Construction of an alternation-based English valency dictionary

Ben Hutchinson,* Francis Bond† and Timothy Baldwin‡
* **University of New South Wales** <ben@sultry.arts.usyd.edu.au>
† **NTT Communication Science Laboratories** <bond@cslab.kecl.ntt.co.jp>
‡ **Tokyo Institute of Technology** <tim@cs.titech.ac.jp>

Abstract

This paper describes the construction of an English valency dictionary which lists a wide range of alternations for each verb sense. Information is automatically extracted from the on-line version of GoiTaikei — a Japanese Lexicon, and an extended COMLEX, incorporating information from EVCA and WordNet, as well as additional features such as argument status, and an augmented case-role representation. Senses are distinguished on semantic grounds, depending on the core lexical meaning of the verb. Each sense may have one or more alternations, thus keeping the number of senses manageable, while allowing for systematic variation in the lexical realization.

1 Introduction

Our goal is to create an English dictionary useful for both analysis and generation in machine translation. We base our dictionary on the English half of the GoiTaikei valency dictionary (Ikehara *et al.* 1997).

This dictionary has detailed information about the structure of English verbs and their complements, but is somewhat rigid, normally only one possible pattern of complements is given. To improve the flexibility we incorporate ideas and information from four other lexical resources: COMLEX (Grishman *et al.* 1994), WordNet (Fellbaum 1998), EVCA (English Verb Classes and Alternations: Levin (1993)) and Jing and McKeown's (1998) combined lexicon, which incorporates information from COMLEX, WordNet and EVCA.

Our dictionary follows the basic architecture set out in Baldwin *et al.* (1999) in dividing each verb entry into one or more senses, each of which has one or more syntactic realizations or frames.

In this paper we first compare the existing GoiTaikei with COMLEX, WordNet, EVCA and the combined lexicon, using as an example the word *gather*. We then motivate our decision to divide verbs into senses and frames, and finally discuss the automatic construction of an enriched lexicon.

1.1 Linguistic resources

There are now several large-scale on-line resources for English, in this section we compare four of them, showing the strengths and weaknesses of each.

Goitaikei's Japanese/English valency dictionary

GoiTaikei's valency dictionary is made up of pairs of linked Japanese and English sentence patterns, as shown in Figure 1.[1] Each pair of patterns is considered to be a different sense. In principle, this means that there is a well motivated test for how many senses a word should have: a word has as many senses as it has different translations.

In practice there are two problems. The first is that the dictionary is uni-directional, so that even though two Japanese words may map to the same word in English, the English words are considered to be different. In a computational lexicon where each word has a great deal of information associated with it, this redundancy is undesirable and there is a real risk that relevant information may only be entered for one of the entries.

The second is that semantic constraints on the Japanese side are used both for Japanese analysis and transfer into English.[2] This means that they may have to be retuned whenever a new translation is added, making extension of the dictionary difficult and time-consuming. In addition many of the distinctions made cannot be motivated on mono-lingual grounds.

These problems aside, GoiTaikei contains many features that other dictionaries lack. The most obvious are the semantic constraints on verb complements, and the bilingual links. In addition GoiTaikei also contains many idiomatic constructions (such as *come and go* "be intermittent"), as well as marking of case-roles, domain, genre and many other features.

The COMLEX syntax dictionary

The COMLEX verb dictionary contains a rich source of syntactic information about the possible patterns

[1] Only two of the 19 patterns that include English *gather* are shown.

[2] These semantic constraints are used in word sense disambiguation. If two patterns have the same syntactic structure but different constraints on their arguments, the one whose constraints match those of the actual arguments best will be chosen. Over 2500 semantic categories allow fine differentiations between constraints.

```
Pattern ID: -201263-00-                U_SENT   (動作)
 ┌ N1 (主体) (が)                        ├ PREDICATE - VERB "gather"
 ├ N2 (主体 具体物 証拠) (を)            ├ CASE S  N1
 ├ N5 (場 場所) (に へ)                  ├ CASE DO N2 OBJ-form [det MAS]
 └ 集める                               └ CASE PP — U_PP "in" N5 OBJ-form

Pattern ID: -201257-00-                U_SENT   (状態 受身不可)
 ┌ N1 (主体 自然 生物 自然物) (が)       ├ PREDICATE - VERB "gather"
 ├ N3 (具体 場所) (に へ)                ├ CASE S  N1 [det MAS]
 ├ N5 (人間活動) (に)                    ├ CASE PP — U_PP "in/on/at" N3 OBJ-form
 └ 集まる                               └ CASE PP — U_PP "for" N5 OBJ-form
```

Figure 1: Sample GoiTaikei patterns for *gather*

```
(VERB :ORTH "gather" :SUBC ((INTRANS-RECIP)
                (PP :PVAL ("around" "inside" "with"))
                (S)
                (PART-NP :ADVAL ("up" "together"))
                (PART-PP :ADVAL ("together" :PVAL ("in"))
                (PART :ADVAL ("around" "together"))
                (NP-PP :PVAL ("into" "in"))
                (NP))
```

Figure 2: COMLEX entry for *gather*

```
1. gather, garner, collect, pull together -- (get together; "gather some stones";
   "pull your thoughts together")
2. meet, gather, assemble, forgather, foregather -- (collect in one place;
   "We assembled in the church basement"; "Let's gather in the dining room")
3. gather, congregate, collect -- (move together)
4. accumulate, cumulate, conglomerate, pile up, gather, amass --
   (collect or gather; "Journals are accumulating in my office")
7. assemble, gather, get together -- (get people together; "assemble your colleagues";
   "get together all those who are interested in the project";
   "gather the close family members")

5. gather -- (conclude from evidence; "I gather you have not done your homework")
8. understand, gather, infer -- (believe to be the case;
   "I understand you have no previous experience?")

6. gather, pucker, tuck -- (draw fabric together and sew it tightly)
```

Figure 3: WordNet senses of *gather*

verbs can appear in. This makes it very useful for syntactic analysis. Another strength of the dictionary is that it has been extensively checked against a corpus, and is annotated with many examples. The entry for *gather* (without its examples) is given in Figure 2.

Unfortunately, the syntactic frames are not grouped into senses: only *gather* "understand" can take a sentential complement, while only *gather* "collect" takes *around*, but this distinction is not made by COMLEX.

The WordNet on-line lexical database

WordNet is an online resource which lists a number of different senses for nouns, adjectives and verbs as well as numerous links between them. This makes WordNet useful for a variety of semantic tasks.

However we claim that many of these senses are unnecessary distinctions and lead to difficulties in sense disambiguation. For example WordNet senses 1., 2., 3., 4. and 7. or 5. and 8. of *gather* seem equivalent (shown in Figure 3, with senses grouped together by us). We discuss this further in the next section.

English Verb Classes and Alternations

Levin (1993) proposes the use of alternations as a useful tool in the study of a verb's meaning and its syntactic behavior. An alternation is a relation between pair of similar syntactic frames, involving a rearrangement or change in the number of arguments. A typical alternation is the **causative/inchoative** alternation: in verbs that undergo this alternation the subject of the intransitive verb is related to the object of the transitive. For example: *I gathered* <u>*the students*</u> ↔ <u>*The students*</u> *gathered*.

Levin divides verbs into classes on the basis of which syntactic alternations they can take, and proposes that these classes also reflect a common core in meaning. The classes are grouped into 49 families. Verbs are not explicitly separated into senses.

The word *gather* appears in three classes: the "Get" subclass of the "Verbs of Change of Possession" family, the "Shake" class of the "Verbs of Combining and Attaching" and the "Herd" class of the "Verbs of Existence" family. We claim that the fact that it appears in three different classes indicates that it has three different senses.

Jing and McKeown's (1998) combined lexicon

Jing and McKeown's (1998) dictionary incorporates syntactic frames from COMLEX and alternation pairs from EVCA into WordNet senses, along with frequency information for each sense. The combined dictionary has the strengths of all three resources, and has been successfully used in generation (Jing 1998). It has some rudimentary semantic constraints on arguments, but only at the level of `something` or `somebody`.

1.2 A definition of sense

In order to avoid spurious ambiguities, we keep the number of senses to a minimum, as argued for by Wierzbicka (1996:244).[3] Each sense has a core meaning, the "semantic invariant" but can be realized in different frames, which may differ in their thematic properties, aspect or even valency. Our architecture therefore stores information about the core meaning, such as semantic constraints, at the sense level.

This definition of sense allows us to make the following claims.

Claim 1 EVCA alternations do not alter the sense of a verb.

Claim 2 If two apparent senses have the same sets of alternations, then they are in fact a single sense.

Claim 3 If a phrase P1 in sentence S1 of an alternation has certain semantic constraints C_1 then the corresponding phrase P_2 in the other sentence S_2 of the alternation has the same semantic constraints C_2=C_1.

2 Generation of a structured lexicon

Generation of the new lexicon was in three stages: identification of alternations in GoiTaikei; clustering of GoiTaikei patterns into senses; mapping our senses onto Jing and McKeown's (1998) by comparison of frames and alternations.

In preparation, we copied the semantic constraints from the Japanese verbs to the English verbs. There were 2,284 different English verbs, with 13,052 patterns. There were 1,136 identical patterns, which we merged.

2.1 Identification of alternations

We estimated which alternations would most likely occur in the English half of GoiTaikei. We then searched for these alternations in the following way. First we classified the alternations according to (1) how many arguments were introduced or omitted, and (2) the subcategorizations of their patterns. This classification produced a number of small sets of alternations. One such set was {**causative/inchoative**, **unspecified object**}.

The semantic constraints were used to link arguments in separate patterns. For example the direct object of the first English pattern in Figure 1 was linked to the subject in the second pattern. This linking of arguments allowed us to choose the correct alternation from the appropriate set, in this case the **causative/inchoative** alternation. We found only 191 instances of alternations.

[3]Most human readable dictionaries, and WordNet, take the opposite approach, when in doubt, a new sense is created. This means that disambiguation is extremely difficult, even for humans. Is, for example, the meaning of *They gathered* "they moved together" or "they collected in one place"?

There are two reasons for this. The first is that the GoiTaikei lexicon has only been used for English generation and normally only offers one canonical form as a translation. This is one of the weaknesses of the lexicon.[4] The second is that Claim 3 is too strong for the dictionary as it now stands. Semantic constraints can differ for two reasons not related to the core meaning of the English verb: (1) the constraints are necessary for disambiguation during transfer, (2) there is variation between lexicographers. In order to identify all alternations it is necessary to match with much less restrictive conditions, and then check the results, either by hand, or against another lexicon.

In some cases the argument of an intransitive pattern had the same constraints as both arguments of a transitive pattern. In these cases it was not possible to tell if the alternation was **causative/inchoative** or **unspecified object**. These cases were flagged as being **intrans-trans** alternations to be later checked by hand. We hypothesize that a verb cannot alternate in both the **causative/inchoative** and **unspecified object** alternations, as it would lead to unresolvable ambiguity. When we have finished creating this lexicon, we can validate this empirically.

2.2 Clustering of patterns

If two patterns were in an alternation, then according to our Claim 2 they are of the same sense, and so were clustered together. Thus the transitive extension of the alternation relation (if we include the null alternation) produces a number of equivalence classes.

During the clustering, we also identified canonical frame-types, where the syntactic frame could be generated from the sense-level information by using a simple template that mapped semantic arguments onto syntactic ones. For example, a **transitive** frame maps the first argument to the subject, and the second to the object, while an **intransitive-ergative** frame maps the second semantic argument to the subject. Around 50% of the frames which appeared in the alternations could be reduced in this way, suggesting that the introduction of such templates can lead to a drastic reduction in the size of the dictionary.

2.3 Mapping senses onto WordNet's

We were able to map our senses onto WordNet's through a straightforward comparison of frames and alternations to those in Jing and McKeown's (1998) combined lexicon. Two senses were linked only if every GoiTaikei frame and alternation appeared in the combined lexicon list.

This means that the WordNet semantic information and COMLEX syntactic information combined by Jing and McKeown can be used together with GoiTaikei's semantic constraints on arguments.

[4] In fact some alternations, such as **passive**, are handled during the English generation in ALT-J/E.

3 Conclusions

In this paper we compared the strengths and weaknesses of four large scale computational English lexicons. We then described the first stages of the automatic construction of a dictionary that combines the strengths of all the resources, based on the architecture outlined in Baldwin *et al.* (1999).

The new lexicon offers both theoretical and practical advantages. All senses are motivated, different senses will only be created if they allow different syntactic realizations. Previous verb sense entries can act as templates for new sense entries leading to fewer errors in dictionary production. Using templates entry can be done on a sense, rather than frame, level, ensuring that all possible frames are considered.

Acknowledgments

We would like to thank Jing and McKeown for sharing their dictionary with us, Adam Meyers for sharing the COMLEX documentation source files and the members of the NTT Machine Translation Research Group for their comments and support.

References

Baldwin, Timothy, Francis Bond, and Ben Hutchinson. 1999. An alternation-based Japanese valency dictionary architecture. In *5th Annual Meeting of the Association for Natural Language Processing*. The Association for Natural Language Processing. (this volume).

Fellbaum, Christine (ed.) 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Grishman, Ralph, Catherine Macleod, and Adam Myers. 1994. COMLEX syntax: Building a computational lexicon. In *15th International Conference on Computational Linguistics: COLING-94*, volume I, 268–272, Kyoto, Japan.

Ikehara, Satoru, Masahiro Miyazaki, Satoshi Shirai, Akio Yokoo, Hiromi Nakaiwa, Kentaro Ogura, Yoshifumi Ooyama, and Yoshihiko Hayashi. 1997. *Goi-Taikei — A Japanese Lexicon*. Tokyo: Iwanami Shoten. 5 volumes.

Jing, Hongyan. 1998. Usage of WordNet in natural language generation. In *COLING-ACL '98 Workshop on Usage of WordNet in Natural Language Processing Systems*, 128–134, Montreal.

——, and Kathleen McKeown. 1998. Combining multiple, large-scale resources in a resuable lexicon for natural language generation. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics: COLING/ACL-98*, volume I, 607–613, Montreal, Canada.

Levin, Beth. 1993. *English Verb Classes and Alternations*. Chicago, London: University of Chicago Press.

Wierzbicka, Anna. 1996. *Semantics: Primes and Universals*. Oxford University Press.