Integrating Probabilistic and Knowledge-based Approaches to Corpus Parsing

John Carroll
Computer Laboratory
University of Cambridge
Pembroke Street, Cambridge CB2 3QG
(*jac@cl.cam.ac.uk*)

Ted Briscoe
Rank Xerox Research Centre, Grenoble
6, chemin de Maupertius
38240 Meylan, France
(*Ted.Briscoe@xerox.fr*)

Abstract

We have developed a prototype system for syntactic parsing of corpus text based on a wide-coverage unification-based grammar of English and domain-independent statistical techniques for selecting the most plausible parses from the typically large number licensed by the grammar. Although the results from initial experiments are promising, the system is 'brittle', relying particularly on the correctness and completeness of lexical entries. We are currently concentrating on parsing large amounts of tagged text with a relatively simple, but robust, grammar of tag sequences and punctuation. This grammar produces coarse phrasal analyses of sentences from which possible complementation patterns can be extracted, allowing omissions in the set of lexical entries to be remedied[1].

1 The Probabilistic LR Parsing System

Briscoe & Carroll (1993) describe an approach to probabilistic parse selection using a large unification-based grammar of English. The grammar contains approximately 800 phrase structure rules written in the Alvey Natural Language Tools (ANLT) formalism (Briscoe *et al.* 1987), a syntactic variant of the Definite Clause Grammar formalism (Pereira & Warren 1980). The ANLT grammar has wide coverage and has been shown, for instance, to be capable of assigning a correct analysis to 96.8% of a corpus of 10,000 noun phrases extracted randomly from manually analysed corpora (Taylor, Grover & Briscoe 1989). The grammar is linked to a lexicon containing about 64,000 entries for 40,000 lexemes, including

[1] This research is supported in part by ESPRIT BRA 7315 'The Acquisition of Lexical Knowledge' (ACQUILEX II).

detailed subcategorisation information appropriate for the grammar, built semi-automatically from the *Longman Dictionary of Contemporary English* (LDOCE, Procter 1978).

1.1 The Probabilistic Model

The probabilistic parsing model developed by Briscoe & Carroll represents a refinement of probabilistic context-free grammar (PCFG). A maximally informative context-free 'backbone' is derived automatically from the ANLT grammar (in which all categories are represented as feature bundles). This backbone is used to construct a generalised, non-deterministic LR parser (e.g. Tomita 1987) based on a LALR(1) table. Unification of the 'residue' of features not incorporated into the backbone grammar is performed at parse time in conjunction with reduce operations. Unification failure results in the reduce operation being blocked and the associated derivation being assigned a probability of zero.

Probabilities are assigned to transitions in the LALR(1) action table via a process of supervised training based on computing the frequency with which transitions are traversed in a corpus of parse histories constructed using a user-driven, interactive version of the parser. The result is a probabilistic parser which, unlike a PCFG, is capable of probabilistically discriminating derivations which differ only in terms of order of application of the same set of CF backbone rules (within a context defined by the LALR(1) table) but which remains a stochastic first-order Markov model, because the LALR(1) table defines a non-deterministic finite-state machine (FSM) and the total probability of an analysis is computed from the sequence of transitions taken to construct it.

The parser is based on Kipps' (1989) LR recogniser (a re-formulation of Tomita's algorithm), generalised for the case of unification grammars; Carroll (1993) describes the parser in detail. The parser constructs a packed parse forest representation of the complete set of analyses licensed by the ANLT grammar for a given input. In this representation identical sub-analyses are shared between differing superordinate analyses (as in chart parsing and other tabular parsing techniques) and sub-analyses covering the same portion of input are packed if the subsumption relation defined on unification-based formalisms holds between their root categories.

Ideally, the computation of the most probable analysis or the *n*-best analyses defined by the probabilistic LR parser should not involve exhaustive search of the space of syntactically legitimate analyses defined by the ANLT grammar for any given input. However, it is not possible to introduce any Viterbi-style optimisation into the computation of local maximal paths through the probabilistic non-deterministic FSM defined by the parse table, because at any point in a derivation a maximal path may receive a probability of zero through unification failure, rendering a hitherto non-maximal local path maximal again. Unfortunately, the effects of feature propagation cannot be localised with respect to the computation of most probable sub-analyses, whilst any attempt to incorporate featural information into the probabilistic component of the grammar would result either in an intractably large grammar, or a model with too many free parameters, or both.

The full parse forest must therefore be constructed. Although this computation can be exponential in sentence length for some relatively unnatural grammars (Johnson 1989), in practice we have been able to create packed parse forests for sentences containing over 30 words having many thousands of analyses. In the packed parse forest the probabilities of sub-analyses are associated with each node in the forest and, in the case of packed nodes, a distinct probability is maintained for each distinct sub-analysis at that node. Carroll & Briscoe (1992) describe a practical Viterbi-like algorithm for unpacking the n-best analyses from this form of probabilistic parse forest, and a strategy for normalising partial derivations of differing lengths.

1.2 Empirical Results

Carroll (1993) describes a preliminary experiment using a subset of LDOCE noun definitions as the test corpus. A total of 246 definitions, selected without regard for their syntactic form, were parsed semi-automatically. One hundred and fifty were parsed successfully, the results of which were used to construct the probabilistic component of the system. Reparsing the training corpus and automatically comparing the most highly ranked analysis with the original parse, for the 89 definitions between two and ten words in length inclusive (mean length 6.2), in 68 cases (76%) the correct analysis (as defined by the training corpus) was also the most highly ranked. Reparsing the further set of 55 LDOCE noun definitions not drawn from the training corpus, each containing up to ten words (mean length 5.7), in 41 cases the correct parse was the most highly ranked, giving a correct parse / sentence measure of 75%.

These experiments suggest that this system is able to rank parses in a comparable fashion to systems based on PCFG (Fujisaki *et al.* 1989), probabilistic ID/LP CFG (Sharman, Jelinek & Mercer 1990) or simulated annealing (Sampson, Haigh & Atwell 1989), whose grammars are couched in a linguistically less adequate formalism and in two cases derived directly from manual analyses of the training and test corpus. The results are achieved solely on the basis of statistics concerning the conditional probability of syntactic rules in a syntactically-defined (LR) parse context, therefore a significant number of errors involve incorrect attachment of PPs, analyses of compounds, coordinations, and so forth, where lexical (semantic) information plays a major role. In many of these cases, the correct analysis is in the three highest ranked analyses. Both Sharman et al and Fujisaki *et al.* achieve slightly better results (about 85% correct parse / sentence), but their grammars integrate information concerning the probability of a lexeme occurring as a specific lexical syntactic category. Using a tree similarity measure, such as that of Sampson *et al.*, the most probable analyses achieve a better than 96% fit to the correct analyses (as opposed to 80% for Sampson *et al.*'s simulated annealing parser).

2 Improving Lexical Robustness

Although the preliminary results reported above are encouraging, the system is 'brittle': one major shortcoming is that it requires that the definitions of lexical entries be correct and complete. If this is not the case, then the system will either fail to produce any analysis at all, or will produce one that differs markedly from the correct one[2]. For example, Figure 1 shows the highest ranked analysis assigned to one definition in LDOCE for the noun *aid*. In this case, there is no lexical entry for *support* as an intransitive verb. Consequently, the parser finds, and ranks highest, an analysis in which *supports* and *helps* are treated as transitive verbs forming verb phrases with object noun phrase gaps, and *that supports or helps* as a zero relative clause with *that* analysed as a prenominal subject (compare *a person or thing that* **that** *supports or helps*). It is difficult to fault this analysis on syntactic grounds and the same is true for other 'false positives' of this type observed to date. Although the analyses assigned to such examples sometimes have low probabilities relative to the most probable correct analyses of other examples, this trend is not sufficiently pronounced, unfortunately, to allow false positives to be identified automatically.

2.1 Robust Phrasal Parsing

One solution to the problem of lexical coverage that we are currently investigating is to use a simpler type of grammar that does not require such detailed lexical information, and to acquire appropriate lexical entries automatically from large

[2] Although the lexicon was derived from the machine-readable version of a published dictionary, errors and omissions in the dictionary are reflected in the lexicon (Carroll & Grover 1989).

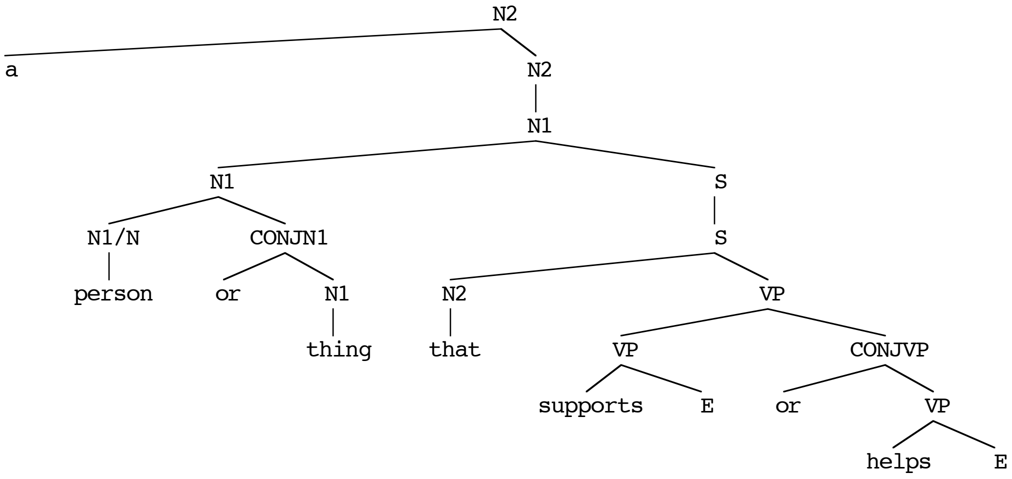

Figure 1: Parse Tree for *a person or thing that supports or helps.*

amounts of corpus data. The second author has developed such a grammar for English within the ANLT framework. The grammar contains around 250 phrase structure (PS) rules. As an indication of the level of analysis performed by this grammar, the following specific types of verb are distinguished:

- intransitive (possibly followed by a particle),
- transitive with a single NP object (possibly followed by a particle),
- taking one or two PP complements, optionally preceded by an NP direct object,
- taking a VP complement (base uninflected, infinitival, or past/present participle or gerund),
- taking an NP direct object followed by an infinitival or present participle VP complement,
- taking a sentential complement, optionally preceeded by an NP direct object (maybe with particle), and
- taking an AP object, optionally preceeded by an NP direct object.

For comparison, the full ANLT grammar contains 782 PS rules, and distinguishes approximately 60 different patterns of verb complementation.

Punctuation is an important source of grammatical constraint (Nunberg, 1990); for example, in the noun phrase *Western military, research and industrial computer systems*, the comma between *military* and *research* means that *military research* cannot be a compound noun in this context, whereas the possibility would have to be considered if the comma were ignored. Punctuation is invariably ignored in large syntactic grammars since it can appear in almost any position but is very often optional, making the grammar much harder to write, more verbose, and thus more difficult to maintain and extend. However, it is possible to treat punctuation in a principled and exhaustive manner within a simpler grammar of the type discussed here. The grammar contains two main types of category: syntactic and textual, the latter corresponding to the structure imposed on the text by punctuation. In the grammar, some commas are integrated with the analysis of specific constructions (e.g. noun compounds or coordination), but the rest of punctuation is just 'folded in' by treating text categories and syntactic categories as overlapping and dealing with the properties of each in terms of disjoint sets of unification grammar features; the two sets of features behave differently with respect to the way their values are passed around analysis trees.

The grammar resolves attachment ambiguities (e.g. when there are multiple prepositional phrases) in favour of a single canonical structure, thus producing relatively coarse phrasal analyses. Moreover, several syntactic distinctions that are necessary for principled semantic processing are finessed (so the grammar could not be used on its own in the syntactic stage of a complete NL analysis system). Indeed, our attempts to use the grammar to drive even a simplified treatment of NL semantics have failed, mainly due to the fact that textual (punctuation) units do not correspond directly to conventional syntactic constituents.

Instead of explicitly acquiring appropriate lexical entries from corpus data, the grammar is written to analyse sequences of part-of-speech tags[3]. Parsing consists of taking as input the results produced by a lexical tagger, throwing away the words, and parsing just the tags. The lexical stage is therefore exactly as robust as the tagger, and there is no need to construct any sort of lexicon; in effect the grammar uses the tagger's internal lexicon. Since the tagger disambiguates the set of possible tags for each word to the single highest ranking alternative (given the surrounding words), lexical ambiguity has been eliminated, making parsing more efficient.

Using an optimised non-deterministic chart-like parser (the Alvey NL Tools parser; Carroll, 1993) in conjunction with the grammar, we have parsed the com-

[3] The grammar is currently set up for the CLAWS-2 tagset (Garside, 1987; Taylor & Knowles, 1988).

plete Spoken English Corpus (SEC; Taylor & Knowles, 1988) producing parses for almost exactly half of the 2229 sentences. The maximum number of parses for a single sentence is 4500; 47 sentences are assigned 1000 or more analyses, 160 sentences 100–999 analyses, 404 sentences 10–99 analyses, and 377 sentences receive between one and nine analyses. Typical throughput is around fifty words per second on a Hewlett Packard PA-RISC workstation.

2.2 Inferring Complementation Patterns

We are currently concentrating on extracting complementation patterns for particular, common, verbs (e.g. *expect*, *swing* etc.) from the results of parsing sentences taken from corpora. These patterns will be used to:

- correct the ANLT lexicon and augment it with relative probabilities of lexical entries' various complementation patterns, and to

- help lexicographers detect collocations to aid the process of constructing of accurate printed dictionaries.

For example, the following sentence (taken from the Suzanne corpus):

> there_EX was_VBDZ no_AT debate_NN1 as_CSA the_AT senate_NNJ1 passed_VVD the_AT bill_NN1 on_RP to_II the_AT house_NNL1.

is assigned two parses (differing only in the point at which the final prepositional phrase is attached), from which the (verbal) patterns

> passed_VVD <Noun Phrase> on_RP <Prep Phrase>
> was_VBDZ <Noun Phrase>

are extracted.

We are also applying the same probabilistic techniques described in the first part of this paper to this grammar in order to rule out implausible analyses. To minimise the ill-effects of mistagging—a tagger will on average tag 5% of ambiguous words incorrectly—multiple alternative word-tag hypotheses will be fed into the parsing system (the hypotheses filtered to retain only those within a given threshold of the highest ranked alternative); the parser will use the probabilities assigned to each word-tag pair to carry forward the tagger's ordering of the alternatives.

A tag grammar for French based on the same principles as the one described above for English is in development at Rank Xerox Grenoble. The same techniques will be experimented with and the results compared.

3 Further Related Work

If the punctuation portion of the combined grammar of punctuation and tag sequences is removed, the resulting grammar parses a larger proportion of sentences in the SEC successfully. A standalone version of just the punctuation portion shows an even more dramatic improvement, covering all but 66 sentences (and assigning a single analysis to more than half). This suggests that overall coverage could be improved by separating the two grammars into two loosely interacting parsing stages, the punctuation grammar returning the set of possible bracketings imposed by punctuation on the input, and the tag grammar constrained not to return complete parses containing constituents which violate this structure. This approach might also allow a simplified treatment of semantics to be used with the tag grammar (whereas before the treatment of punctuation in it prevented this).

We will also be investigating using a module based around the tag and punctuation grammar as a front-end to constrain the full ANLT grammar. The module will identify phrase boundaries and provide less detailed fall-back analyses for input fragments outside the coverage of the full grammar.

References

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

Briscoe, E., Grover, C., Boguraev, B. and Carroll, J. (1987) A formalism and environment for the development of a large grammar of English. In *Proceedings of the 10th International Joint Conference on Artificial Intelligence*, 703–708. Milan, Italy.

Briscoe, E. and Waegner, N. (1992) Robust stochastic parsing using the inside-outside algorithm. In *Proceedings of the AAAI Workshop on Statistically-based Techniques in Natural Language Processing.* San Jose, California.

Carroll, J. (1993) *Practical unification-based parsing of natural language.* Cambridge University, Computer Laboratory, TR-314.

Carroll, J. and Briscoe, E. (1992) Probabilistic normalisation and unpacking of packed parse forests for unification-based grammars. In *Proceedings of the AAAI Fall Symposium on Probabilistic Approaches to Natural Language*, 33–38. Cambridge, MA.

Carroll, J. and Grover, C. (1989) The derivation of a large computational lexicon for English from LDOCE. In Boguraev, B. and Briscoe, E. eds. *Computational Lexicography for Natural Language Processing*. Longman, London: 117–134.

Fujisaki, T., Jelinek, F., Cocke, J., Black, E. and Nishino, T. (1989) A probabilistic method for sentence disambiguation. In *Proceedings of the 1st International Workshop on Parsing Technologies*, 105-114. Carnegie-Mellon University, Pittsburgh.

Garside, R. (1987) The CLAWS word-tagging system. In R. Garside, G. Leech and G. Sampson eds. *The Computational Analysis of English: A Corpus-based Approach*. London, UK: Longman: 30–41.

Johnson, M. (1989) The computational complexity of Tomita's algorithm. In *Proceedings of the 1st International Workshop on Parsing Technologies*, 203–208. Carnegie-Mellon University, Pittsburgh.

Kipps, J. (1989) Analysis of Tomita's algorithm for general context-free parsing. In *Proceedings of the 1st International Workshop on Parsing Technologies*, 193–202. Carnegie-Mellon University, Pittsburgh.

Nunberg, G. (1990) *The linguistics of punctuation*. CSLI Lecture Notes 18, Stanford, CA.

Pereira, F. and Warren, D. (1980) Definite clause grammars for language analysis – a survey of the formalism and a comparison with augmented transition networks. *Artificial Intelligence* 13.3: 231–278.

Procter, P. ed. (1978) *The Longman dictionary of contemporary English*. Longman, London.

Sampson, G., Haigh, R. and Atwell, E. (1989) Natural language analysis by stochastic optimization: a progress report on Project APRIL. *Journal of Experimental and Theoretical Artificial Intelligence* 1: 271–287.

Sharman, R., Jelinek, F. and Mercer, R. (1990) Generating a grammar for statistical training. In *Proceedings of the DARPA Speech and Natural Language Workshop*, 267–274. Hidden Valley, Pennsylvania.

Taylor, L., Grover, C. and Briscoe, E. (1989) The syntactic regularity of English noun phrases. In *Proceedings of the 4th European Meeting of the Association for Computational Linguistics*, 256–263. UMIST, Manchester.

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

Tomita, M. (1987) An efficient augmented-context-free parsing algorithm. *Computational Linguistics* 13.1: 31–46.