Dictionary-driven analysis of Japanese verbal alternations

Timothy Baldwin*, Francis Bond‡ and Kentaro Ogura‡

* Tokyo Institute of Technology <tim@cl.cs.titech.ac.jp>
‡ NTT Communication Science Laboratories <{bond,ogura}@cslab.kecl.ntt.co.jp>

Abstract

We present a method for extracting verbal (diathesis) alternations from a valency dictionary, based on comparison of selectional restrictions. The quality of match between selectional restrictions is evaluated according to an entropy-based measure with backing-off facility. We use the proposed method to derive a provisional listing of the range and distribution of verbal alternations in Japanese.

1 Introduction

This research represents a component of ongoing work on the reconstruction of a Japanese–English valency dictionary, as described in Baldwin et al. (1999). In the proposed valency dictionary design, dictionary entries are hierarchically described by way of the word, sense and frame levels. The basic set of arguments associated with a sense is described at the sense level and annotated by way of selectional restrictions and lexical fillers. Frames then take the form of expression filling features/constraints and a list of case slots, linked back to the sense-level argument description.

One key feature of the proposed dictionary structure is that inter-frame correspondences are explicitly identified in the form of alternations. We define a (diathesis) alternation to be a directed 1-to-1 relation from a source to a target frame. Alternation can affect a range of features including overall expression style, focus, and also case slot content and realisation. Individual case slots are affected by three operations: (1) transfer, where one case slot is mapped onto another, potentially undergoing modification of a range of morpho-syntactic features in the process; (2) deletion, where a given case slot is realised in the source but not the target frame; and (3) insertion, where a given case slot is realised in the target but not the source frame. With the unexpressed object alternation, for example, the source frame is made up of a subject and direct object case slot (i.e. is transitive), from which the direct object case slot is deleted and the subject case slot transferred unaltered to produce an intransitive frame. A prototypical effect of this comes with the English verb eat, as in He ate breakfast alternating with He ate.

As mentioned above, extraction of alternations takes place over a dictionary, namely the Goi-Tailei pattern-based valency dictionary (Ikehara et al. 1997; Shirai et al. 1997). The Goi-Tailei pattern-based valency dictionary is a rich source of Japanese–English predicate phrase transfer pairs, in which lexical selection is based on the selectional restrictions governing each case slot. Selectional restrictions take the form of concept nodes in the Goi-Tailei thesaurus tree and/or lexical fillers.

Throughout this paper, we will focus exclusively on Japanese verbal alternations. One characteristic of

1 Adjectival alternation is also possible in Japanese, typically...
Japanese verbal alternation which sets it apart from languages such as English is that it is commonly accompanied by relatively predictable lexical alternation (Jacobsen 1990). To give an example, "break" translates as kousu in its transitive form and kowaru in its intransitive form, and the causative-inchoative alternation is thus marked by -s/-re- lexical alternation. Other lexical alternation types to co-occur with verbal alternations include -φ-/ar-, -te/-s- and -φ-/e-. In order to be able to capture the full range of verbal alternations, therefore, the extraction process must be able to accommodate comparison of lexically similar but not necessarily identical verbs. This can be achieved without making any assumptions as to the manner of lexical alternation, through the observation that all such verb pairs invariably share the same kouji stem.

As stated above, case slot alternation can take the form of insertion, deletion or insertion. We assume alternations to be monotonic in terms of valency, that is a single alternation cannot involve both case slot insertion and deletion. In order to be able to merge together candidate alternations with maximum effectiveness, we apply the constraint that alternations must be either valency-reducing or valency-preserving (no insertions or deletions); with valency-preserving alternations, arbitrary means are used to normalise direction. Effectively all this does is to guarantee that the same analysis is obtained for frames displaying the same basic case slot correspondence. Note that it does not in any way restrict the descriptive power of the alternation extraction mechanism. We do, however, recognise that this can lead to a misconstrual of alternation direction (Baldwin and Tanaka 2000; Dorr and Olsen 1996).

3 Method

The extraction process involves taking all verb pairs sharing a common kouji stem, and exhaustively identifying all possible alternations between them through comparison of the selectional restriction-based characterisation of each case slot. We provide a backing-off mechanism to cope with near-misses. This is intended to pick up on miscellaneous lexicographic errors, but also provides the means to isolate any regularities in selectional restriction alternation (see below).

3.1 Scoring method

We score each set of matching selectional restrictions according to the final "quality of match" (after backing-off). This is achieved by fashioning an entropy-based score for each subtree of the Goi-Taikei thesaurus, based on its inverse token density, such that matches over sparser regions of the thesaurus structure are scored higher than those in the upper reaches of the tree. This is intended to reflect the plausibility of the match, in the sense that highly-demarcated selectional restrictions (low token density, high entropy) tend to indicate high annotational confidence on the part of the lexicographer. The chances of a match at such a level of specificity are lower than for selectional restrictions of greater token coverage and lower lexicographic commitment, a fact which we wish to reflect in a higher overall score for that alternation.

First, we derived the scores for each node of the thesaurus tree. We generated a list of morpheme types occurring in the EDR corpus (EDR 1995) and the frequency of each. For each such morpheme contained in the Goi-Taikei thesaurus, we next distributed the frequency between each of its senses. In Goi-Taikei, noun senses are listed in decreasing order of salience, such that the sense listed first is the most accessible sense and can be expected to occur most frequently. We draw on this sense ranking in distributing the morpheme token frequencies between the various senses, according to Zipf's law. Zipf's law states that the nth-ranking sense tends to occur with \( \frac{1}{n} \) the frequency of the first-ranking sense. We put this observation to practice in calculating a normalised weight \( w(l, x, j, k) \) for each sense of morpheme \( l, x, j, k \), such that \( \sum_l w(l, x, j, k) = 1 \).

In this way, we were able to estimate the total frequency of occurrence of each sense node in the thesaurus. We next calculated the "inverse token density" of each subtree of the thesaurus as the entropy of the combined frequencies of each node subsumed by it, in the manner of Resnik (1999). The inverse token density (itd) of node \( n_0 \) is calculated as:

\[
\text{itd}(n_0) = -\log p(n_0) = -\log \frac{\sum_{l, x, j, k \in n_0} \text{freq}(l, x, j, k)}{\sum_{l, x, j, k} \text{freq}(l, x, j, k)}
\]

where \( p(n_0) \) is the probability that an arbitrary lexeme of given sense will be subsumed under node \( n_0 \); \( l, x, j, k \in n_0 \) indicates that sense \( n_0 \) of lexeme \( l, x, j, k \) is contained in the subtree described by \( n_0 \); and \( n_0 \) is the root node. Nodes describing sparsely-populated subtrees are thus given higher weights than densely-populated subtrees, and as we ascend the tree, the node weights decrease monotonically, right down to weight 0 for the root node.

The quality of match between sense nodes \( n_0 \) and \( n_n \) is determined by way of the following equation:

\[
\text{match}(n_0, n_n) = 3 \text{itd}(\text{sub}(n_0, n_n)) - \text{itd}(n_0) - \text{itd}(n_n)
\]

where \( \text{sub}(n_0, n_n) \) is the least common hyponym node subsuming both \( n_0 \) and \( n_n \). In the case that \( n_0 \) and \( n_n \) are coincident, \( \text{sub}(n_0, n_n) = n_0 = n_n \), such that the overall match score becomes itd(\( n_0 \)) = itd(\( n_n \)). It is important to realise that match can be negative in the face of high levels of backing-off up the tree structure in order to reach the least common hyponym node.

Naturally, a single set of selectional restrictions can include multiple sense nodes. In matching a pair of selectional restrictions, we determine the spanning bipartite mapping between them for which the mean match score for connected sense nodes is maximised. The overall score for a given frame pair is then determined as the sum of the averaged match scores for each selectional restriction pairing.

3.2 Clustering candidate alternations

Alternation is defined to be an 1-to-1 relation, that is, multiple case slot transfer to a single case slot cannot take place. Additionally, alternations must be valency monotonous, as described above. In exhaustively determining all possible candidate alternations between a given frame pair, therefore, we take the frame of lower valency and derive all surjective mappings from the frame of higher valency, and treat any residue case
slots in the source frame as having been deleted. In the case that the frames are of equal valency, we use an arbitrary method based on the alphabetic order of case markers to select one frame as the source and the other as the target frame, and generate all isomorphic mappings between them.

In order to be able to pick up on different types of lexical alternation, each candidate alternation is tagged with the directed pair of non- coincident suffixes of the source and target frame head verbs, which we term \textit{SUFF}. In the case of the verb pair \textit{kousu}/\textit{kousuru}, for example, \textit{SUFF} takes the form \textit{sw}/\textit{veru}. This combines with the case slot mapping and score to make up a triple for each candidate alternation. The case slot mapping is described by way of the part-of-speech of each case slot (NP or S in our case), the selective restrictions on the source and target case slots, and the source of and target case markers. For the causative- inchoative alternation, this takes the form:

\[(NP_1[-\{ga\}]-\phi)(NP_2[-\{o\}--\{ga\}])\]

where "\(\cdot\)" indicates that the selective restrictions on the corresponding case slots coincide. In presenting alternations below, we omit explicit description of selective restrictions in the case that they are preserved under alternation.

As noted above, the \textit{match} function can return a negative value, meaning that the overall score for a given alternation can be negative given sufficiently high levels of backing-off. If this occurs, that candidate alternation is automatically removed from processing on the grounds of it being too implausible to warrant consideration. After having pruned off negatively-scoring candidate alternations, we next select the best-scoring candidate alternation for each frame pair. In the case that a tie in score is produced, we select that candidate alternation which preserves case marking for the most case slot transfers. If the tie still remains, then we have no reasonable grounds for selecting between the candidate alternations, and no output is produced.

Having selected a unique candidate alternation for each frame pair (or no candidate alternation in the case of the top-ranking candidate alternation having a negative score or a tie not having been broken), we next turn to the clustering of alternation tokens. In the first step of clustering, we combine together the scores for all candidate alternations with the same \textit{SUFF} value and alternation mapping. This goes some way to detecting lexical alternations, but \textit{toru}/\textit{toreru} "move/come away" and \textit{tokoro}/\textit{tokoreru} "solve/be solved", for example, would not be merged together despite conforming to the \(-\{o\}--\{e\}\) lexical alternation. We thus convert \textit{SUFF} tuples into lexical alternation paradigms and recluster.\footnote{Lexical alternation paradigms take the form of both simple lexical alternations such as \(-\{o\}--\{e\}\) and verb morpheme tags such as \textit{PASSIVE}, \textit{ACTIVE} and \textit{CAUSATIVE}, as appropriate for the given \textit{SUFF} pairing.}

In order to extract core alternations, for alternations where non-alternating case slots occur at the tail of the source frame, we remove each final non-alternating case slot one at a time, and add the combined score for the original alternation to the most basic alternation produced in this way which is observed in the data. At the same time, we retain the original alternation within the output.

4 Results

In this section, we present the results of alternation extraction in the form of the 10 alternations with the highest cumulative scores out of a total of 71 extracted alternation types (Fig. 1), and 10 alternations with the highest average scores (Fig. 2); in the latter case, any alternations occurring less than 3 times have been excluded from the data. For each presented alternation, any lexical alternation accompanying the (case slot) alternation is indicated, and a name given. We also present a selection of representative verbs undergoing that alternation. With all alternations presented in the two figures, selectional restrictions were preserved under alternation.

Overall, the results are credible, although we do not have any direct means of evaluating them empirically. While they do not appear in the presented data, a number of synthetic and auxiliary verb co- occurrence-based alternations were also observed in the data. Surprisingly few lexical alternations appeared in the cumulative score ranking, partly due to the wide range of lexical alternation apparent in the dictionary and limited number of instances for each. A few alternations involving alternation of selectional restrictions as well as case marking were observed, although they tended to be feature well down in both rankings. Typically, selectional restriction alternation was between a daughter node and its parent or involved the insertion of a node such as \textit{ANIMAL} to complement a more general node such as \textit{AGENT}.

One effect that is clear in Fig. 2 is that the same basic alternations are on occasion repeated in the presence of non- alternating peripheral case slots, such as occurs with the final alternation where the expressed object in alternation is produced with a third, dative-marked (ni) case slot. The names of all such alternations are marked with an asterisk. This effect was not as obvious within the cumulative score ranking, as core alternations were scored up based on such alternation variants.

A more subtle effect reflected in the example verbs is that our decision to base analysis of alternations on case marker and selectional restriction alternation can group together verbs with disparate semantics. To take the example of the unexpressed dative alternation (Fig. 1), the alternating dative case slot with \textit{kakunin}suru "to confirm" and \textit{noru} "get on" represent quite different case-roles, with the first being an experiencer (indirect object) and the second a local allative. One immediate means of splitting apart such verbs would be to include the original case-role mark-up within the alternation description, a possibility we leave for future research.

5 Discussion

While it is difficult to directly compare this work to past research, it is worthwhile outlining other methods applied to the same basic task. Baldwin and Tanaka (2000) proposed a total of three alternation extraction procedures, using a full match or edge count- based similarity measure rather than entropy. Schule (2000) similarly derived subcategorisation frames with selectional restrictions through the device of selectional preference, and then classified verbs into Levin classes according to the subcategorisation frames they occur with. Interestingly, she gained better results based on simple syntactic behaviour than when adding in selectional restrictions. McCarthy (2000) similarly derived subcategorisation frames with selectional restrictions from a corpus, using the minimum distance length principle, and then classified verbs as participating in a select set of alternations according to similarity in selectional restrictions on corresponding case slots. One aspect of this research which sets it apart from that of both Schule and McCarthy is that we compare selectional restrictions between alternating case slots for a given verb, rather than comparing corre-
responding case slots in equivalent frames across different verbs. Also, we are seeking to posit a set of alternations, rather than simply classifying according to an
pre-defined alternation set.
To summarise, this research is targeted at the extraction of Japanese verbal alternations from a dictionary annotated with selectional restrictions, based on the assumption that selectional restrictions are preserved under alternation. We proposed an entropy-based scoring method for evaluating both the degree of similarity and quality of match of a pair of selectional restrictions.
This was used to score candidate alternations, and the candidate alternations clustered together through analysis of lexical alternation and the core component of each candidate alternation.

References