An alternation-based Japanese valency dictionary architecture

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

Abstract

This research is aimed at developing a valency dictionary architecture to comprehensively list the full range of alternations associated with a given predicate sense, both efficiently and robustly. The architecture is designed to incorporate all relevant information included in the on-line version of GoiTaikei — a Japanese Lexicon, as well as additional features such as argument status, grammatical relations, and an augmented case-role representation. Alternations are represented as individual syntactic case frames indexed back to the basic semantic argument component of the given predicate sense. A proposal is given for the use of the dictionary in a machine translation system.

1 Introduction

The short-term aim of this research is to formulate a hierarchical dictionary structure to map the Japanese component of the GoiTaikei Japanese-English valency dictionaries (Ikehara *et al.* 1997) onto, which is able to describe structural consistencies in the Japanese data. The English component of the original dictionary set is similarly mapped into an analogous structure (Hutchinson *et al.* 1999), and a set of transfer links devised to indicate correspondences from Japanese to English. Clearly, the new structure must include all information described in the target dictionaries such that it is possible to reconstruct the original dictionary content from the three newly-generated structures, but the new dictionaries are equally exploited in extending the original informational content.

The advantages of maintaining the Japanese and English dictionaries separately based around alternations, and having an external link set, include efficiency, maintainability, robustness and scalability. These merits arise partly because of the sense-independence of the derived monolingual dictionaries (see below), and partly because of the possibilities for the clustering of lexical alternates of the same sense. Despite the obvious successes of the the original GoiTaikei valency dictionary architecture, the combination of Japanese and English correlates within a single entry has meant that unnecessarily fine-grained sense distinctions have had to be made in both languages. By considering the two languages separately, we are able to broaden our handling of mono-lingual predicate sense to a level more cognitively justifiable, reducing the number of dictionary entries. Also, by clustering lexical alternates, we are able to employ inheritance for the core pool of semantic and lexical data, improving maintainablity, alleviating redundancy of annotation, and enhancing scalability by way of reducing the informational requirement when annotating new alternates and predicate senses.

In the existing valency dictionary structure, the linking of inter-language sense within a single structure has led to the generation of extraneous senses. It is certainly true for closely related language pairs that overlap of sense for corresponding lexemes in the two languages can partially release us from consideration of word sense disambiguation. However, in the case of Japanese-English machine translation, we are not able to rely on the same effect. Rather, for a given source–target language translation pair, we are commonly faced with the situation of having only partial sense overlap for either a single sense or a restricted number of senses in the source language. Here, the exact degree of overlap must be described through selectional preferences (in the case of the GoiTaikei valency dictionary), and alternative translations found for any sub-usages of the source language lexeme not covered by the original translation.

An example of this general phenomenon can be seen for the Japanese verb *tenkai-suru* "to develop". Within GoiTaikei, *tenkai-suru* is associated with 4 distinct Japanese-to-English translation pairs, as depicted in Figure 1, with usage p_3 sense-subsumed by p_2 according to the selectional preferences on corresponding argument slots B and C. The reason for the partitioning off of a sub-usage of p_2 is that the "develop" translation of *tenkai-suru* is inappropriate for the semantic region described by p_3. As such, p_3 is an artificial sense of *tenkai-suru* used to increase accuracy in translation, and an unavoidable side-effect of having Japanese and English described within a single dictionary framework. By separating the descriptions of the two languages, we are able to remove such artificial senses, and relocate interlingual sense-based idiosyncrasies to the linking lexicon.

Closer observation of Figure 1 reveals that p_1 is the **causative/inchoative** alternate of p_2. In the original dictionary formulation, no explicit representation of this alternation relation between p_1 and p_2 is possible, and that the corresponding case slots (C and A, respectively) bear identical selectional restrictions re-

In Proceedings of the Fifth Annual Meeting of the Association of Natural Language Processing (Japan), Tokyo, Japan, pp. 193-6.

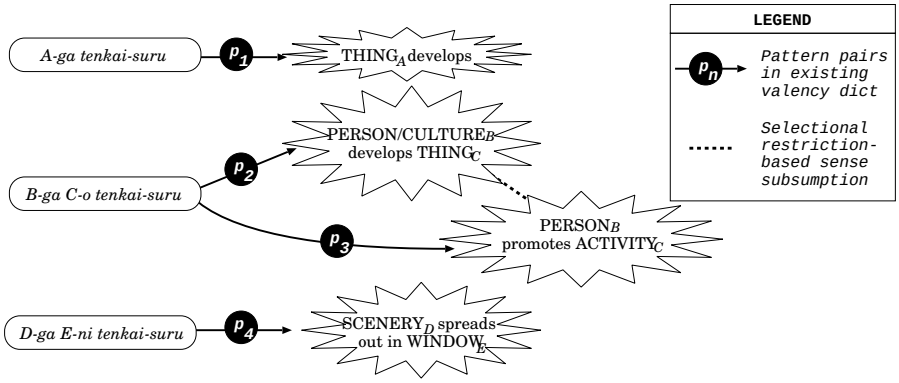

Figure 1: Japanese-English sense correspondence

flects more on the aptitude of the lexicographer than the inherent dictionary structure. Within our proposed architecture, however, p_1 and p_2 would be clustered together at the sense level and the alternation-based relation that exists between them explicitly expressed, producing co-indexing of the corresponding case slots. For this purpose, we clearly require a well-defined set of Japanese predicate alternations, in the manner of Levin's 80-fold set of alternation types for English (Levin 1993). The fleshing out of such a full set of Japanese alternations remains a longer-term aim of this research, with Fukui *et al.* (1985) providing a good start in this direction. For the time being, we have placed emphasis on the most readily occurring and well-documented alternations, namely the **object/argument**,[1] **causative/inchoative**, **passive** (*-rare*) and **causative** (*-sase*) alternations.

A longer-term advantage of maintaining the various dictionaries separately is that it becomes considerably easier to both reverse the translation direction and incorporate new languages into a single system architecture. In this, instead of effectively having $2C_2^n$ unidirectional transfer dictionaries between individual language pairs for n languages, we can have n monolingual lexicons connected with $2C_2^n$ linking lexicons, noting however that the overhead for constructing a linking lexicon is considerable less than that for constructing a transfer dictionary anew.

Given that we are justified in wanting to separate the Japanese and English components of the GoiTaikei valency dictionary and apply alternations in the sense clustering process, we are next faced with the issue as to what structure the new dictionaries should take. Here, we were inspired in the main part by the LISP-based hierarchical structure employed in the verb component of the COMLEX syntax (Macleod *et al.* 1998), and particularly in the forms of data encapsulation and cross-indexing utilised to describe constituent, control and grammatical structure across the subcategorisation frames for each verb alternation. Unfortunately, however, the COMLEX syntax was not immediately tenable with the descriptive needs of the GoiTaikei lexicon, largely because of the need to consider sense and selectional preferences, which are ignored in the COMLEX syntax. We thus adapted the basic structure to meet the needs of the task at hand.

The remainder of this paper is structured as follows. Section 2 describes the proposed dictionary architecture and the inter-relation between the various levels of representation. Section 3 presents preliminary results in converting the existing GoiTaikei dictionary into the given format, and Section 4 details a number of implementation issues related to the linking lexicon.

2 Japanese dictionary architecture

The proposed dictionary architecture is hierarchical, comprising, in descending order, of the word, sense and frame levels; these correspond to entries being clustered according to lexical stem, sense, and argument content/realisation, respectively.

Word level

At the highest level, entries sharing a common predicate stem are lexically clustered, as for conventional dictionaries. This enables us to give a single annotation of the basic stem orthography, part-of-speech (verb, adjective or adjectival noun) and conjugational class. Additionally, a regular expression representation of the predicate stem is given to counter the effects of *maze-gaki* (systematic variation in the Japanese orthography through the mixed use of kanji and kana).

Sense level

At the second level of description, entries are clustered according to sense, again in the manner of a conventional dictionary. Senses contain a sense ID, a list of sentences and/or indices to sentences in a corpus exemplifying the basic predicate sense, and a set of features including characteristic domains/genres of use of that sense. Most importantly, however, senses contain a description of the maximum argument content of that verb sense (`:sem`), by way of selectional preferences (`:res`) and/or a list lexical fillers (`:lexarg`). This represents the core meaning of the sense.

[1] This refers to the situation of the argument content of α comprising a proper subset of that of β, and being fully sense-subsumed by the intersecting argument content (see below).

```
(word :pos verb :orth "展開する"
  :features (:stem "(展|てん)(開|かい)"
             :conj "suru")
  :senses
  ((sense :senseid JP-tenkaisuru001
     :sem ((arg 1 :res (C0003 C1002))
           (arg 2 :res (C1000)))
     :features (:domain (01))
     :ex ("彼が新しい数学理論を展開した")
     :frames
     ((frame :index JP-tenkaisuru001-01
        :frame-type transitive
        :alt (:cause-inch (01 02))
        :features (:pid 101264 :vsa ((25/1)))
        :ex ("彼が新しい数学理論を展開した")
        :slots
        ((slot 1 :cs (np :cmark ("が"))
             :gs subject  :role agent
             :stat 3  :sem-arg 1)
        (slot 2 :cs (np :cmark ("を"))
             :gs dobj  :role changed
             :stat 3  :sem-arg 2)))
     (frame :index JP-tenkaisuru001-02
        :frame-type intransitive-erg
        :alt (:cause-inch (01 02))
        :features (:pid 101261 :vsa ((16/1)))
        :ex ("事件が面白い方向に展開した")
        :slots
        ((slot 1 :cs (np :cmark ("が"))
             :gs subject  :role agent
             :stat 3  :sem-arg 2)))))))
```

Figure 2: An example dictionary entry

The LISP list representation of argument content allows us to describe complex structures by way of nested structures, including optional or obligatory modifiability of arguments, and the manner of modification.

By including arguments at the sense level, we are taking the stance that, within the context of a single sense, a given argument has the same basic scope for lexical/semantic variance irrespective of its lexical realisation. That is not to say that the full range of arguments must appear in all usages of that sense, but simply that, given argument compatibility with a given alternation, that argument will be associated with a fixed set of selectional preferences and/or lexical fillers.[2]

As with the GoiTaikei lexicon, selectional preferences are indicated by way of a list of indices to nodes in the GoiTaikei thesaurus (Ikehara *et al.* 1997).

Frame level

The lowest level in the dictionary describes each individual case frame realisation. Frames are listed with an index, optional inflectional constraints, an optional description of the alternation types the current lexical realisation takes linked to the alternating frames, a list of example sentences characterising the alternation, and a list of features of the expression including its set of verbal semantic attributes (Nakaiwa and Ikehara 1997). What is undoubtedly the most integral component of alternation description, however, is a listing of individual case slots and associated features.

Case slots are presented in canonical ordering and annotated with: constituent structure (`:cs`), including case marker, an optional obligatoriness flag, and a phrase-level part-of-speech; grammatical relation (`:gs`); case-role (`:role`), based on the case grid representation proposed by Somers (1987) and suggested as being appropriate for inclusion in an expanded form of the GoiTaikei valency dictionary by Bond and Shirai (1997); argument status (`:stat`), based on a 7-level adaptation of the Somers (1987) valency binding hierarchy again proposed by Bond and Shirai (1997); and an index back to the sense-level list of argument constraints (`:sem-arg`).

3 Analysis

To get an idea of the level of alternation and sense subsumption in the GoiTaikei valency dictionary, we set out to automatically cluster GoiTaikei entries into the sense and frame levels. This clustering process has two facets, that of removing fully subsumed senses for a given predicate orthography, and combining and further reducing the final collapsed senses through alternation-based analysis.

In the first stage of processing, we assume that **full sense subsumption** is an indication of the artificial partitioning of sense to generate a more appropriate English translation, and discard all fully subsumed entries. α is fully sense subsumed by β sharing the same predicate stem and inflectional constraints, iff for each case slot α_i, the corresponding case slot β_i displays at least the same scope for surface case marker alternation, and the selectional restrictions on α_i are equivalent to or subsumed by the selectional restrictions on β_i. On removal of all fully sense-subsumed predicate entries, therefore, for every pair of residue predicate entries sharing the same predicate stem, there will be some usage particular to each.

The second stage of processing then involves comparing each pair $\langle \alpha, \beta \rangle$ of residue entries to determine if there is any alternation relation which exists between them such that the base form of alternating entry α is fully sense subsumed by β.[3] In the case that sense subsumption is detected, the entries is question are clustered together as being alternates of the same predicate sense, and appropriately marked for form of alternation.

[2] That pragmatic effects such as empathy can affect the relative acceptability of differing lexical contexts is not seen as a threat to this claim, but more evidence that pragmatics can override semantics in determining the felicity of an utterance. There is, however, facility to override the selectional preferences at the frame level.

[3] The implementation of this reverse alternation process is trivial for valence-maintaining alternations. In the case of valence-reducing and valence-expanding alternations, however, we devise a description of the prototypical semantics of affected case slots, and check that any added or removed case slots correspond in content with such prototypes.

Initially, the Japanese component of the basic GoiTaikei valency dictionary (verbs, adjectives and adjectival nouns) was transferred across to the proposed dictionary architecture, without consideration of sense- or frame-level clustering. This produced a total of 6484 word-level clusters, made up of 16488 senses (with 3593 word-level clusters containing a single sense) at an average of 2.54 senses per word cluster.

Basic analysis of fully sense-subsumed senses within the provisionally derived dictionary revealed 1229 instances of sense subsumption. The remaining 15259 senses were then further analysed for occurrences of the **object/argument**, **causative/inchoative**, **passive** and **causative** alternations, and all detected alternation pairs clustered together at the sense level. The number of instances of each alternation type are as follows:

Object/argument alternation: 1183
Causative/inchoative alternation: 171
Passive alternation: 11
Causative alternation: 24

In this way we were able to further reduce the number of senses by a factor of 1389 to 13870, a combined reduction of more than 15% over the original number of senses. This figure is particularly significant given that analysis was based on full sense subsumption, meaning that any slight variation in the case marking paradigm or instance of non-subsumption of selectional restrictions meant the senses in question were considered fully independent of another.

4 Use in MT: the linking lexicon

In order to use the mono-lingual alternation-based lexicon for machine translation, it must be linked to another such dictionary. To do this we use a linking lexicon. The basic idea is that lexical choice is left to the generation stage, but constrained by the input text. This allows flexible, fluent generation.

There are also several practical advantages. The lexicon is easy to update — for example a single sense entry may be adjusted rather than changing several pattern entries. All frames of a single verb or a single verb sense can be viewed at a glance, allowing errors and inconsistencies to be detected easily.

Ideally, verbs are linked at the sense level, information about which frame was used is passed along with the verb, but does not determine the frame used in the target language. However, the architecture allows links to be placed at any level.

The links should allow for additional syntactic and semantic constraints, for example the verb *warau* "smile/laugh" should be translated as *smile* if it is modified by the adverb *nikoniko*. There is no need to create an additional sense in the Japanese lexicon, it is sufficient to create an entry in the linking lexicon.

Finally, the linking lexicon should allow for pragmatic constraints on genre, domain or politeness.

5 Conclusion

In this paper we introduced an alternation-based valency dictionary structure for Japanese and discussed the relative merits of the proposed structure and separate linking lexicon over a transfer-style dictionary structure. We then went on to provisionally map the Japanese component of the GoiTaikei valency dictionary onto the new structure and analyse the descriptional efficiency of the produced dictionary over the original formulation.

Admittedly, this research is limited by the variety and scope of alternations applied in analysis, and thus further work is required to extend our set of Japanese alternation types. Once this set begins to grow in size, it should be possible to apply it in the analysis of the syntax/semantics interface, after (Levin 1993), and also lexical selection in generation (Dorr and Olsen 1996; Stede 1996). These are left as matters for future research.

References

Bond, F., and S. Shirai, 1997. *Practical and Efficient Organization of a Large Valency Dictionary.* Handout at the *Workshop on Multilingual Information Processing, held in conjunction with NLPRS'97.*

Dorr, B.J., and M.B. Olsen. 1996. Multingual generation: The role of telicity in lexical choice and syntactic realization. *Machine Translation* 11.37–74.

Fukui, N., S. Miyagawa, and C. Tenny, 1985. *Verb Classes in English and Japanese: A Case Study in the Interaction of Syntax, Morphology and Semantics.* Lexicon Working Papers #3, Center for Cognitive Science, MIT.

Hutchinson, B., F. Bond, and T. Baldwin. 1999. Construction of an alternation-based English valency dictionary. In *Proc. of the Fifth Annual Meeting of the Japanese Association for Natural Language Processing*, 197–200.

Ikehara, S., M. Miyazaki, A. Yokoo, S. Shirai, H. Nakaiwa, K. Ogura, Y. Ooyama, and Y. Hayashi. 1997. *Nihongo Goi Taikei – A Japanese Lexicon.* Iwanami Shoten. 5 volumes. (In Japanese).

Levin, B. 1993. *English Verb Classes and Alterations.* University of Chicago Press.

Macleod, C., R. Grishman, and Adam Meyers, 1998. *COMLEX Syntax Reference Manual.* Proteus Project, NYU, `ftp://cs.nyu.edu/pub/html/comlex.html/refman.ps`.

Nakaiwa, H., and S. Ikehara. 1997. A system of verbal semantic attributes in Japanese focused on syntactic correspondence between Japanese and English. *Journal of the Information Processing Society of Japan* 38.215–25. (In Japanese).

Somers, H. L. 1987. *Valency and Case in Computational Lingustics.* Edinburgh University Press.

Stede, M. 1996. Lexical paraphrases in multilingual sentence generation. *Machine Translation* 11.75–107.