Can Subcategorisation Probabilities Help a Statistical Parser?

John Carroll and Guido Minnen
School of Cognitive and Computing Sciences
University of Sussex, Brighton, BN1 9QH, UK
{johnca,guidomi}@cogs.susx.ac.uk

Ted Briscoe
Computer Laboratory, University of Cambridge
Pembroke Street, Cambridge CB2 3QG, UK
ejb@cl.cam.ac.uk

Abstract

Research into the automatic acquisition of lexical information from corpora is starting to produce large-scale computational lexicons containing data on the relative frequencies of subcategorisation alternatives for individual verbal predicates. However, the empirical question of whether this type of frequency information can in practice improve the accuracy of a statistical parser has not yet been answered. In this paper we describe an experiment with a wide-coverage statistical grammar and parser for English and subcategorisation frequencies acquired from ten million words of text which shows that this information can significantly improve parse accuracy[1].

1 Introduction

Recent work on the automatic acquisition of lexical information from substantial amounts of machine-readable text (e.g. Briscoe & Carroll, 1997; Gahl, 1998; Carroll & Rooth, 1998) has opened up the possibility of producing large-scale computational lexicons containing data on the relative frequencies of subcategorisation alternatives for individual verbal predicates. However, although Resnik (1992), Schabes (1992), Carroll & Weir (1997) and others have proposed 'lexicalised' probabilistic grammars to improve the accuracy of parse ranking, no wide-coverage parser has yet been constructed which explicitly incorporates probabilities of different subcategorisation alternatives for individual predicates. It is therefore an open question whether this type of information can actually improve parser accuracy in practice.

In this paper we address this issue, describing an experiment with an existing wide-coverage statistical grammar and parser for English (Carroll & Briscoe, 1996) in conjunction with subcategorisation frequencies acquired from 10 million words of text from the British National Corpus (BNC; Leech, 1992). Our results show conclusively that this information can improve parse accuracy.

2 Background

2.1 Subcategorisation Acquisition

Several substantial machine-readable subcategorisation dictionaries exist for English, either built semi-automatically from machine-readable versions of conventional learners' dictionaries, or manually by (computational) linguists (e.g. the Alvey NL Tools (ANLT) dictionary, Boguraev *et al.* (1987); the COMLEX Syntax dictionary, Grishman, Macleod & Meyers (1994)). However, since these efforts were not carried out in tandem with rigorous large-scale classification of corpus data, none of the resources produced provide useful information on the relative frequency of different subcategorisation frames.

Systems which are able to acquire a small number of verbal subcategorisation classes automatically from corpus text have been described by Brent (1991, 1993), and Ushioda *et al.* (1993). Ushioda *et al.* also derive relative subcategorisation frequency information for individual predicates. In this work they utilise a part-of-speech (PoS) tagged corpus and finite-state NP parser to recognise and calculate

[1]This work was funded by UK EPSRC project GR/L53175 'PSET: Practical Simplification of English Text', CEC Telematics Applications Programme project LE1-2111 'SPARKLE: Shallow PARsing and Knowledge extraction for Language Engineering', and by an EPSRC Advanced Fellowship to the first author. Some of the work was carried out while the first author was a visitor at the Tanaka Laboratory, Department of Computer Science, Tokyo Institute of Technology, and at CSLI, Stanford University; the author wishes to thank researchers at these institutions for many stimulating conversations.

the relative frequency of six subcategorisation classes. They report that for 32 out of 33 verbs tested their system correctly predicts the most frequent class, and for 30 verbs it correctly predicts the second most frequent class, if there was one.

Manning (1993) reports a larger experiment, also using a PoS tagged corpus and a finite-state NP parser, attempting to recognise sixteen distinct complementation patterns—although not with relative frequencies. In a comparison between entries for 40 common verbs acquired from 4.1 million words of text and the entries given in the *Oxford Advanced Learner's Dictionary of Current English* (Hornby, 1989) Manning's system achieves a precision of 90% and a recall of 43%.

Gahl (1998) presents an extraction tool for use with the BNC that is able to create subcorpora containing different subcategorisation frames for verbs, nouns and adjectives, given the frames expected for each predicate. The tool is based on a set of regular expressions over PoS tags, lemmas, morphosyntactic tags and sentence boundaries, effectively performing the same function as a chunking parser (c.f. Abney, 1996). The resulting subcorpora can be used to determine the (relative) frequencies of the frames.

Carroll & Rooth (1998) use an iterative approach to estimate the distribution of subcategorisation frames given head words, starting from a manually-developed context-free grammar (of English). First, a probabilistic version of the grammar is trained from a text corpus using the expectation-maximisation (EM) algorithm, and the grammar is lexicalised on rule heads. The EM algorithm is then run again to calculate the expected frequencies of a head word accompanied by a particular frame. These probabilities can then be fed back into the grammar for the next iteration. Carroll & Rooth report encouraging results for three verbs based on applying the technique to text from the BNC.

Briscoe & Carroll (1997) describe a system capable of distinguishing 160 verbal subcategorisation classes—a superset of those found in the ANLT and COMLEX Syntax dictionaries—returning relative frequencies for each frame found for each verb. The classes also incorporate information about control of predicative arguments and alternations such as particle movement and extraposition. The approach uses a robust statistical parser which yields complete though 'shallow' parses, a comprehensive subcategorisation class classifier, and *a priori* estimates of the probability of membership of these classes. For a sample of seven verbs with multiple subcategorisation possibilities the system's frequency rankings averaged 81% correct. (We talk about this system further in section 3.2 below, describing how we used it to provide frequency data for our experiment).

2.2 Lexicalised Statistical Parsing

Carroll & Weir (1997)—without actually building a parsing system—address the issue of how frequency information can be associated with lexicalised grammar formalisms, using Lexicalized Tree Adjoining Grammar (Joshi & Schabes, 1991) as a unifying framework. They consider systematically a number of alternative probabilistic formulations, including those of Resnik (1992) and Schabes (1992) and implemented systems based on other underlying grammatical frameworks, evaluating their adequacy from both a theoretical and empirical perspective in terms of their ability to model particular distributions of data that occur in existing treebanks.

Magerman (1995), Collins (1996), Ratnaparkhi (1997), Charniak (1997) and others describe implemented systems with impressive accuracy on parsing unseen data from the Penn Treebank (Marcus, Santorini & Marcinkiewicz, 1993). These parsers model probabilistically the strengths of association between heads of phrases, and the configurations in which these lexical associations occur. The accuracies reported for these systems are substantially better than their (non-lexicalised) probabilistic context-free grammar analogues, demonstrating clearly the value of lexico-statistical information. However, since the grammatical descriptions are induced from atomic-labeled constituent structures in the training treebank, rather than coming from an explicit generative grammar, these systems do not make contact with traditional notions of argument structure (i.e. subcategorisation, selectional preferences of predicates for complements) in any direct sense. So although it is now possible to extract at least subcategorisation data from large corpora[2] with

[2]Grishman & Sterling (1992), Poznanski & Sanfilippo (1993), Resnik (1993), Ribas (1994), McCarthy (1997) and others have shown that it is possible also to ac-

some degree of reliability, it would be difficult to integrate the data into this type of parsing system.

Briscoe & Carroll (1997) present a small-scale experiment in which subcategorisation class frequency information for individual verbs was integrated into a robust statistical (non-lexicalised) parser. The experiment used a test corpus of 250 sentences, and used the standard GEIG bracket precision, recall and crossing measures (Grishman, Macleod & Sterling, 1992) for evaluation. While bracket precision and recall were virtually unchanged, the crossing bracket score for the lexicalised parser showed a 7% improvement. However, this difference turned out not to be statistically significant at the 95% level: some analyses got better while others got worse.

We have performed a similar, but much larger scale experiment, which we describe below. We used a larger test corpus, acquired data from an acquisition corpus an order of magnitude larger, and used a different quantitative evaluation measure that we argue is more sensitive to argument/adjunct and attachment distinctions. We summarise the main features of the 'baseline' parsing system in section 3.1, describe how we lexicalised it (section 3.2), present the results of the quantitative evaluation (section 3.3), give a qualitative analysis of the analysis errors made (section 3.4), and conclude with directions for future work.

3 The Experiment

3.1 The Baseline Parser

The baseline parsing system comprises:

- an HMM part-of-speech tagger (Elworthy, 1994), which produces either the single highest-ranked tag for each word, or multiple tags with associated forward-backward probabilities (which are used with a threshold to prune lexical ambiguity);
- a robust finite-state lemmatiser for English, an extended and enhanced version of the University of Sheffield GATE system morphological analyser (Cunningham *et al.*, 1995);
- a wide-coverage unification-based 'phrasal' grammar of English PoS tags and punctuation;
- a fast generalised LR parser using this grammar, taking the results of the tagger as input, and performing disambiguation using a probabilistic model similar to that of Briscoe & Carroll (1993); and
- training and test treebanks (of 4600 and 500 sentences respectively) derived semi-automatically from the SUSANNE corpus (Sampson, 1995);

The grammar consists of 455 phrase structure rule schemata in the format accepted by the parser (a syntactic variant of a Definite Clause Grammar with iterative (Kleene) operators). It is 'shallow' in that no attempt is made to fully analyse unbounded dependencies. However, the distinction between arguments and adjuncts is expressed, following X-bar theory, by Chomsky-adjunction to maximal projections of adjuncts ($XP \rightarrow XP\ Adjunct$) as opposed to 'government' of arguments (i.e. arguments are sisters within X1 projections; $X1 \rightarrow X0\ Arg1\ ...\ ArgN$). Furthermore, all analyses are rooted (in S) so the grammar assigns global, shallow and often 'spurious' analyses to many sentences. Currently, the coverage of this grammar—the proportion of sentences for which at least one analysis is found—is 79% when applied to the SUSANNE corpus, a 138K word treebanked and balanced subset of the Brown corpus.

Inui *et al.* (1997) have recently proposed a novel model for probabilistic LR parsing which they justify as theoretically more consistent and principled than the Briscoe & Carroll (1993) model. We use this new model since we have found that it indeed also improves disambiguation accuracy.

The 500-sentence test corpus consists only of in-coverage sentences, and contains a mix of written genres: news reportage (general and sports), *belles lettres*, biography, memoirs, and scientific writing. The mean sentence length is 19.3 words (including punctuation tokens).

3.2 Incorporating Acquired Subcategorisation Information

The test corpus contains a total of 485 distinct verb lemmas. We ran the Briscoe & Carroll (1997) subcategorisation acquisition system on the first 10 million words of the BNC, for each of these verbs saving the first 1000 cases in which a possible instance of a subcategorisation frame

quire selection preferences automatically from (partially) parsed data.

AP	NP_PP_PP	PP_WHPP	VPINF
NONE	NP_SCOMP	PP_WHS	VPING
NP	NP_WHPP	PP_WHVP	VPING_PP
NP_AP	PP	SCOMP	VPPRT
NP_NP	PP_AP	SINF	WHPP
NP_NP_SCOMP	PP_PP	SING	
NP_PP	PP_SCOMP	SING_PP	
NP_PPOF	PP_VPINF	VPBSE	

Table 1: *VSUBCAT* values in the grammar.

was identified. For each verb the acquisition system hypothesised a set of lexical entries corresponding to frames for which it found enough evidence. Over the complete set of verbs we ended up with a total of 5228 entries, each with an associated frequency normalised with respect to the total number of frames for all hypothesised entries for the particular verb.

In the experiment each acquired lexical entry was assigned a probability based on its normalised frequency, with smoothing—to allow for unseen events—using the (comparatively crude) *add-1* technique. We did not use the lexical entries themselves during parsing, since missing entries would have compromised coverage. Instead, we factored in their probabilities during parse ranking at the end of the parsing process.

We ranked complete derivations based on the product of (1) the (purely structural) derivation probability according to the probabilistic LR model, and (2) for each verb instance in the derivation the probability of the verbal lexical entry that would be used in the particular analysis context. The entry was located via the *VSUBCAT* value assigned to the verb in the analysis by the immediately dominating verbal phrase structure rule in the grammar: *VSUBCAT* values are also present in the lexical entries since they were acquired using the same grammar. Table 1 lists the *VSUBCAT* values. The values are mostly self-explanatory; however, examples of some of the less obvious ones are given in (1).

(1) *They made* (NP_WHPP) *a great fuss about what to do.*
They admitted (PP_SCOMP) *to the authorities that they had entered illegally.*
It dawned (PP_WHS) *on him what he should do.*

Some *VSUBCAT* values correspond to several of the 160 subcategorisation classes distinguished by the acquisition system. In these cases the sum of the probabilities of the corresponding entries was used. The finer distinctions stem from the use by the acquisition system of additional information about classes of specific prepositions, particles and other function words appearing within verbal frames. In this experiment we ignored these distinctions.

In taking the product of the derivation and subcategorisation probabilities we have lost some of the properties of a statistical language model. The product is no longer strictly a probability, although we do not attempt to use it as such: we use it merely to rank competing analyses. Better integration of these two sets of probabilities is an area which requires further investigation.

3.3 Quantitative Evaluation

3.3.1 Bracketing

We evaluated parser accuracy on the unseen test corpus with respect to the phrasal bracketing annotation standard described by Carroll *et al.* (1997) rather than the original SUSANNE bracketings, since the analyses assigned by the grammar and by the corpus differ for many constructions[3]. However, with the exception of SUSANNE 'verb groups' our annotation standard is bracket-consistent with the treebank analyses (i.e. no 'crossing brackets'). Table 2 shows the baseline accuracy of the parser with respect to (unlabelled) bracketings, and also with this model when augmented with the extracted subcategorisation information. Briefly, the evaluation metrics compare unlabelled bracketings derived from the test treebank with those derived from parses, computing *recall*, the ratio of matched brackets over all brackets in the treebank; *precision*, the ratio of matched brackets over all brackets found by the parser; *mean crossings*, the number of times a bracketed sequence output by the parser overlaps with one from the treebank but neither is properly contained in the other, averaged over all sentences;

[3]Our previous attempts to produce SUSANNE annotation scheme analyses were not entirely successful, since SUSANNE does not have an underlying grammar, or even a formal description of the possible bracketing configurations. Our evaluation results were often more sensitive to the exact mapping we used than to changes we made to the parsing system itself.

	Zero crossings (% sents.)	Mean crossings per sent.	Bracket recall (%)	Bracket precision (%)
'Baseline'	57.2	1.11	82.5	83.0
With subcat	56.6	1.10	83.1	83.1

Table 2: Bracketing evaluation measures, before and after incorporation of subcat information

and *zero crossings*, the percentage of sentences for which the analysis returned has zero crossings (see Grishman, Macleod & Sterling, 1992).

Since the test corpus contains only in-coverage sentences our results are relative to the 80% or so of sentences that can be parsed. In experiments measuring the coverage of our system (Carroll & Briscoe, 1996), we found that the mean length of failing sentences was little different to that of successfully parsed ones. We would therefore argue that the remaining 20% of sentences are not significantly more complex, and therefore our results are not skewed due to parse failures. Indeed, in these experiments a fair proportion of unsuccessfully parsed sentences were elliptical noun or prepositional phrases, fragments from dialogue and so forth, which we do not attempt to cover.

On these measures, there is no significant difference between the baseline and lexicalised versions of the parser. In particular, the mean crossing rates per sentence are almost identical. This is in spite of the fact that the two versions return different highest-ranked analyses for 30% of the sentences in the test corpus. The reason for the similarity in scores appears to be that the annotation scheme and evaluation measures are relatively insensitive to argument/adjunct and attachment distinctions. For example, in the sentence (2) from the test corpus

(2) *Salem (AP) - the statewide meeting of war mothers Tuesday in Salem will hear a greeting from Gov. Mark Hatfield.*

the phrasal analyses returned by the baseline and lexicalised parsers are, respectively (3a) and (3b).

(3) a ... *(VP will hear (NP a greeting) (PP from (NP Gov. Mark Hatfield))) ...*

b ... *(VP will hear (NP a greeting (PP from (NP Gov. Mark Hatfield)))) ...*

The latter is correct, but the former, incorrectly taking the PP to be an argument of the verb, is penalised only lightly by the evaluation measures: it has zero crossings, and 75% recall and precision. This type of annotation and evaluation scheme may be appropriate for a *phrasal* parser, such as the baseline version of the parser, which does not have the knowledge to resolve such ambiguities. Unfortunately, it masks differences between such a phrasal parser and one which can use lexical information to make informed decisions between complementation and modification possibilities[4].

3.3.2 Grammatical Relation

We therefore also evaluated the baseline and lexicalised parser against the 500 test sentences marked up in accordance with a second, grammatical relation-based (GR) annotation scheme (described in detail by Carroll, Briscoe & Sanfilippo, 1998).

In general, grammatical relations (GRs) are viewed as specifying the syntactic dependency which holds between a head and a dependent. The set of GRs form a hierarchy; the ones we are concerned with are shown in figure 1. *Subj*(ect) GRs divide into clausal (*xsubj/csubj*), and non-clausal (*ncsubj*) relations. *Comp*(lement) GRs divide into *clausal*, and into non-clausal direct object (*dobj*), second (non-clausal) complement in ditransitive constructions (*obj2*), and indirect object complement introduced by a preposition (*iobj*). In general the parser returns the most specific (leaf) relations in the GR hierarchy, except when it is unable to determine whether clausal subjects/objects are controlled from within or without (i.e. *csubj* vs. *xsubj*, and *ccomp* vs. *xcomp* respectively), in which case it

[4]Shortcomings of this combination of annotation and evaluation scheme have been noted previously by Lin (1996), Carpenter & Manning (1997) and others. Carroll, Briscoe & Sanfilippo (1998) summarise the various criticisms that have been made.

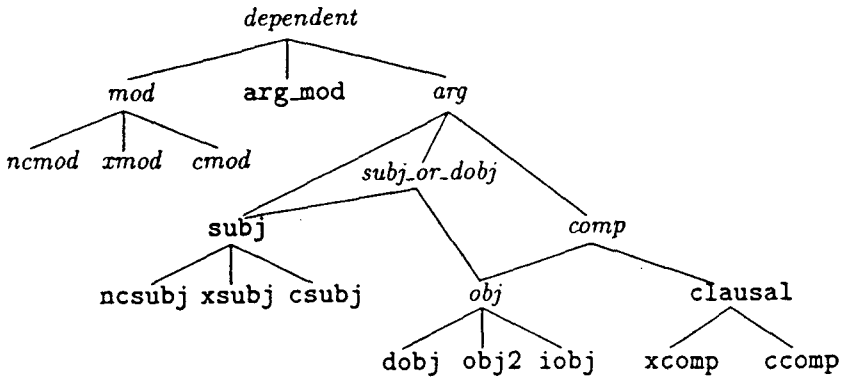

Figure 1: Portions of GR hierarchy used. (Relations in *italics* are not returned by the parser).

returns *subj* or *clausal* as appropriate. Each relation is parameterised with a head (lemma) and a dependent (lemma)—also optionally a type and/or specification of grammatical function. For example, the sentence (4a) would be marked up as in (4b).

(4) a *Paul intends to leave IBM.*

b *ncsubj (intend,Paul,_)*
xcomp (to,intend,leave)
ncsubj (leave,Paul,_)
dobj (leave,IBM,_)

Carroll, Briscoe & Sanfilippo (1998) justify this new evaluation annotation scheme and compare it with others (constituent- and dependency-based) that have been proposed in the literature.

The relatively large size of the test corpus has meant that to date we have in some cases not distinguished between *c/xsubj* and between *c/xcomp*, and we have not marked up modification relations; we thus report evaluation with respect to argument relations only (but including the relation *arg_mod*—a semantic argument which is syntactically realised as a modifier, such as the passive 'by-phrase'). The mean number of GRs per sentence in the test corpus is 4.15.

When computing matches between the GRs produced by the parser and those in the corpus annotation, we allow a single level of subsumption: a relation from the parser may be one level higher in the GR hierarchy than the actual correct relation. For example, if the parser returns *clausal*, this is taken to match both the more specific *xcomp* and *ccomp*. Also, an unspecified filler (_) for the type slot in the *iobj* and *clausal* relations successfully matches any actual specified filler. The head slot fillers are in all cases the base forms of single head words, so for example, 'multi-component' heads, such as the names of people, places or organisations are reduced to one word; thus the slot filler corresponding to *Mr. Bill Clinton* would be *Clinton*. For real-world applications this might not be the desired behaviour—one might instead want the token *Mr._Bill_Clinton*. This could be achieved by invoking a processing phase similar to the conventional 'named entity' identification task in information extraction.

Considering the previous example (2), but this time with respect to GRs, the sets returned by the baseline and lexicalised parsers are (5a) and (5b), respectively.

(5) a *ncsubj (hear,meeting,_)*
dobj (hear,greeting,_)
iobj (from,hear,Hatfield)

b *ncsubj (hear,meeting,_)*
dobj (hear,greeting,_)

The latter is correct, but the former, incorrectly taking the PP to be an argument of the verb, *hear*, is penalised more heavily than in the bracketing annotation and evaluation schemes: it gets only 67% recall. There is also no misleadingly low crossing score since there is no analogue to this in the GR scheme.

Table 3 gives the result of evaluating the baseline and lexicalised versions of the parser on the GR annotation. The measures compare the set of GRs in the annotated test corpus with those returned by the parser, in terms of *recall*, the percentage of GRs correctly found by the parser out of all those in the treebank; and *precision*,

	Recall (%)	Precision (%)
'Baseline'	88.6	79.2
With subcat	88.1	88.2

Table 3: GR evaluation measures, before and after incorporation of subcategorisation information. Argument relations only.

	Number
Complementation	124
Modification	134
Co-ordination	30
Textual	30
Misbracketing	40

Table 5: Numbers of errors of each type made by the lexicalised parser.

the percentage of GRs returned by the parser that are actually correct. In the evaluation, GR recall of the lexicalised parser drops by 0.5% compared with the baseline, while precision increases by 9.0%. The drop in recall is not statistically significant at the 95% level (*paired t-test*, 1.46, 499 *df*, $p > 0.1$), whereas the increase in precision is significant even at the 99.95% level (*paired t-test*, 5.14, 499 *df*, $p < 0.001$).

Table 4 gives the number of each type of GR returned by the two models, compared with the correct numbers in the test corpus. The baseline parser returns a mean of 4.65 relations per sentence, whereas the lexicalised parser returns only 4.15, the same as the test corpus. This is further, indirect evidence that the lexicalised probabilistic system models the data more accurately.

3.4 Discussion

In addition to the quantitative analysis of parser accuracy reported above, we have also performed a qualitative analysis of the errors made. We looked at each of the errors made by the lexicalised version of the parser on the 500-sentence test corpus, and categorised them into errors concerning: complementation, modification, co-ordination, structural attachment of textual adjuncts, and phrase-internal misbracketing. Of course, multiple errors within a given sentence may interact, in the sense that one error may so disrupt the structure of an analysis that it necessarily leads to one or more other errors being made. In all cases, though, we considered all of the errors and did not attempt to determine whether or not one of them was the 'root cause'. Table 5 summarises the number of errors of each type over the test corpus.

Typical examples of the five error types identified are:

complementation ... *decried the high rate of unemployment in the state* misanalysed as *decry* followed by an NP and a PP complement;

modification in ... *surveillance of the pricing practices of the concessionaires for the purpose of keeping the prices reasonable*, the PP modifier *for the purpose of* ... attached 'low' to *concessionaires* rather than 'high' to *surveillance*;

co-ordination the NP *priests, soldiers, and other members of the party* misanalysed as just two conjuncts, with the first conjunct containing the first two words in apposition;

textual in *But you want a job guaranteed when you return, I continued my attack*, the (textual) adjunct *I ... attack* attached to the VP *guaranteed ... return* rather than the S *But ... return*; and

misbracketing *Nowhere in Isfahan is this rich aesthetic life of the Persians ...* has *of* misanalysed as a particle, with *the Persians* becoming a separate NP.

There are no obvious trends within each type of error, although some particularly numerous sub-types can be identified. In 8 of the 30 cases of textual misanalysis, a sentential textual adjunct preceded by a comma was attached too low. The most common type of modification error was—in 20 of the 134 cases—misattachment of a PP modifier of $\overline{N}$ to a higher VP. The majority of the complementation errors were verbal, accounting for 115 of the total of 124. In 15 cases of incorrect verbal complementation a passive construction was incorrectly analysed as active, often with a following 'by' prepositional phrase erroneously taken to be a complement.

Other shortcomings of the system were evident in the treatment of co-ordinated verbal

	arg_mod	*ccomp*	*clausal*	*csubj*	*dobj*	*iobj*	*ncsubj*	*obj2*	*subj*	*xcomp*
'Baseline'	16	39	202	4	415	327	1054	53	14	202
With subcat	9	20	138	3	429	172	1058	39	15	195
Correct	32	16	136	2	428	160	1064	23	13	203

Table 4: Numbers of each type of grammatical relation.

heads, and of phrasal verbs. The grammatical relation extraction module is currently unable to return GRs in which the verbal head alone appears in the sentence as a conjunct—as in the VP *... to challenge and counter-challenge the authentication.* This can be remedied fairly easily. Phrasal verbs, such as *to consist of* are identified as such by the subcategorisation acquisition system. The grammar used by the shallow parser analyses phrasal verbs in two stages: firstly the verb itself and the following particle are combined to form a sub-constituent, and then phrasal complements are attached. The simple mapping from *VSUBCAT* values to subcategorisation classes cannot cope with the second level of embedding of phrasal verbs, so these verbs do not pick up any lexical information at parse time.

4 Conclusions

We surveyed recent work on automatic acquisition from corpora of subcategorisation and associated frequency information. We described an experiment with a wide-coverage statistical grammar and parser for English and subcategorisation frequencies acquired from 10 million words of text which shows that this information can significantly improve the accuracy of recovery of grammatical relation specifications from a test corpus of 500 sentences covering a number or different genres.

Future work will include: investigating more principled probabilistic models; addressing immediate lower-level shortcomings in the current system as discussed in section 3.4 above; adding *mod*(ification) GR annotations to the test corpus and extending the parser to also return these; and working on incorporating selectional preference information that we are acquiring in other, related work (McCarthy, 1997).

References

Abney, S. (1996). Partial parsing via finite-State cascades. *Natural Language Engineering*, 2(4), 337–344.

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*, 193–200. Stanford, CA.

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

Brent, M. (1993). From grammar to lexicon: unsupervised learning of lexical syntax. *Computational Linguistics*, 19(3), 243–262.

Briscoe, E. & Carroll, J. (1993). Generalized probabilistic LR parsing for unification-based grammars. *Computational Linguistics*, 19(1), 25–60.

Briscoe, E. & Carroll, J. (1997). Automatic extraction of subcategorization from corpora. In *Proceedings of the 5th ACL Conference on Applied Natural Language Processing*. Washington, DC.

Carpenter, B. & Manning, C. (1997). Probabilistic parsing using left corner language models. In *Proceedings of the 5th ACL/SIGPARSE International Workshop on Parsing Technologies*. MIT, Cambridge, MA.

Carroll, J. & Briscoe, E. (1996). Apportioning development effort in a probabilistic LR parsing system through evaluation. In *Proceedings of the 1st ACL/SIGDAT Conference on Empirical Methods in Natural Language Processing*, 92–100. University of Pennsylvania, Philadelphia, PA.

Carroll, J., Briscoe, E., Calzolari, N., Federici, S., Montemagni, S., Pirrelli, V., Grefenstette, G., Sanfilippo, A., Carroll, G. & Rooth, M. (1997). *SPARKLE WP1 specification of phrasal parsing.* <http://www.ilc.pi.cnr.it/sparkle.html>.

Carroll, J., Briscoe, E. & Sanfilippo, A. (1998). Parser evaluation: a survey and a new proposal. In *Proceedings of the 1st International Conference on Language Resources and Evaluation*, 447–454. Granada, Spain.

Carroll, J. & Weir, D. (1997). Encoding frequency information in lexicalized grammars. In *Proceedings of the 5th ACL/SIGPARSE International Workshop*

on Parsing Technologies (IWPT-97), 8–17. MIT, Cambridge, MA.

Carroll, G. & Rooth, M. (1998). Valence induction with a head-lexicalized PCFG. In *Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing.* Granada, Spain.

Charniak, E. (1997). Statistical parsing with a context-free grammar and word statistics. In *Proceedings of the 14th National Conference on Artificial Intelligence (AAAI-97)*, 598–603. Providence, RI.

Collins, M. (1996). A new statistical parser based on bigram lexical dependencies. In *Proceedings of the 34th Meeting of the Association for Computational Linguistics*, 184–191. Santa Cruz, CA.

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 ACL Conference on Applied Natural Language Processing.* Stuttgart, Germany.

Gahl, S. (1998). Automatic extraction of subcorpora based on subcategorization frames from a part-of-speech tagged corpus. In *Proceedings of the COLING-ACL'98.* Montreal, Canada.

Grishman, R., Macleod, C. & Meyers, A. (1994). Comlex syntax: building a computational lexicon. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING-94)*, 268–272. Kyoto, Japan.

Grishman, R., Macleod, C. & Sterling, J. (1992). Evaluating parsing strategies using standardized parse files. In *Proceedings of the 3rd ACL Conference on Applied Natural Language Processing*, 156–161. Trento, Italy.

Grishman, R. & Sterling, J. (1992). Acquisition of selectional patterns. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING-92)*, 658–664. Nantes, France.

Hornby, A. (1989). *Oxford Advanced Learner's Dictionary of Current English.* Oxford, UK: OUP.

Inui, K., Sornlertlamvanich, V., Tanaka, H. & Tokunaga, T. (1997). A new formalization of probabilistic GLR parsing. In *Proceedings of the 5th ACL/SIGPARSE International Workshop on Parsing Technologies (IWPT-97)*, 123–134. Cambridge, MA.

Joshi, A. & Schabes, Y. (1991). Tree-adjoining grammars and lexicalized grammars. In M. Nivat & A. Podelski (Eds.), *Definability and Recognizability of Sets of Trees.* Elsevier.

Leech, G. (1992). 100 million words of English: the British National Corpus. *Language Research*, 28(1), 1–13.

Lin, D. (1996). Dependency-based parser evaluation: a study with a software manual corpus. In R. Sutcliffe, H-D. Koch & A. McElligott (Eds.), *Industrial Parsing of Software Manuals*, 13–24. Amsterdam, The Netherlands: Rodopi.

Magerman, D. (1995). Statistical decision-tree models for parsing. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics.* Boston, MA.

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*, 235–242. Columbus, Ohio.

Marcus, M., Santorini, B. & Marcinkiewicz (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2), 313–330.

McCarthy, D. (1997). Word sense disambiguation for acquisition of selectional preferences. In *Proceedings of the ACL/EACL'97 Workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, 52–61. Madrid, Spain.

Poznanski, V. & Sanfilippo, A. (1993). Detecting dependencies between semantic verb subclasses and subcategorization frames in text corpora. In B. Boguraev & J. Pustejovsky (Eds.), *SIGLEX ACL Workshop on the Acquisition of Lexical Knowledge from Text.* Columbus, Ohio.

Ratnaparkhi, A. (1997). A linear observed time statistical parser based on maximum entropy models. In *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing.* Brown University, Providence, RI.

Resnik, P. (1992). Probabilistic tree-adjoining grammar as a framework for statistical natural language processing. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING-92)*, 418–424. Nantes, France.

Resnik, P. (1993). *Selection and information: a class-based approach to lexical relationships.* University of Pennsylvania, CIS Dept, PhD thesis.

Ribas, P. (1994). An experiment on learning appropriate selection restrictions from a parsed corpus. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING-94).* Kyoto, Japan.

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

Schabes, Y. (1992). Stochastic lexicalized tree-adjoining grammars. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING-92)*, 426–432. Nantes, France.

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