Detecting Alternation Instances in a Valency Dictionary*

Francis Bond‡, Timothy Baldwin† and Sanae Fujita‡
† Center for the Study of Language and Information (CSLI)
210 Panama Street, Stanford, CA 94305-4115, USA
<tbaldwin@csli.stanford.edu>
‡ NTT CS Labs., Nippon Telegraph and Telephone Corporation
2-4 Hikari-dai, Seika-cho, Soraku-gun, Kyoto, 619-0237, JAPAN
<{bond,sanae}@cslab.kecl.ntt.co.jp>

Abstract

This research is aimed at developing a hierarchical alternation-based lexical architecture for machine translation, through analysis of diathesis alternation patterns in an existing Japanese–English transfer valency dictionary. We present a basic method for extracting alternations, and propose enhancements through analysis of both the morphological correspondence between the participant verbs and the English translations.

1 Introduction

This paper draws together and expands upon previous word on the automatic extraction of alternating case frames from a dictionary (Baldwin and Tanaka, 2000; Baldwin and Bond, 2002), and focuses particularly on means for enhancing the alternation extraction process.

Currently, alternations are extracted fully automatically and without the use of a bootstrap or any other external data. That is, we do not start with a pre-conception of how verbs alternate (e.g. of the type of Levin's English alternation inventory (Levin, 1993), or Jacobsen's list of basic Japanese alternations (Jacobsen, 1992)) and identify particular instances of each alternation type. Rather, we make assumptions about the nature of alternation, and extract all inter-valency frame mappings that fit in with these assumptions. We then apply a number of constraints over the extracted data to reduce noise as far as is possible.

This research is targeted at the Goi-Taikei Japanese–English valency dictionary (Ikehara et al., 1997), as used in the ALT-J/E machine translation (MT) system (Ikehara et al., 1991). The Goi-Taikei valency dictionary describes each Japanese verbal expression as a case frame headed by the verb in question, linked to an English translation "skeleton". Each case slot is annotated with a discrete set of prototypical case markers, part of speech (NP or S), an obligatoriness flag, and a list of selectional restrictions and lexical fillers. The selectional restrictions take the form of nodes within the Goi-Taikei thesaurus tree. The Goi-Taikei thesaurus is an unbalanced tree of 2,710 nodes, connected by links showing either hyponymic (is-a) or meronymic (has-a) relations.

We define a (diathesis) **alternation** to be a 1-to-1 relation from a source to a target frame, which involves at least one of: (i) case marking variation between corresponding case slots, (ii) case slot deletion, and (iii) case slot insertion. To give an example, the causative-inchoative alternation occurs between a transitive and intransitive frame, as shown in (1) for *akeru/aku.*[1]

(1) Kim-ga doa-o aketa / doa-ga aita
Kim-NM door-AC opened door-NM opened
'Kim opened the door' / 'The door opened'

This example illustrates the nature of case marking variation, the principal form of morphological variation considered in this research.

The remainder of this paper is structured as follows. Section 2 discusses theoretical issues and assumptions surrounding alternations. In Section 3, we then discuss the methodology employed to derive alternation data and briefly evaluate the alternation extraction method. We conclude the paper in Section 4.

2 Alternations: issues and assumptions

In the case of Japanese, alternation can be: (i) unmarked on the verb (**analytical** alternation, as seen for *hiraku* "open$_{\text{intrans/trans}}$"), (ii) marked on the verb stem by often-predictable lexical variation (**lexical** alternation, such as between *akeru* "open$_{\text{trans}}$" and *aku* "open$_{\text{intrans}}$"), (iii) marked by way of verbal inflection or a verb morpheme

*構文意味辞書における類似構文の融合方法

[1] The following abbreviations are used in glosses: -NM = nominative and -AC = accusative.

(**synthetic** alternation, such as occurs with the passive morpheme *(r)are*).

We take after Baldwin and Bond (2002) in making a number of assumptions about alternations in this research:

1. The selectional restrictions and lexical fillers on matching case slots are preserved under alternation
2. Alternations are monotonic in valency terms
3. A given alternation type has fixed direction

The first of these assumptions states that corresponding case slots in the two alternants of a given alternation token, display the same selectional restrictions and lexical fillers (Baldwin and Tanaka, 2000). This provides the means for extracting alternation candidates from the valency dictionary.

The second assumption states that a given alternation type cannot involve both case slot insertion and deletion, and constrains the space of alternation mappings we must consider between a given valency frame pair.

The third and final assumption constrains the direction of a given alternation type in all its realisations. We have no immediate means of determining for each alternation token which is the base and which the derived form, however. Our solution is to impose direction on the alternation type, and apply this to all instances thereof. This is achieved by stipulating that all alternations are either valency-decreasing or valency-maintaining, and arbitrarily normalising the direction of valency-maintaining alternations using the alphabetic order of the case markers on case slots which undergo modification.

3 Alternation extraction method

Alternations are extracted by first taking all pairs of case frames from the base valency dictionary which share some (kanji) prefix, and identifying the most plausible (if any) alternation for each from the set of all possible valence-monotonic case slot mappings between them. We then analyse trends in the alternation data and apply a number of filters over the extracted data to reduce noise.

3.1 Scoring case slot matches

The quality of match between case slots is quantitatively described by comparing the relative proximity of the selectional restrictions describing each, within the Goi-Taikei thesaurus tree. As stated above, selectional restrictions are provided as thesaurus node indices, and the greater the topological overlap between and conceptual cohesion within the regions described for the two case slots, the higher the match quality. This is intended to reflect the intuition that the higher the specificity of the selectional restrictions, the greater the confidence of the lexicographer in their integrity.

The scoring method adopted in this paper is identical to that of Baldwin and Bond (2002), to which the reader is referred for full details.

The conceptual cohesion of the subtree described by a given node is modelled by way of the relative entropy of the token population of that region, as determined from corpus occurrence statistics (in the manner of Resnik (1999)). Having determined the score for each node, we can evaluate the semantic proximity of two nodes as the relative disparity between the conceptual cohesion of each and their least common hypernym (*lch*); this is determined by way of the weighted difference between the conceptual cohesion value for the *lch* node, and that of the two original nodes (in the form of $match_c$). Importantly, $match_c$ can be negative in the face of high levels of backing-off up the tree structure in order to reach the *lch*.

In scoring a pair of selectional restrictions, we determine the spanning bi-partite mapping between them for which the mean $match_c$ score for connected sense nodes is maximised. The overall score for a given alternation is defined as the sum of slot-wise scores for only those case slots which are mapped from/to, i.e. deleted case slots do not enter into calculations. As noted above, negative values can be returned for individual slot mappings if there is a large discrepancy in the selectional restrictions. In the case that the overall score is negative, we reject that alternation candidate outright.

3.2 Resolving alternation ambiguity

For each case frame pair, we return the best-scoring alternation(s), recalling that any negatively-scoring alternation candidates are automatically filtered off. In the case that a tie in score is produced, we select that alternation candidate which preserves case marking for the highest proportion of case slot mappings, under the assumption that alternations are conservative in their scope for modification. If the tie still remains, then we have no reasonable grounds for selecting between the alternation candidates, and no output is produced. With case frames of

	Analytical alts.		*Lexical alts.*		*Synthetic alts.*		*Overall*		
	Prec	*Rec*	*Prec*	*Rec*	*Prec*	*Rec*	*Prec*	*Rec*	*F-score*
Basic	76.2%	100%	62.5%	100%	25.7%	100%	60.9%	100%	75.7%
No compound verbs	78.2%	100%	69.8%	100%	28.6%	100%	66.4%	100%	79.8%
Cutoff = 2	81.8%	64.9%	82.4%	75.7%	33.3%	64.3%	73.2%	67.3%	70.1%
Cutoff = 2/1	82.2%	79.3%	82.9%	78.4%	36.7%	78.6%	74.4%	79.0%	76.6%
Cutoff = 2/1/0	82.4%	80.2%	83.8%	83.8%	36.7%	78.6%	74.9%	80.9%	77.8%

Table 1: Evaluation of extracted alternation candidates

equal valency, it can happen that in the highest-scoring "alternation", no case marker variation has in fact taken place (i.e. the case frames are identical modulo linear reordering of case slots).[2] Based on our assumptions on the nature of alternations in Section 2, this does not construe a true alternation. Given that this is the best analysis for the given case frame pairing, we consider that no alternation exists.

Additionally, in the case of multiple alternations mapping onto a single valency frame, we take the single best alternation candidate involving that target valency frame.

3.3 Filter the alternation data

Having extracted the alternation data, we next propose a number of filters to filter out noise (i.e. errant alternation candidates) from the system output.

The only stipulation on verb pairs we consider for alternation extraction is that they share some common prefix. This is intended to ensure that there is at least a tenuous semantic link between them. For a verb pair such as 申し合わせる *mōshiawaseru* "to arrange" and 申し出る *mōshideru* "to report", our matching algorithm proposes alternation candidates, but the semantics of the verbs are too far removed to justify an alternation analysis. In cases such as this, we assume that if an alternation is to exist, it will be between the verb-verb compound and the root verb, that is 申す *mōsu* "to say". We therefore first apply a filter stating that no alternation may exist between two (non-identical) verb-verb compounds.

Secondly, it is possible to raise the cutoff score on alternations above 0, and assuming that there is some correlation between higher scores and better quality alternations, attain more reliable results.

Third, given that each valency frame is linked to an English translation, it is possible to analyse whether the root English verb translation for the two valency frames is the same, to gain some insight into the semantic similarity of the two valency frames. As with our requirement on the two verbs in question sharing a common kanji prefix, this is expected to increase extraction accuracy. This requirement can be combined with a modified cutoff. Closely related to this method of common English translation root, some valency frames are linked to the exact same translation, which is a strong indicator of high alternation quality.

3.4 Results of alternation extraction

A total of 2,777 alternation tokens were detected in the valency dictionary, from a total of 13,880 verbal case frames. After removing competing alternations (i.e. lower-scoring alternations mapping onto a common valency frame), 1,653 alternation tokens remained, making up a total of 373 alternation types.

We evaluated the quality of the extracted alternations based on a random sample of 261 alternation tokens from the 1,653 remaining alternations. Each alternation token is manually classified according to its type (analytical/lexical/synthetic), and scored as correct or incorrect. "Correct" alternations are defined as those which are either motivated alternations analogous to Levin's set of English alternations (Levin, 1993), or transfer dictionary quirks such as adjunct optionality being described by way of two valency frames, with and without the adjunct. Based on this definition, we calculated the precision and recall for the basic system, and then went on to apply the various filters described in Section 3.3. Precision (*Prec*) is defined as the

[2] In fact, the detection of identical case frames in the dictionary may prove valuable in the grander scheme of the lexicon overhaul. We would still not want to treat them as alternations, however.

percentage of extracted alternations of each basic type which is in fact a true alternation according to the annotated data. Recall (*Rec*), on the other hand, is relative to the full set of 261 annotated alternation tokens, and a measure of how many of the total body of correct alternations are extracted by each procedure. It does not reflect the absolute recall of each method over the full dictionary file, because we do not have annotation data for valency frame pairs which the basic method does not extract as alternation candidates.

The results of extraction are presented in Table 1. Firstly, the basic system (*Basic*) resulted in an overall precision of 60.9% (and recall of 100%, with 161 correct alternations), as reported by Baldwin and Bond (2002). Analytical and lexical alternations show considerably higher accuracy than synthetic alternations, at a precision of only 25.7%. Removing all alternations between compound verbs (*No compound verbs*) results in an appreciable gain in overall precision, up to 66.4%, with no loss in recall. That is, this filter removed noise without alleviating any correct alternation data. Next, an across-the-board cutoff score of 2 for alternation candidates (*Cutoff = 2*) bumps up overall precision to 73.2%, but leads to a sharp fall in recall to 67.3%. That is, the filter is able to remove some errant alternation candidates, but sacrifices many correct alternation candidates in the process. If we introduce a 2-tier cutoff approach, with a threshold of 1 on verb pairs with a common English translation root and threshold of 2 for other verb pairs (*Cutoff = 2/1*), overall recall recovers to 79.0% and precision also appreciates slightly to 74.4%. Finally, if we implement a three-tier scoring system, with the first two tiers identical to those above, and a third tier at the original threshold of 0 for valency frames which are translated identically, then we achieve a further slight gain in both precision and recall, to 74.9% and 80.9%, respectively.

Based on the overall F-score (*F-score* – the harmonic mean of precision and recall), the simple no V-V compound filter produces the best results, although if we wished to generate reference data for the analysis of Japanese alternation types, precision would become our main concern, in which case the final system configuration would be optimal.

4 Conclusion

We extracted alternations in an unsupervised manner, relying 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 case frame mappings and analyse whether an alternation could be found between a given case frame pairing. Through the application of a range of filters operating over the composition of the Japanese verbs participating in each alternation pair, and also the make-up of their English translations, an overall precision of 74.9% and recall of 80.9% were attained.

Acknowledgements

This research was supported in part by the Research Collaboration between NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation and CSLI, Stanford University. We would like to thank Tomoko Kawaguchi for her help with the alternation annotation.

References

Timothy Baldwin and Francis Bond. 2002. Alternation-based lexicon reconstruction. In *Proc. of the 9th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI 2002).* (to appear).

Timothy Baldwin and Hozumi Tanaka. 2000. Verb alternations and Japanese — how, what and where? In *Proc. of the 14th Pacific Asia Conference on Language, Information and Computation (PACLIC 14)*, pages 3–14.

EDR, 1995. *EDR Electronic Dictionary Technical Guide.* Japan Electronic Dictionary Research Institute, Ltd. (In Japanese).

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 *Proc. of the Third Machine Translation Summit (MT Summit III)*, pages 101–106, Washington DC.

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

Wesley M. Jacobsen. 1992. *The Transitive Structure of Events in Japanese.* Kurosio Publishers.

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

Philip Resnik. 1999. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research (JAIR)*, 11:95–130.