracted referents are 'representative arguments', prototypical examples of the real-world referents that are likely to fill the argument slots. It is a fact about the real-world that things like *police* are likely to be the subject of the verb *arrest* and things like *criminal* are likely to be its object. These representative arguments can be used as the basis for selectional preferences, which allow room for rhetorical and other deviated usages.

In general, we should prefer an interpretation where the referents of the arguments are semantically similar to the representative arguments. Because arguments only have to be similar, not subsumed by, it is possible for the representative arguments to be actual words, although word senses would be preferred.

In contrast, processing using selectional restrictions must use broader semantic classes, otherwise non-typical sentences would be rejected. For example, *Goi-Taikei*'s valency dictionary[1] has the semantic classes *agent* and *person* as selectional restrictions for *taiho* 'arrest'. These subsume the words *police* and *criminal* but are much less informative.

The goal of our research is to extract more specific referents than what *Goi-Taikei* provides. This will expand *Goi-Taikei*'s contents (rather than replace it) by adding complementary referential information that is more useful for ellipsis resolution. Moreover, it is intended that the inventory of representative arguments we extract also provides world knowledge in many cases, such that we can draw an inference from a sentence, for example 'John arrested the criminal', that John is a police officer or has a related occupation.

The rest of the paper is organised as follows. We elaborate on the significance of representative arguments, the process of extracting them from a dictionary, and the architecture for ellipsis resolution in §2. §3 describes our scheme for the extraction (combining Shallow and Deep Parsing), followed by Evaluation (§4), Discussions and remaining work (§5), and Related research (§6).

2 Extracting representative arguments from dictionaries

We see several advantages in using dictionary definition sentences for collecting referential knowledge. Dictionaries are created to provide information about words from cross-domain in lay terms with little contextual information to be comprehensible, while often providing world knowledge as well. The abundance of Japanese ellipses is also reflected in the dictionary definition sentences, thus most arguments are not expressed. Nonetheless, when the arguments are expressed, they tend to be very specific and useful for ellipsis resolution, as seen in (2) 'arrest'.

We use the Japanese semantic database Lexeed (Bond et al. 2004). This is a hand-built self-contained lexicon, consisting of definition words and their definitions for the most familiar 28,270 words, as measured by native speakers, comprising a total of 46,347 different senses. This set is large enough to include most basic level words and covers over 72% of the common words in a typical Japanese newspaper.

Lexeed has been enhanced by manual sense disambiguation of all the open class words. Further, the senses are linked in an ontology (Nichols et al. 2005), which allows us to measure the semantic distance between words or senses using a variety of methods.

2.1 Resolving ellipsis in MT using representative arguments

The representative arguments we extract enhance the performance of ellipsis resolution in many ways. They:

- Give *more detailed descriptions* of the semantic attributes of referents than currently available
- Hence, *narrow down candidates* for the referent of ellipsis
- Handle the *split of antecedents*: e.g. (3)

The subject of the last predicate 'arrest' is 'j', not the same as the subject 'j+k' in the preceding clauses. This reading can be achieved only if the representative arguments of the predicate 'arrest' are used in ellipsis resolution.

(3) *Soosain*$_j$ *ya sinzoku*$_k$ *ga koishitsu ni hairi,* $\varnothing_{j+k}$ *Nakagawa yogisha*$_i$ *o settokushi,* $\varnothing_j$ $\varnothing_i$ *taiho shita.*
捜査員$_j$や親族$_k$が 更衣室に入り、$\varnothing_{j+k}$〃 中川容疑者$_i$ を説得し、$\varnothing_j$〃 $\varnothing_i$〃 逮捕した。
'Detectives$_j$ and relatives$_k$ entered the locker room, $\varnothing_{j+k}$ persuaded Nakagawa$_i$, the suspect, and $\varnothing_j$ arrested $\varnothing_i$.'

- Handle *word sense disambiguation*

*Gohan*ご飯 has two senses: rice and meal. However, as shown in (4), the predicate *taku* 炊く 'cook' is defined as 'to cook rice or vegetable, thereby the referential information resolves the ambiguity. In contrast, a MT system translates (4) incorrectly translating *gohan* as 'meal'.

(4) *Asa* $\varnothing_i$ *taite, hiru* $\varnothing_i$ *tabenai kara, gohan*$_i$ *ga nokoru.*
朝 $\varnothing_i$ 炊いて、昼 $\varnothing_i$ 食べないからご飯$_i$が残る。
'(I) cook (rice$_i$) in the morning, but (I) don't eat (it$_i$) for lunch, so the rice$_i$ is left over.'

MT: 'It cooks in the morning and, because it does not eat at noon, a meal remains.'

- Handle *a mismatch of generic referential ellipsis*

There is an asymmetry of implicit generic pronouns between languages. For example, in (5) the Japanese verb *kaeru*帰る'return' implies returning 'home', which cannot be inferred in the corresponding English verb 'return'. On the other hand, in English (6) implies 'dinner' and (7) 'letter', without which the correct translations of the

[1] *Goi-Taikei* was published in 1997 primarily for parsing and disambiguating different senses of predicates in MT, and the semantic granularity of arguments was deemed sufficient for their initial purposes.

Japanese sentences cannot be derived. These representative arguments can be extracted from dictionary definition sentences.

(5) *Taro ga sore o katte kaetta.* 太郎がそれを買って帰った.
' Taro bought it and returned.'

(6) 'He usually eats at 7pm.'
 # Karewa itsumo 7ji ni taberu. 彼はいつも7pm に食べる

(7) 'I will write to you.'
 # Watashi wa kakimasu. 私は書きます.

- Provide a *default argument* in case of no candidate is available in the context, such as the case for generic referential ellipsis.

2.2 Architecture for ellipsis resolution

The architecture for making referential information useful for ellipsis resolution comprises fours steps, as shown below, followed by an explanation of each step:

{Preparing for ellipsis resolution}
[1] **Extract representative arguments from the definition sentences**,
[2] **Draw referential categories and their values** to cluster the definition words by the types of those extracted arguments using *Goi-Taikei*,
[3] **Assign candidate nouns** for the ellipsis selected from the context **with the referential categories and their values**, and

{Resolving ellipsis}
[4] **Give a high preference** as the referent of ellipsis **to the candidate noun** that has **the same referential category and the same value** as that of the definition word (i.e. the predicate of the given sentence where ellipsis is contained).

[1] Extraction of representative arguments

We examine the definition words in Lexeed that are predicates, i.e. verbs (V), verbal nouns (VN), and adjectives. We automatically extract from the definition sentences representative arguments of these predicates that are Nominative (marked by *ga*), Dative/Locative (*ni*), and for V and VN also Accusative (*o*)[2]. Then we manually verify the extracted arguments in order to make a reliable inventory (see §3.2).

[2] We tested a sample of nouns with all other case particles. The number of useful referents extracted from those cases was deemed insignificant and precluded from the initial test.

[2] Drawing referential categories and their values using *Goi-Taikei* thesaurus

When resolving ellipsis, the candidate referent of ellipsis should carry the same semantic information (referential category and it value) as the one that is preferred by the predicate (i.e. the definition word). In the case of *arrest*, the candidate noun must contain occupational information (referential category) to be matched with *police* (value). Hence, we create referential categories and values to cluster the definition words by the referential types of the extracted arguments. A sample of the categories and values that have been identified as useful for ellipsis resolution is shown below:

Referential categories

Gender: [F] 妊娠 *ninshin* 'pregnant'
[F] 男勝り *otokomasari* 'mannish
[M] 男らしい *otokorashii* 'manlike'
[MtoF/FtoM] 求婚 *kyukon* 'propose a marriage'
Occupation: [doctor, patient] 往診 *ooshin* 'examine'
Generation [infant, boy/girl, youth, adult, elderly]:
[adult] 大人げない *otonagenai* 'unlike grown-up'
[boys/girls] いたずらっ子 *itazurakko* 'a little imp'
Social hierarchy [company, school, family, etc]:
[employer, employee] 解雇 *kaiko* 'dismissal'
Person, Number, In/out group, Domain, Others

These referential categories are drawn from the semantic classes in *Goi-Taikei*, such as Agent and Person. *Goi-Taikei* has a semantic feature tree with over 3,000 nodes of semantic classes organised with a maximum of 12 levels (see Figure 1). It includes in its semantic classes information on gender, occupation, generation; exactly the sort of information needed for this assignment.

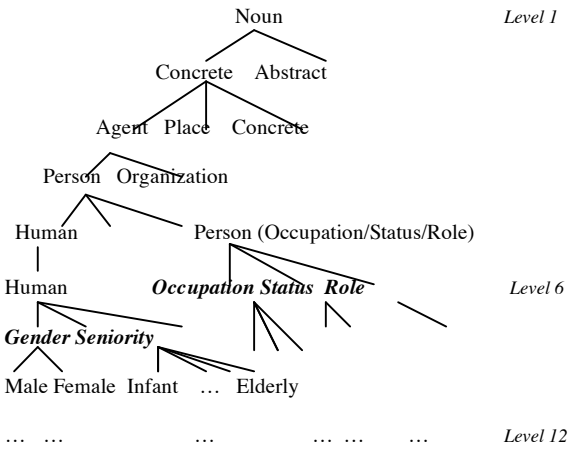

Figure 1: Excerpt from Goi-Taikei thesaurus

Two tasks are required for gathering referential categories from the semantic classes in *Goi-Taikei*.

1) ***Group*** *some semantic classes together from different nodes.*

For example, gender information is found in different classes, not only under Gender, but also under Women, Mothers; in addition, Employees, Proper names, Titles and others have separate classes for gender subclasses beneath them, but these classes are spread over different nodes in the tree. The same goes for Social hierarchy; Occupation (status), Organization, Family, etc, are scattered in the tree. These classes are grouped together under a name to help resolving ellipsis.

2) ***Rank*** *some semantic classes*

The use of honorifics and some predicates require information on seniority ranking; e.g., 'senior' is higher than 'youth', and 'employer' is higher than 'employee'. To be more concrete, *insotusuru* 引率する'A takes B out somewhere' requires in Japanese that A person is senior than B person in terms of age or social status.

Japanese has no syntactic coding of the Subject-Verb agreement in terms of Person, Number and Gender seen in many European languages, which causes a problem in MT to translate sentences to languages that have such agreement. However, these referential categories function similar to the agreement. Only they extend to more categories in the form of Argument (objects as well as subjects, i.e. extracted referents) - Predicate (i.e. definition words) selectional preferences. They can be lexico-semantically drawn to fill the gap; e.g. 'butler' in English entails 'a male person', so it has Gender category with Male value.

[3] Assigning referential categories to candidate nouns

Referential categories can be assigned to candidate nouns only when they hold that information, since not every definition word has 'representative' arguments or dictionaries list those arguments. But candidate nouns can be assigned for more than one referential category. For example, the subject of 妊娠 *ninshin* 'pregnant' is assigned at least for [Female] and [Adult].

[4] Resolving ellipsis

This paper deals mainly with the first step of referent extraction. After completing the three steps, we are able to add the referential information to our algorithms for ellipsis resolution (Nakaiwa 1999, Nariyama 2003) that selects the referent most semantically similar to the representative argument (see examples in §2.1).

3 Scheme for extraction

3.1 Which arguments to extract

Although dictionary definition sentences contain many arguments in them, not all of them are useful in ellipsis resolution. This subsection describes the criteria for 'useful arguments' and the complexity of that extraction due to the nature of dictionaries containing different functions described in definition sentences.

Dictionary definition sentences can be divided into four types according to what they describe:

1) explaining the meaning of the definition word,

2) paraphrasing the definition word using a predicate with the same argument structure of the definition word,

3) paraphrasing the definition word using a predicate with a different argument structure of the definition word, and

4) providing world knowledge about the word

Identifying precisely to which function a definition sentence pertains is currently beyond the scope of a computer program[3], although we have incorporated some constraints into the algorithm capturing the differences (see § 3.2).

The referents useful for ellipsis resolution can be found mostly in sentence types 2) and 3), whereby the extracted arguments can comprise a grammatical sentence with the definition word; namely, the arguments can be used as the arguments of the definition word.

The definition word *taiho* 逮捕 'arrest' seen in (2) is such an example. The arguments *police* and *criminal* in the definition sentence can be the exact arguments having the same case and semantic roles as the definition word has, i.e. '*police* arrest *criminal*' forms a grammatical sentence. On the other hand, an example of unsuitable arguments is the definition word *aisuru* 愛する 'to love', which has the argument structure 'A-subj B-obj Verb', while its definition sentence is *jo o motsu* 情を持つ 'to have love' with the argument structure 'A-subj for B C-obj Verb'.' If the argument in the definition sentence is adopted to be the argument of the definition word, we get *jo-o ai-suru* 情を愛する, 'to love love', which is infelicitous and thus discarded.

[3] Many definition sentences have more than one type of description within one complex sentence. In addition, the distinctions among the different types are not always clear.

3.2 How to extract

In essence, the following procedure was taken to find representative arguments.

1. [Auto] Extract candidate arguments using both Shallow parsing and Deep parsing for optimising the extraction of referential information (see §3.3 for details).
2. [manual] Select the argument if the definition word can take it as its argument (e.g. *taiho* 'arrest'), and go to 3.
   If not (e.g. *aisuru* 愛する 'love'), discard it.
3. [manual] Compare the argument structure of the definition sentence with that of definition word.
   - o if identical, then take the argument
   - o if involving different cases, then take it but with the correct case particle
4. [Auto] Select the argument if it is more specific than that in *Goi-Taikei*.

Given the complexities of dictionary definition sentences, hand verification is mandatory to make a reliable list of representative arguments. In order to minimize the inclusion of irrelevant arguments and maximize the extraction of useful arguments in auto-extraction for ellipsis resolution, we developed an algorithm with the following constraints:

- Exclude arguments that have functional predicates, which tend to occur when definition sentences explain definition words (type 1)): *aru, suru, naru, you, yousu* (meaning 'be', 'do', etc)
- Exclude arguments that are general: *mono, koto, hito, joutai, tokoro, teido, ten, kanji, tame* (meaning 'thing', 'person', 'state', 'place', etc)
- Extract the preceding noun when *nado* 'etc', *nomi/dake* 'only', *to* 'and' are to be extracted.

3.3 Combined use of Shallow Parsing and Deep Parsing

Shallow parsing (SP) allows us to extract information from more data, but with less precision. Deep parsing (DP) gives us more accurate information, but only for those sentences that can be parsed. Combining the results, as suggested by, for example the Deepthought project (Frank 2004), gives us the advantages of both. Hence, we combine DP and SP to extract a greater number of representative arguments while maintaining a high level of accuracy.

3.3.1 Shallow Parsing

We used the Japanese morphological analyser, ChaSen (Matsumoto et al. 2002), as the base for our SP. We tagged the words in the definition sentences, and identified the predicates and arguments to be extracted using the following heuristic:

- Predicate are identified as having POS verb (V) or verbal noun (VN) (excluding auxiliary verbs)[4]
- Arguments are identified as nouns preceded by Nominative (*ga*) or Dative/Locative (*ni*) case, and Accusative (*o*) as well for the definition words that are V and VN
- Arguments are assumed to attach to nearest following predicate[5]
- First nominative arguments plus final predicate and its arguments are extracted for each sentence.

3.3.2 Deep Parsing

We used a combination of the PET parsing system (Callmeier 2002) and the JaCY Japanese HPSG grammar (Siegel and Bender 2002). PET is an open source[6], highly efficient unification parser. JaCY is broad-coverage, freely available HPSG grammar that produces semantic anaysis in Robust Minimal Recursion Semantics (RMRS, Frank 2004). JaCY has a lexicon containing over 36,000 entries, allowing it to cover the entire Lexeed corpus. It currently produces RMRS structures for 61,462 of the Lexeed definition sentences for a coverage of 81.9%.

RMRS is an algebra for specifying predicate relations. It uses a number of semantically bleached 'ARG' slots to store the arguments of a relation. The essence of the RMRS structure for (2) is given in Figure 2.[7] In the example RMRS, ARG1 and ARG2 of **_toraeru_v_1** *'to capture'* point to **_keisatsu_n_1** '*police officer*' and **_hannin_n_1** '*criminal*' respectively.

---

[4] Adjectives were initially classified as predicates. However, arguments would attach to adjectives as well, creating a great deal of noise. Thus, adjectives are extracted as predicates only in DP.

[5] It has been suggested that using a dependency analyser could be useful in more accurately determining the arguments of a predicate, an area for future work.

[6] PET can be downloaded at: http://wiki.delph-in.net/moin/PetTop

[7] In Figure 2, **real relations** (events and objects) are given in **bold font**, and *grammatical relations* are given in *italics*. Relation names are formed by joining the word, its POS, and its sense together with underscores. The **HOOK** for sentence (2) points to the **_koto_n** relation identifying it as the relation with the highest scope.

proposition_m_rel(h1,h3)
qeq(h3,h23)
_keisatsu_n_1(h4,x5)
_hannin_n_1(h9,x10)
_toraeru_v_1(h14,e15:present:)
ARG1(h14,x5)
ARG2(h14,x10)
_koto_n(h16,x17)
ARG1(h16,h18)
proposition_m_rel(h18,h22)
qeq(h22,h14)
unknown_rel(h23,e2:present:)
ARG2(h23,x17)

Figure 2: RMRS structure for (2)

「逮捕: 警察が、犯人 を 捕える こと 。」
'Arrest: A police officer captures a criminal.'

Our algorithm for extraction using DP is essentially the same as for SP; extract the semantically most relevant predicate ('main/final predicate') and its arguments. However, the criteria for determining predicates and arguments are quite different. We produce RMRS structures for the definition sentences by parsing them with PET and JaCY, and use the semantic head extraction algorithm given in Nichols et al. (2005) to determine the main predicate for each sentence and extract its arguments directly from the RMRS structure. In short, our DP algorithm is as follows:[8]

1. Extract the main predicate that are a V, VN, or adjective using semantic head extraction algorithm
2. Extract ARG1, ARG2, ARG3 and Dative/ Locative marked (_ni_p,に) from main predicate
3. Filter out arguments that are not N or VN.

4 Evaluation

By following the method described in §3, we obtained the following results, shown in Tables 1 and 2. The total number of extracted arguments is 10,076. Of these 6,550 (65.0%) are representative arguments that are more specific than those in *Goi-Taikei* or new to *Goi-Taikei*.

Table 1 gives the precision (the rate of representative arguments extracted over total extraction) per POS and parsing method, and Table 2 the proportion of each case among the representative arguments extracted per POS.

Table 1 shows promising results, except for Adjectives, perhaps because the definition sentences for adjectives tend to *explain* more than *paraphrase*. Table 2 shows that the representative arguments are found mostly in Accusative, except for Adjective in Nominative.

	Adjective	Verb	Verbal N	All
DP only	69.3%	76.6%	72.8%	74.1%
SP only	49.9%	63.7%	49.9%	55.3%
Extracted by Both	56.8%	72.4%	72.9%	70.0%
Total (number)	**57.8%** (841/15455)	**71.5%** (3041/4252)	**66.0%** (2883/4370)	**67.4%** 6765/10076)

Table 1: Precision per POS and parsing method

	Adjective	Verb	Verbal noun
Nom. *ga*	86.4%	26.0%	22.5%
Acc. *o*	N/A	47.0%	57.7%
Dative *ni*	11.7%	27.0%	20.0%

Table 2: The proportion of each case among representative arguments per POS

Filtering by *Goi-Taikei*

Finally, we compare the specificity of the extracted arguments with that of the corresponding words per sense in *Goi-Taikei* with the following classification. The results are shown in Table 3.

① > GT: more specific than GT
② = GT: same specificity as GT
③ no entry of the definition word in GT
④ no sense entry of the definition word in GT
⑤ < GT: less specific than GT

	Adj.	Verb	VN	All
> GT ①	**48.8%**	**57.4%**	**46.6%**	***51.7%***
= GT ②	.8%	3.4%	3.1%	2.9%
no GT entry ③	41.1%	22.7%	39.8%	32.3%
No sense GT ④	9.3%	16.2%	10.2%	12.8%
< GT ⑤	0%	.3%	.3%	.3%
∑	100% (841)	100% (3,041)	100% (2,883)	100% (6,765)
N(①+③+④) / ∑ extracted	98.9%	98.2%	98.5%	*96.8% (6,550)*

Table 3: Comparing specificity of extracted arguments with that in Goi-Taikei (GT)

The results show that 51.7% of the arguments we selected provide more specific referential information than those in GT. If those arguments

[8] Using our algorithm for the example RMRS structure, step 1 identifies **_toraeru_v_1** as the main and only predicate because **_koto_n** is treated as a semantically empty predicate. In step 2 **_keisatsu_n_1** and **_hannin_n_1** are extracted as arguments. Steps 3 and 4 do not filter out any of the results because the predicate and arguments have POS types consistent with the above restrictions.

This approach is similar to that of Hoelter (1999), who used an HPSG parse of definitions from COBUILD to extract sortal restrictions on arguments.

that are not listed in GT are to be included, i.e. ③+④, it goes up to 96.8%. In other words, virtually every argument extracted from the proposed method provides new or more specific referential information than what exists in GT.

5 Discussions and remaining work

We draw the following from the results in §4.

- The precision can still be improved by dealing with the areas of improvement noted below.
- Nonetheless, those representative arguments that were extracted provide new or more specific referential knowledge for most predicates than what exists in *Goi-Taikei,* which is currently the most informative resource in Japanese.
- From the view point of ellipsis resolution, the information on the nominative is particularly welcome, as they are by far most frequently omitted (e.g. Nakaiwa 1999, Nariyama 2003).

Areas of improvement

- Most definition words have multiple senses, some of which are a nominal, i.e. not a predicate, hence do not take arguments. Nonetheless arguments are extracted because Lexeed does not note the nominal use and does not excluded them. This caused a decrease in the precision.
- Some extracted arguments are representative but take a case particle other than *ga, wo, ni* that we set to extract from. These are not selected under this experiment, which lowered the precision.
- Coordinate structures are not handled well by DP and not at all by SP; e.g. in 'A, B, and C', often only C is extracted.
- *A no B* 'B of A' problem (e.g. 病気の質 *byoki no shitsu* 'type of illness'): both DP and SP predominantly picked only B when A is the true argument and B is an attribute.
- The inventory must expand by checking the entire dictionary instead of the most familiar words that covers 72% and further by combining the referential information from other dictionaries.

Extended usage of referential knowledge

- Apart from enhancing the performance of ellipsis resolution, this list of referential information is useful for word sense disambiguation, and will also be useful for resolving pronominal and zero anaphora in other languages.
- The method to extract representative arguments from dictionaries can be applied for other languages; e.g. Oxford Advanced Learner's dictionary 'If the *police* arrest someone, the person is taken to a *police station*'.
- Many representative arguments are cross-linguistically valid and readily transferable to other languages, as the meanings are based on world knowledge, e.g. 'arrest', while some are not, e.g. generic ellipsis 'return (home)'.
- Adding the information about representative arguments to the ontology extracted by Nichols et al. (2005) enriches it to become a more general ontology, comparable to MindNet (Richardson et al. 1998).

Remaining work

The work described in this paper covers mainly Step 1 in the architecture of ellipsis resolution outlined in §2.2. Once the referential categories and values are determined, we are able to proceed to the next step, that is, to assign the categories and values to candidate nouns, and use that information in line with our algorithms to resolve ellipses.

6 Related work

The use of dictionaries for acquiring ontology has been the method taken by many in various languages (Tsurumaru 1991, Wilks et al. 1996, Richardson et al. 1998, inter alia).

In terms of work that focuses on extracting referential information, many studies use newspaper corpora. The two notable work in Japanese are the new EDR *Verb valency dictionary* (Hagino et al. 2003, listing verbs only) and Case frame dictionary (Kawahara and Kurohashi 2004).

The corpora of the latter are particularly impressive covering 20 years of newspaper articles (21 million sentences). Nonetheless, as the Japanese language contains an abundance of ellipses, the nominative arguments in particular, lists based on corpora may still be influenced by this phenomenon to some degree.

Kawahara and Kurohashi draw a semantic feature for every sense of predicate from the arguments collected from the corpora to deal with word sense disambiguation. This approach is inductive, and is also seen in the work for English (e.g. Agirre and Martinez 2002). Our approach is the reverse, i.e. deductive, in that representative (prototypical) arguments for a predicate is extracted, which can then be expanded to include words that are similar to them.

Kawahara and Kurohashi's inductive method has a possibility of taking words that are not so representative, which may cause some deviation of meaning. For example with *shinsatsu* 診察 'medical examination', we get '*medical doctor* examines *patient*' from Lexeed. The arguments extracted

from corpora, on the other hand, must include titles and proper names. Even after discarding those non-central/indirect referents, they will need to disambiguate different senses of the title *sensei* that appears with *shinsatsu,* which can be a medical *doctor*, school/university *teacher*, or *politician*. Hence, we believe our method to be more efficient and accurate.

The particular advantage of our approach is that the referential categories drawn from representative arguments resolve what is particularly difficult in ellipsis resolution. That is, to resolve ellipses referring to the same semantic class, such as 'person'. For example with *kyukon* 求婚 'A propose (marriage) to B', both A and B are 'person'. Using Kawahara's Case frame dictionary drawn from corpora, because the arguments under A and those under B have both genders in them, when A and B are ellipses, it is quite possible to come out as 'John proposed to Bill.' Our method using representative arguments and referential categories, on the other hand, imposes the gender constraint that 'Male person proposes Female person' or vice versa (save for gay marriage).

Nevertheless, our approach also has one disadvantage, in that it cannot hold for all arguments, since not every definition word has 'representative' arguments or dictionaries list them. Merging the results from both methods is deemed beneficial.

7 Conclusions

Given the fact that even the state of the art of NLP has difficulty accounting for contextual information and world knowledge, rendering ellipses to overt forms seems a prohibitive task at present. Referential information in the form of representative argument gathered in this research is a step forward towards achieving that task.

We presented a method to automatically extract representative arguments from dictionary definition sentences by using an algorithm that combines deep and shallow parsing techniques in order to maximize the number of referents extracted without decreasing accuracy. Initial results showed promise and were made immediately usable in applications via human evaluation. Our next task is to improve the areas noted in §5 and to evaluate the result fully in ellipsis resolution.

References

E Agirre and D. Martinez. 2002. Integrating selectional preferences in WordNet. The first *International WordNet Conference*.

F. Bond et al. 2004. The Hinoki Treebank: A Treebank for Text Understanding, In Proc. of the First *IJCNLP*, *Lecture Notes in Computer Science*. Springer Verlag

U. Callmeier. 2002. Preprocessing and encoding techniques in PET. In Oepen et al. (eds), *Collaborative Language Engineering*, 127–143. CSLI Publications, Stanford

A. Frank. 2004. Constraint-based RMRS construction from shallow grammars. In Proc. of *COLING*. 1269–1272, Geneva

T. Hagino et al. 2003. *Japanese verb valency dictionary*. Sanseido Publishing

M. Hoelter. 1999. *Lexical-semantic information in Head-Driven Phrase Structure grammar and natural language processing*. Lincom Theoretical Linguistics

R. Iida et al. 2003. Incorporating contextual cues in tranable models for coreference resolution. In Proc. of the 10th *EACL Workshop on the computational treatment of anaphora*. 23-30

S. Ikehara et al. (eds). 1997. *Japanese Lexicon*. Iwanami Publishing

H. Isozaki and T. Hirano. 2003. Japanese zero pronoun resolution based on ranking rules and machine learning. In Proc. of *EMNLP*.184-191

D. Kawahara and S. Kurohashi. 2004. Improving Japanese zero pronoun resolution by global word sense disambiguation. In Proc. of *COLING*. 343-349. Geneva

Y. Matsumoto et al. 2000. *Nihongo Keitaiso Kaiseki* System: Chasen, version 2.2.1 manual

H. Nakaiwa. 1999. Automatic extraction of rules for anaphora resolution of Japanese zero pronouns in Japanese-to-English Machine Translation from aligned sentence pairs. *Machine translation*. 14(3-4):247-279

H. Nakaiwa and Seki S. 1999. Automatic addition of verbal semantic attributes to a Japanese-to-English valency transfer dictionary. In Proc. of the 8th International conference on *Theoretical and Methodological issues in MT*. 185-195

S. Nariyama. 2003. *Ellipsis and Reference-tracking in Japanese*. SLCS 66, Amsterdam: John Benjamins

E. Nichols, F. Bond, and D. Flickenger. 2005. Robust Ontology Acquisition from Machine-Readable Dictionaries. In Proc. of *the International Joint Conferences on Artificial Intelligence*. Edinburgh

S. Richardson et al. 1998. MindNet: acquiring and structuring semantic information from text. In Proc. of *COLING 1998*

M. Siegel and E M. Bender. 2002. Efficient deep processing of Japanese. In Proc. of the 3rd Workshop on Asian Language Resources and International Standardization at *COLING*, Taipei

H. Tsurumaru et al. 1991. An approach to thesaurus construction from Japanese language dictionary. In IPSJ SIG Notes Natural Language, vol.83-16. 121–128 (in Japanese).

Y.A. Wilkes et al. 1996. *Electric Words*. MIT Press.