Automatic Extraction of Subcategorization Frames from Corpora - Improving Filtering with Diathesis Alternations

Anna Korhonen
Computer Laboratory
University of Cambridge
Pembroke Street, Cambridge CB2 3QG, UK
alk23@cl.cam.ac.uk

Abstract

Attempts to extract subcategorization information from textual corpora by shallow parsing followed by statistical filtering of alternatives proposed for specific predicates have met with some success (Briscoe & Carroll, 1997) but are not yet accurate enough. Examination of the errors suggests that the filtering of spurious hypotheses is the source of most errors in the system. This paper builds on the framework described in (Briscoe and Carroll, 1997) and proposes a knowledge-based approach for improvement of the filtering phase of the system.

1 Background

Manual development of large subcategorised lexicons has proved very difficult because predicates change behaviour between sublanguages, domains and across time. Yet current parsers depend crucially on such information, and probabilistic parsers would greatly benefit from accurate information concerning relative likelihood of different subcategorization frames of a given predicate. This suggests that automatic construction of subcategorization dictionaries from textual corpora is a more promising method to apply. Briscoe & Carroll (1997) propose a technique and implemented system for constructing a subcategorization dictionary from textual corpora. Their system is capable of distinguishing 160 subcategorization classes, and able to both assign classes to individual verbal predicates and to rank them according to relative frequency. As described in Briscoe & Carroll (1997), the system consists of six overall components which are applied in sequence to sentences containing a specific predicate in order to retrieve a set of subcategorization classes for that predicate: a tagger, a lemmatizer, a probabilistic LR parser, a patternset extractor, a pattern classifier, and a patternsets evaluator. Even though the system has met with some success it is not yet accurate enough. The experimental evaluation performed by Briscoe & Carroll shows that the filtering of spurious hypotheses in the patternsets evaluator stage is the weak link of the system.

2 Filtering

In the current filter, the set of putative classes are filtered, following Brent (1993) by hypothesis testing on binomial frequency data. The system first records the total number of patternsets n for a given predicate, the number of these patternsets containing a pattern supporting an entry for given class m, and estimates of the probability that a pattern for a class i will occur with a verb which is not a member of subcategorization class i.

Briscoe & Carroll estimate the above probability by first extracting the number of verbs which are members of each class in the ANLT dictionary (Boguraev *et al.* 1987), and converting this

to a probability of class membership by dividing by the total number of verbs in the dictionary; and secondly, by multiplying the complement of these probabilities by a probability of a pattern for class i, defined as the number of patterns for i extracted from the Susanne corpus divided by the total number of patterns. According to this, the probability of verb v not of class i occurring with a pattern for class i is calculated by:

$$p(v-i) = \left(1 - \frac{|anlt\_verbs\_in\_class\_i|}{|anlt\_verbs|}\right) \frac{|patterns\_for\_i|}{|patterns|} \quad (1)$$

The probability of an event with probability p happening exactly m times out of n attempts is given by the following binomial distribution:

$$P(m,n,p) = \frac{n!}{m!(n-m)!} p^m (1-p)^{n-m} \quad (2)$$

The probability of the event happening m or more times is:

$$P(m+,n,p) = \sum_{i=m}^{n} P(i,n,p) \quad (3)$$

Finally, *P(n,m,p(v-i))* is the probability that m or more occurrences of patterns for i will occur with a verb which is not a member of i, given n occurrences of that verb. A threshold is set of less than or equal to 0.05 which yields a 95% or better confidence that a high enough proportion of patterns for i have been observed for the verb to be in class i.

Briscoe and Carroll showed that the performance of this filter for classes with less than 10 exemplars is around chance and a simple heuristic of accepting all classes with more than 10 examples would have produced broadly similar results to the filter for the verbs examined. There are several reasons why the filter may be performing poorly. One reason may be that in the current filter, the probability of generating a subcategorization class for a given verb is often lower than the error probability for that class. So where the probability of the class given the verb is lower than the estimate, the filter incorrectly rejects the class of that verb. Another reason may have to do with the way the filter defines the amount of evidence required for some verb to be a member of that class. It does this partly on the basis of the estimates of the hypothesis generator. The patterns returned by the hypothesis generator may be inaccurate if the sample from the Susanne corpus was not representative enough. It is also possible that to achieve better results the number of verbs in ANLT dictionary class i should be weighted by the frequency of these verbs.

3 Improving Filtering with Diathesis Alternations

As the hypothesis filter is the main source of error in the system, we tried adding a knowledge-based component seeded with general linguistic information to improve the filtering performance. We decided to make use of *diathesis alternations*. In the linguistic literature, diathesis alternations mean alternative ways in which verbs can express their arguments. Different types of alternations are identified, and verbs can be classified according to which alternations they participate in. A simple way of looking at alternations is to examine whether there are likely or unlikely correlations between two patterns of complementation. For example, we could consider an English verb which takes both a noun phrase and a sentence as its complement:

a) *It bothered John that Bill was so clever*

On the basis of linguistic intuition, we could predict that it is very unlikely that such a verb would only take a sentential complement:

b) **It bothered that Bill was so clever*

If the system proposes these two complementation patterns for the same verb, this would tell us that one is probably wrong. So the rule NP (SFIN or NP + SFIN) is very likely to hold. These type of observations are called *alternation generalisations* and can be expressed as *alternation rules*:

a) NP $\longrightarrow$ SFIN

b) NP $\longrightarrow$ NP + SFIN

For example, the verb *bother* participates in the alternation NP $\longrightarrow$ NP + SFIN but not in the alternation NP $\longrightarrow$ SFIN. It was decided to add 'alternation rules' of this general form to the subcategorization extraction system to aid the filter's choice of correct subcategorization classes for a verb. The idea was to equip the current system with more conditions for the acceptance of classes, and thus produce more accurate results.

3.1 Alternation Rules

To develop the idea, we used the seven manually analysed test verbs from (Briscoe and Carroll, 1997) as training data. We analysed the errors the system had made with them, and examined whether alternations could be found to correct the mistakes. The filter performed especially badly with nine subcategorization classes. For these classes we considered possible alternations. Mainly linguistic intuition was used when searching alternations but we also checked manually that the proposed rules would correct most of the incorrect occurrences of a class. Table 3 shows the results of the examination, and the alternation rules found. The first column shows the number and the code of the problematic class[1]. The second column gives an example of this class. The third column shows the alternating class, and the fourth column an example of this. The rules proceed from left to right. For example, the rule in the first row indicates that the intransitive class, 22, is accepted for a verb only if the transitive class, 24, is also proposed for this verb.

After finding the set of appropriate alternation rules, the next task was to rate how reliable the individual rules are. Writing them all as 100% rules would have corrected most of the errors in the training data but would have had only a slight chance of generalising to any other set of verbs. We decided not to rate the rules purely on linguistic grounds but rather to focus on empirical approximations of the productivity of these alternations. To get the required empirical approximations, we extracted alternation information from the ANLT dictionary. The method used was to extract all the verbs taking a problematic subcategorization class in the ANLT dictionary, and then all the verbs taking its proposed alternating class. Then, an intersection of these was built, and we calculated how many common verbs took both classes with respect to the intersection. Using this method, the nine alternation rules were assigned the probabilities shown in the fifth column of Table 3.

In addition to these nine alternation rules developed on basis of the training data, an extended set of alternation rules was introduced. No training data was used but it was decided to examine alternation correlations between all the possible subcategorization classes occurring in ANLT

[1] When we have several classes separated by a slash that means that the classifier was not able to decide on one correct analysis. In these cases, the correct class is one or the other, or in a few cases both.

dictionary. This was achieved by deriving a matrix of all classes against all classes. An extended set of 5468 rules was generated using this method.

Class	**Example**	**Alternating Class**	**Example**	**Rule Probability**
22: INTRANS	I went	**24:** NP	He loved her	.7065
25/26: NP-ADJP / NP-ADJP-PRED	He painted the car black/ She considered him foolish	**24:** NP	He loved her	.9103
29/30: NP-AS-NP / NP-AS-NP-SC	I sent him as a messenger/ She served the firm as a researcher	**24:** NP	He loved her	.3333
37/38/144: NP-NP / NP-NP-PRED /SUBCAT OCNP, SUBTYPE RAIS	She asked him his name/ They appointed him professor/ He considers Fido a fool	**24:** NP	He loved her	.8519
49/50: NP-PP / NP-PP-PRED	She added the flowers to the bouquet/ I considered that problem of little concern	**24:** NP	He loved her	.3819
53/55: NP-TO-INF-OC / NP-TO-INF-VC	I advised Mary to go / They badgered him to go	**24:** NP	He loved her	.9735
74/3/4: PART / ADVP / ADVP-PRED-RS	She gave up / He meant well / He seems well	**76:** PART-NP / NP-PART	I looked up the entry/I looked the entry up	.1369
87/96: PP / PP-PRED-RS	They apologized to him/ The matter seems in dispute	**22:** INTRANS	I went	.2435
106/33/32: S-SUBJUNCT / NP-INF-OC / NP-INF	She demanded that he leave/ he helped her bake the cake/ he made her sing	**24:** NP	He loved her	.9815

Table 3: Alternation rules

3.2 New Filtering Method

After constructing the rules, a decision had to be made how to apply them to the system. The sentences in the data were tagged, lemmatized and parsed, patternsets were extracted and patterns classified. After this, the highest ranked classified pattern from a patternset was assigned a probability based on the binomial filter representing the probability that it is a correct class. However, no filtering was done using a confidence threshold. As a result, all the patterns go through the filter. Thus the probability assigned by the binomial hypothesis test is used as an estimate of a *class probability* i.e. a probability a verb is a member of a class. Our aim is to further correct this probability by making use of alternation rules. This is done using the following method.

Let $p(Cfc)$ be the class probability representing the confidence measure of a class, $p\,(rfc)$ the probability representing a revised confidence measure of a class, and $p(c|c')$ the probability of seeing class c given its alternating class c'. We assume that the following conditions hold:

a) $p(Cfc) < p\,(rfc) \qquad$ iff $\qquad p(c|c') > 0$

b) $p(Cfc) > p\,(rfc) \qquad$ iff $\qquad p(c) > 0 \quad$ and $\quad p(c') = 0$

These conditions indicate that if we have $p(c \rightarrow c')$ (i.e., we have seen both class c and its alternating class c'), we should improve the probability of c by the probability of the alternation rule. On the other hand, if we have seen c but not c' we should lower the probability of c by the probability of the alternation rule. I further assume that the following condition holds:

c) $p(c|c') > 0 \qquad$ iff $\qquad p(c') < 0.2 \quad$ and $\quad prel(c') > 0.45$

This condition states that even if we have $p(c \rightarrow c')$, the conditional implication of the rule counts only if the class probability, $p(c')$, and the class reliability, $prel(c')$, of the alternating class c' are high enough. Thus class probability must be less than 0.2 and class reliability more than 0.45[2]. The purpose of this condition is to consider the evidence value of the alternating class c'. This assures that alternation rules apply only if the alternating class is good evidence enough. Assuming the above conditions a) - c), the following algorithm is used to adjust the class probability:

$$\text{If} \quad p(c|c') > 0, \qquad p\,(rfc) \quad = p(Cfc) - w(p(Cfc)\ .\ p(c \longrightarrow c')) \tag{4}$$

$$\text{If} \quad p(c) > 0 \quad \text{and} \quad p(c') = 0, \qquad p\,(rfc) \quad = p(Cfc) + w(p(Cfc)\ .\ p(c \longrightarrow c')) \tag{5}$$

where w is a weighted sum defined empirically. For example, if we have the values $p(c) = 0.6$, $p(c') = 0.1$, $prel(c') = 0.5$, and the conditional implication is $p(c \longrightarrow c') = 0.4$, then the revised probability of c is:

$0.36 = 0.6 - 1(0.6\ .\ 0.4)$

After revising the class probabilities in the above way, the entries are filtered using a confidence of 95% (0.05) as before.

[2] In empirical tests, these values were found to give the best results.

3.3 Experimental Evaluation

The evaluation method used is the same as the one by Briscoe & Carroll (1997). They used the following scheme to indicate whether a class proposed for some verb is correct.

- *True positive* (TP): correct class proposed by the system
- *False positive* (FP): incorrect class proposed by the system
- *False negative* (FN): correct class not proposed by the system

Briscoe and Carroll calculated *precision* (percetage of correct subcategorization classes i.e. true positives to all classes found) and *recall* (percetage of correct classes found in the dictionary entry). They also estimated *accuracy* with which the system ranks true positive classes against the correct ranking of the seven verbs whose corpus input was manually analysed. This measure was computed by calculating the percetage of pairs of classes at positions $(n,\ m)$ s.t. $n < m$ in the system ranking that are ordered the same in the correct ranking. This gives an estimate of the accuracy of the relative frequencies of classes output by the system.

We introduced new test data to provide an adequate and reliable test for the method developed on the basis of the training data. The new test data covers sixteen test verbs, chosen at random but subject to the constraint that they exhibit multiple complementation patterns. The sentences containing the test verbs were tagged and parsed automatically, and the extractor, classifier and evaluator were applied to the resulting successful analyses. The results were evaluated against manual analysis of the corpus data. For each of the manually analysed verbs, 200-300 occurrences from the corpora were analysed. Manual analysis was also performed for the rest of the fourteen test verbs used in (Briscoe & Carroll, 1997).

The results achieved on the training data (the data covering the seven test verbs ranked by Briscoe & Carroll) are shown in Table 4. For comparison, the first row of the table shows the results from the original system run on the training data[3]. The second row shows the results of the improvement method with the nine alternation rules, and the third row with the extended set of rules. Table 5 gives the results achieved on the test data (the data covering 23 test verbs).

Method	Accuracy	Precision	Recall
Original system	53.95%	61.22%	44.70%
New filter + 9 rules	71.29%	63.13%	46.74%
New filter + extended rules	63.81%	69.42%	50.81%

Table 4: Results with the training data

Method	Accuracy	Precision	Recall
Original system	63.89%	55.72%	38.02%
New filter + 9 rules	63.32%	54.54%	38.53%
New filter + extended rules	67.70%	60.03%	43.28%

Table 5: Results with the test data

As the comparison of the results show, the performance on the test data is quite similar to that on the training data. Both the impairments and improvements seem to generalise. It is hardly surprising that the extended set of alternation rules achieves better results than the set of nine rules. Even though the nine rules developed concern frequent classes, these rules correct

[3] In these experiments, the data was extracted using a probabilistic chart parser, while Briscoe & Carroll (1997) used a probabilistic LR parser.

only some of the errors. With the test data, our method improved the original accuracy by 3.81%, precision by 4.31%, and recall by 5.26%.

The results gained on the training data generalised with the test data, and we can thus conclude that the improvements achieved are useful.

4 Conclusions and Further Work

The experiment reported above suggests that the technique proposed for improvement of the filtering component of the system was successful in improving the performance of the system. The proposed filtering method could still be improved. A more sophisticated approach could be taken with the extended set of alternation rules derived from the ANLT dictionary. Firstly, all of the 5468 alternation rules generated are currently applied to the patterns. However, it is possible that these rules contain some linguistically impossible alternations, and this affects the performance. We could try to find such incorrect rules, remove them, and examine whether this affects the results. To proceed with this, we would need to check manually the extended set of rules. Secondly, we also have to note that the alternation rules cover only the classes occurring in the ANLT dictionary. 32 of the classes are missing from the dictionary, and thus have no alternation probability.

However, this whole approach would need some improvements to make a full account of subcategorization phenomena. According to Briscoe & Carroll the system needs further refinement to narrow some subcategorization classes, for example, to choose between differing control options with predicative complements. It also needs supplementing with more accurate information about diathesis alternation possibilities (e.g. Levin, 1993) and semantic selection preferences on argument heads. In future work, we intend to extend the system in these directions.

References

Boguraev, B., Briscoe, E., Carroll, J., Carter, D. & Grover, C. 1987. The derivation of a grammatically-indexed lexicon from the Longman Dictionary of Contemporary English. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*, Stanford, CA. 193–200.

Brent, M. 1991. Automatic acquisition of subcategorization frames from untagged text. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, CA. 209–214.

Brent, M. 1993. From grammar to lexicon: unsupervised learning of lexical syntax. *Computational Linguistics* 19.3: 243–262.

Briscoe, E. & Carroll, J. 1993. Generalised probabilistic LR parsing for unification-based grammars. *Computational Linguistics* 19.1: 25–60.

Briscoe, E. & Carroll, J. 1994. *Parsing (with) punctuation.* Rank Xerox Research Centre, Grenoble, MLTT-TR-007.

Briscoe, E.J. and J. Carroll 1997. Automatic extraction of subcategorisation from corpora. In *Proceedings of the 5th ACL Conf. on Applied Nat. Lg. Proc.*, Washington, DC. 356–363.

Carroll, J. 1993. *Practical unification-based parsing of natural language.* Cambridge University Computer Laboratory, TR-224.

Carroll, J. 1994. Relating complexity to practical performance in parsing with wide-coverage unification grammars. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, NMSU, Las Cruces, NM. 287–294.

Carroll, J. & Briscoe, E. 1996. Apportioning development effort in a probabilistic LR parsing system through evaluation. In *Proceedings of the ACL SIGDAT Conference on Empirical Methods in Natural Language Processing*, University of Pensylvania, Philadelphia, PA. 92–100.

Cunningham, H., Gaizauskas, R. & Wilks, Y. 1995. *A general architecture for text engineering (GATE) - a new approach to language R&D.* Research memo CS-95-21, Department of Computer Science, University of Sheffield, UK.

Elworthy, D. 1994. Does Baum-Welch re-estimation help taggers?. In *Proceedings of the 4th Conf. Applied NLP*, Stuttgart, Germany.

Garside, R., Leech, G. & Sampson, G. 1987. *The computational analysis of English: A corpus-based approach.* Longman, London.

Grishman, R., Macleod, C. & Meyers, A. 1994. Comlex syntax: building a computational lexicon. In *Proceedings of the International Conference on Computational Linguistics, COLING-94*, Kyoto, Japan. 268–272.

Levin, B. 1993. *Towards a lexical organization of English verbs.* Chicago University Press, Chicago.

Manning, C. 1993. Automatic acquisition of a large subcategorisation dictionary from corpora. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Columbus, Ohio. 235–242.

Sampson, G. 1995. *English for the computer.* Oxford, UK: Oxford University Press.

Schabes, Y. 1992. Stochastic lexicalized tree adjoining grammars. In *Proceedings of the International Conference on Computational Linguistics, COLING-92*, Nantes, France. 426–432.

Taylor, L. & Knowles, G. 1988. *Manual of information to accompany the SEC corpus: the machine-readable corpus of spoken English.* University of Lancaster, UK, Ms.

Ushioda, A., Evans, D., Gibson, T. & Waibel, A. 1993. The automatic acquisition of frequencies of verb subcategorization frames from tagged corpora. In Boguraev, B. & Pustejovsky, J. eds. *SIGLEX ACL Workshop on the Acquisition of Lexical Knowledge from Text.* Columbus, Ohio: 95–106.