Detecting Verbal Participation in Diathesis Alternations

Diana McCarthy
Cognitive & Computing Sciences,
University of Sussex
Brighton BN1 9QH, UK

Anna Korhonen
Computer Laboratory,
University of Cambridge, Pembroke Street,
Cambridge CB2 3QG, UK

Abstract

We present a method for automatically identifying verbal participation in diathesis alternations. Automatically acquired subcategorization frames are compared to a hand-crafted classification for selecting candidate verbs. The minimum description length principle is then used to produce a model and cost for storing the head noun instances from a training corpus at the relevant argument slots. Alternating subcategorization frames are identified where the data from corresponding argument slots in the respective frames can be combined to produce a cheaper model than that produced if the data is encoded separately.[1].

1 Introduction

Diathesis alternations are regular variations in the syntactic expressions of verbal arguments, for example *The boy broke the window* ↔ *The window broke*. Levin's (1993) investigation of alternations summarises the research done and demonstrates the utility of alternation information for classifying verbs. Some studies have recently recognised the potential for using diathesis alternations within automatic lexical acquisition (Ribas, 1995; Korhonen, 1997; Briscoe and Carroll, 1997).

This paper shows how corpus data can be used to automatically detect which verbs undergo these alternations. Automatic acquisition avoids the costly overheads of a manual approach and allows for the fact that predicate behaviour varies between sublanguages, domains and across time. Subcategorization frames (SCFs) are acquired for each verb and a hand-crafted classification of diathesis alternations filters potential candidates with the correct SCFs. Models representing the selectional preferences of each verb for the argument slots under consideration are then used to indicate cases where the underlying arguments have switched position in alternating SCFs. The selectional preferences models are produced from argument head data stored specific to SCF and slot.

The preference models are obtained using the minimum description length (MDL) principle. MDL selects an appropriate model by comparing potential candidates in terms of the cost of storing the model and the data stored using that model for each set of argument head data. We compare the cost of representing the data at alternating argument slots separately with that when the data is combined to indicate evidence for participation in an alternation.

2 SCF Identification

The SCFs applicable to each verb are extracted automatically from corpus data using the system of Briscoe and Carroll (1997). This comprehensive verbal acquisition system distinguishes 160 verbal SCFs. It produces a lexicon of verb entries each organised by SCF with argument head instances enumerated at each slot.

The hand-crafted diathesis alternation classification links Levin's (1993) index of alternations with the 160 SCFs to indicate which classes are involved in alternations.

3 Selectional Preference Acquisition

Selectional preferences can be obtained for the subject, object and prepositional phrase slots for any specified SCF classes. The input data includes the target verb, SCF and slot along with the noun frequency data and any prepo-

[1]This work was partially funded by CEC LE1 project "SPARKLE". We also acknowledge support from UK EPSRC project "PSET: Practical Simplification of English Text".

sition (for PPs). Selectional preferences are represented as Association Tree Cut Models (ATCMs) as described by Abe and Li (1996). These are sets of classes which cut across the WordNet hypernym noun hierarchy (Miller et al., 1993) covering all leaves disjointly. Association scores, given by $\frac{p(c|v)}{p(c)}$, are calculated for the classes. These scores are calculated from the frequency of nouns occurring with the target verb and irrespective of the verb. The score indicates the degree of preference between the class (c) and the verb (v) at the specified slot. Part of the ATCM for the direct object slot of *build* is shown in Figure 1. For another verb a different level for the cut might be required. For example *eat* might require a cut at the **FOOD** hyponym of **OBJECT**.

Finding the best set of classes is key to obtaining a good preference model. Abe and Li use MDL to do this. MDL is a principle from information theory (Rissanen, 1978) which states that the best model minimises the sum of *i* the number of bits to encode the model, and *ii* the number of bits to encode the data in the model. This makes the compromise between a simple model and one which describes the data efficiently.

Abe and Li use a method of encoding tree cut models using estimated frequency and probability distributions for the data description length. The sample size and number of classes in the cut are used for the model description length. They provide a way of obtaining the ATCMs using the identity $p(c|v) = A(c,v) \times p(c)$. Initially a tree cut model is obtained for the marginal probability $p(c)$ for the target slot irrespective of the verb. This is then used with the conditional data and probability distribution $p(c|v)$ to obtain an ATCM as a by-product of obtaining the model for the conditional data. The actual comparison used to decide between two cuts is calculated as in equation 1 where C represents the set of classes on the cut model currently being examined and S_v represents the sample specific to the target verb.[2].

$$\frac{|C|}{2} log|S_v| + \sum_{c \epsilon C} -freq_c \times \log \frac{p(c|v)}{p(c)} \qquad (1)$$

In determining the preferences the actual encoding in bits is not required, only the relative cost of the cut models being considered. The WordNet hierarchy is searched top down to find the best set of classes under each node by locally comparing the description length at the node with the best found beneath. The final comparison is done between a cut at the root and the best cut found beneath this. Where detail is warranted by the specificity of the data this is manifested in an appropriate level of generalisation. The description length of the resultant cut model is then used for detecting diathesis alternations.

[2] All logarithms are to the base 2

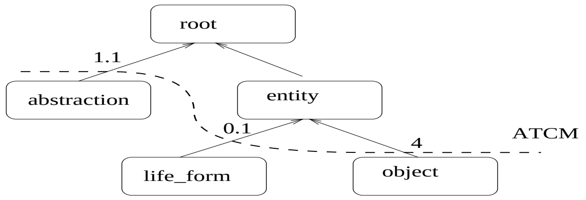

Figure 1: ATCM for *build* Object slot

4 Evidence for Diathesis Alternations

For verbs participating in an alternation one might expect that the data in the alternating slots of the respective SCFs might be rather homogenous. This will depend on the extent to which the alternation applies to the predominant sense of the verb and the majority of senses of the arguments. The hypothesis here is that if the alternation is reasonably productive and could occur for a substantial majority of the instances then the preferences at the corresponding slots should be similar. Moreover we hypothesis that if the data at the alternating slots is combined then the cost of encoding this data in one ATCM will be less than the cost of encoding the data in separate models, for the respective slot and SCF.

Taking the causative-inchoative alternation as an example, the object of the transitive frame switches to the subject of the intransitive frame: *The boy broke the window* $\leftrightarrow$ *The window broke.* Our strategy is to find the cost of encoding the data from both slots in separate ATCMs and compare it to the cost of encoding the combined data. Thus the cost of an ATCM for *i* the sub-

Table 1: Causative-Inchoative Evaluation

	verbs	
true positives	begin end change swing	4
false positives	cut	1
true negatives	choose like help charge expect add feel believe ask	9
false negatives	move	1
total		15

ject of the intransitive and *ii* the object of the transitive should exceed the cost of an ATCM for the combined data only for verbs to which the alternation applies.

5 Experimental Results

A subcategorization lexicon was produced from 10.8 million words of parsed text from the British National Corpus. In this preliminary work a small sample of 30 verbs were examined. These were selected for the range of SCFs that they exhibit. The primary alternation selected was the causative-inchoative because a reasonable number of these verbs (15) take both subcategorization frames involved. ATCM models were obtained for the data at the subject of the intransitive frame and object of the transitive. The cost of these models was then compared to the cost of the model produced when the two data sets were combined.

Table 1 shows the results for the 15 verbs which took both the necessary frames. The system's decision as to whether the verb participates in the alternation or not was compared to the verdict of a human judge. The accuracy was 87% ($\frac{4+9}{4+1+9+1}$). Random choice would give a baseline of 50%. The cause for the one false positive *cut* was that *cut* takes the middle alternation (*The butcher cuts the meat* $\leftrightarrow$ *the meat cuts easily*). This alternation cannot be distinguished from the causative-inchoative because the SCF acquisition system drops the adverbial and provides the intransitive classification.

Performance on the simple reciprocal intransitive alternation (*John agreed with Mary* $\leftrightarrow$ *Mary and John agreed*) was less satisfactory. Three potential candidates were selected by virtue of their SCFs *swing;with add;to* and *agree;with*. None of these were identified as taking the alternation which gave rise to 2 true negatives and 1 false negative. From examining the results it seems that many of the senses found at the intransitive slot of *agree* e.g. *policy* would not be capable of alternating. It is at least encouraging that the difference in the cost of the separate and combined models was low.

6 Conclusions

Using MDL to detect alternations seems to be a useful strategy in cases where the majority of senses in alternating slot position do indeed permit the alternation. In other cases the method is at least conservative. Further work will extend the results to include a wider range of alternations and verbs. We also plan to use this method to investigate the degree of compression that the respective alternations can make to the lexicon as a whole.

References

Naoki Abe and Hang Li. 1996. Learning word association norms using tree cut pair models. In *Proceedings of the 13th International Conference on Machine Learning ICML*, pages 3–11.

Ted Briscoe and John Carroll. 1997. Automatic extraction of subcategorization from corpora. In *Fifth Applied Natural Language Processing Conference.*, pages 356–363.

Anna Korhonen. 1997. Acquiring subcategorisation from textual corpora. Master's thesis, University of Cambridge.

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

George Miller, Richard Beckwith, Christine Felbaum, David Gross, and Katherine Miller, 1993. *Introduction to WordNet: An On-Line Lexical Database.* ftp//clarity.princeton.edu/pub/WordNet/5papers.ps.

Francesc Ribas. 1995. *On Acquiring Appropriate Selectional Restrictions from Corpora Using a Semantic Taxonomy.* Ph.D. thesis, University of Catalonia.

J. Rissanen. 1978. Modeling by shortest data description. *Automatica*, 14:465–471.