ROBUST PARSING — A BRIEF OVERVIEW

John Carroll
Cognitive and Computing Sciences
University of Sussex
Brighton BN1 9QH, UK
john.carroll@cogs.susx.ac.uk

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

Parsing systems able to analyse natural language text robustly and accurately at an appropriate level of detail would be of great value in computer applications ranging from speech synthesis and document style checking to message understanding and automatic translation. A number of research groups worldwide are currently developing such systems, varying in the depth of analysis from lexical parsing or tagging (identifying syntactic features just of individual words), through shallow or phrasal parsing (finding phrases, e.g. NPs, or forming hierarchical syntactic structure but not exploiting subcategorisation), to full parsers (which deal with unbounded dependencies etc., and are able to recover predicate-argument structure).

However, despite over three decades of research effort, no practical domain-independent parser of unrestricted text has been developed. Such a parser should return the correct or a useful 'close' analysis for 90% or more of input sentences. It would need to solve at least the following three problems, which create severe difficulties for conventional parsers utilising standard parsing algorithms with a generative grammar (for general background see Gazdar & Mellish, 1989): appropriate segmentation of text into syntactically parseable units; disambiguation, that is, selecting the unique semantically and pragmatically correct analysis from the potentially large number of syntactically legitimate ones returned; and undergeneration, or dealing with cases of input outside the systems' lexical or syntactic coverage. Conventional parsers typically fail to return any useful information when faced with problems of undergeneration or segmentation, and rely on domain-specific detailed semantic information for disambiguation.

The problem of text segmentation is best exemplified by sentences (beginning with a capital letter and ending with a period) which—and this sentence is an example—contain text adjuncts delimited by dashes, brackets or commas which may not always stand in a *syntactic* relation with surrounding material. There has been very limited work on this issue—Hindle (1983a) describes a system which copes with related problems, such as false starts and 'restarts' in transcribed spontaneous speech, whilst Jones (1994) and Briscoe & Carroll (1995) describe parsers which make limited use of punctuation to constrain syntactic interpretation. Nevertheless, for exmple, out of the 150K word, balanced Susanne Corpus (Sampson, 1995), over 60% of sentences contain internal punctuation marks and of these around 30% contain text-medial adjuncts. Thus the problem is significant, and further research is required building on linguistic accounts of punctuation (Nunberg, 1990).

Disambiguation using knowledge-based techniques requires the specification of too much detailed semantic information to yield a robust domain-independent parser. Yet analysis of the Susanne Corpus with a crude parser suggests that over 80% of sentences are structurally ambiguous. Several parsers yielding a single 'canonical' parse have been developed (Marcus, 1980; Hindle, 1983b; de Marcken, 1990). These are often applied to a (partially) disambiguated sequence of lexical syntactic categories. Simplifying the input to the parser in this way circumvents many

problems of lexical coverage suffered by systems which require rich sets of syntactic subcategories encoding for example valency of verbs (Jensen, 1991) as well as capitalising on the relative success and practicality of lexical category disambiguation (e.g. Garside *et al.*, 1987; DeRose, 1988). Canonical parsers often represent many ambiguities implicitly (Marcus *et al.*, 1983), rather than enumerating possible analyses, and use heuristic disambiguation rules (Hindle, 1989). Such techniques have yielded useful parsers for limited domains but their development is labour intensive and few general principles for their construction have emerged. In recent efforts to construct large 'treebanks' of parsed texts, canonical parsing has been used as a first but small step (Marcus *et al.*, 1993; Leech & Garside, 1991). More limited phrasal canonical parsing, such as systems for phrase spotting which rely on selecting the most likely boundaries of specific phrases (e.g. NPs) using finite-state grammars augmented with probabilities or heuristics (e.g. longest match) have been used in limited parsing tasks such as identifying the contexts of grammatical realisation of predicates for the construction of subcategorisation dictionaries (Manning, 1993; Ushioda *et al.*, 1993)

The availability of treebanks and, more generally, large bodies of machine-readable textual data has provided impetus to statistical approaches to disambiguation. Some approaches use stochastic language modelling inspired by the success of HMM-based lexical category disambiguation. For example, probabilities for a probabilistic version of a context-free grammar (PCFG) can be (re-)estimated from treebanks or plain text (Fujisaki *et al.*, 1989; Sharman *et al.*, 1990; Schabes *et al.*, 1993; Charniak, 1996) and used to rank analyses produced by minimally-modified tabular parsing algorithms (see Charniak, 1993). These techniques yielded promising results but have been largely supplanted by statistical parse decision techniques in which the probabilistic model is sensitive to details of parse context (Magerman & Weir, 1992; Briscoe & Carroll, 1993; Brill, 1993; Magerman, 1995; Collins, 1996) and integrated more closely with the parsing algorithm than with the grammar. These systems have yielded results of up to around 85% 'near correctness' of analyses assigned to (unseen) test sentences from the same source as the unambiguous training material. The barrier to improvement of such results currently lies in the need to use more discriminating models of context, requiring more annotated training material to adequately estimate the parameters of such models.

Models of context can be extended to encompass the whole treebank, the grammar consisting of all subtrees of all depths in it (Bod, 1993). Although returning impressive levels of accuracy, the inefficiency of the parsing algorithms required restricts the treebank size and complexity. This approach, and those of Magerman (1995) and Collins (1996) construct a grammar fully automatically and produce analyses that are patterned on those in the treebank. However, as a side-effect the text phenomena that can be parsed are necessarily limited to those present in the training material, and being able to deal with new texts would normally entail further substantial treebanking efforts, and possibly also major improvements in the efficiency of storage and deployment of derived syntactic knowledge. Other approaches relying on hand-constructed generative grammars (e.g. Magerman & Weir, 1992; Briscoe & Carroll, 1993) are not limited to phenomena that occur in the treebank, but the grammars utilised can be labour-intensive to develop and inevitably suffer from undergeneration. Shallow or phrasal parsers offer a partial solution to the former problem (Carroll & Briscoe, 1996), and Constraint Grammar (Karlsson *et al.*, 1995) is an attractive alternative that is not subject to the latter drawback.

Undergeneration is a significant problem: in one project, a grammar developed over 3 years for sentences from computer manuals containing words drawn from a restricted vocabulary of 3000 words still failed to analyse 4% of unseen examples (Black *et al.*, 1993). This probably represents an upper bound using manual development of generative grammars; most more general grammars have far higher failure rates on this type of text. Early work on undergeneration focussed on knowledge-based manual specification of 'error' rules or rule relaxation strategies (Kwasny &

Sondheimer, 1981; Jensen *et al.*, 1983). This approach, similar to the canonical parse approach to ambiguity, is labour-intensive and suffers from the difficulty of predicting the types of error or extragrammaticality liable to occur. More recently, attempts have been made to use statistical induction to 'learn' the correct grammar for a given (unanalysed) corpus of data, using generalisations of HMM maximum-likelihood re-estimation techniques to PCFGs (Lari & Young, 1990). This extends the application of stochastic language modelling from disambiguation to undergeneration by assuming the 'weakest' grammar for a given category set—that is, the one which contains all possible rules that can be formed for that category set—and using iterative re-estimation of the rule probabilities to converge on the subset of these rules most appropriate to the description of the training corpus.

There are several inherent problems with these statistical techniques which have been partially addressed by recent work. Re-estimation involves considering all possible analyses of each sentence of the training corpus given an (initially) weak grammar, the search space is large and the likelihood of convergence on a useful grammar conforming to any plausible linguistic constraints is low. Pereira & Schabes (1992) and Schabes *et al.* (1993) show that constraining the analyses considered during re-estimation to those consistent with manual parses of a treebank reduces computational complexity and leads to a useful grammar. Briscoe & Waegner (1993) and Briscoe (1994) demonstrate that similar results can be obtained by imposing general linguistic constraints on the initial grammar and biasing initial probabilities to favour linguistically-motivated 'core' rules, whilst still training on plain text. Nevertheless, such techniques are currently limited to simple grammars with category sets of a dozen or so non-terminals or to training on manually parsed data. The induced PCFG can also be used to rank parses, and results of around 80% 'fit' between correct and automatically-generated analyses have been obtained.

It is often difficult to compare reported results for different parsing systems. A wide variety of corpora are in use, ranging in processing difficulty from comparatively simple limited-vocabulary homogeneous texts to unrestricted texts drawn from a variety of different sources (e.g. newspapers, transcribed speech, rules and regulations, novels, etc.) intended to represent the full diversity of a language. Even considering just a single specific corpus, the level of data preparation and method of partitioning it into training and test data sets make inter-system comparison problematic (Goodman, 1996). Several schemes for evaluating accuracy have been used, including exact match, correctness of rule application, and, currently the most popular, the PARSEVAL scheme (Grishman *et al.*, 1992) which measures (labelled) bracketing accuracy and bracketing consistency for parser analyses with respect to a hand-annotated 'standard'. Even this scheme has problems, for instance when evaluating systems incorporating hand-built grammars against existing treebanks which often employ different analysis conventions.

For these reasons, and for the purposes of this workshop, we prepared a small test corpus of English sentences and sent it to all authors of papers in this workshop. The corpus contains thirty sentences selected randomly from some machine-readable text from an Indian English-language newspaper[1] (listed in the appendix). We anticipate that the test corpus will not have been used in the development of any of the systems. We have encouraged authors whose systems parse languages other then English to translate some or all of the sentences into their language. We asked all authors to run the sentences through their systems and then—either in their papers or verbally in their workshop presentations—to:

> "work out answers for as many of the following (simple evaluatory) questions as are applicable:
>
> 1. What proportion of the sentences were processed completely, and what level of representation were you able to construct (e.g. only partial parses for 40% of sentences,

[1]This text was given to us recently by Raman Chandrasekar (Mickey) of the University of Pennsylvania, whom we thank.

full dependency structure for 60%, etc.)

2. What was the mean processing time per word?

3. What does the parser output look like for a typical sentence?

4. What were the main types of mistake the parser made?

5. Can you identify the major causes of these misanalyses?

6. How could you improve your system to deal better with this test set, and/or how easy would it be to tune your system for text of this type?"

We hope that the common test data and evaluation criteria will facilitate comparison between systems, and that the results reported will serve to stimulate focussed discussion on the strengths and weaknesses of the diverse set of approaches currently being investigated, and to discuss areas that require further work.

References

Black, E., Lafferty, J. & Roukos, S. 1992. Development and evaluation of a broad-coverage probabilistic grammar of English-language computer manuals. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, Newark, Delaware. 185–192.

Bod, R. 1993. Using an annotated corpus as a stochastic grammar. In *Proceedings of the 6th Conference of the European Chapter of the Association for Computational Linguistics*, Utrecht, The Netherlands. 37–44.

Brill, E. 1993. Transformation-based error-driven parsing. In *Proceedings of the 3rd International Workshop on Parsing Technologies*, Tilburg, The Netherlands. 13–25.

Briscoe, E. 1994. Prospects for practical parsing: robust statistical techniques. In P. de Haan & N. Oostdijk eds. *Corpus-based Research into Language: A Feschrift for Jan Aarts*. Rodopi, Amsterdam: 67–95.

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

Briscoe, E. & Carroll, J. 1995. Developing and evaluating a probabilistic LR parser of part-of-speech and punctuation labels. In *Proceedings of the 4th ACL/SIGPARSE International Workshop on Parsing Technologies*, Prague, Czech Republic. 48–58.

Briscoe E. & Waegner, N. 1993. Undergeneration and robust parsing. In Meijs, W. eds. *Corpus-based Linguistics: Proceedings of the 6th ICAME Conference*. Amsterdam: Rodopi.

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.

Charniak, E. 1993. *Statistical language learning*. Cambridge, MA: MIT Press.

Charniak, E. 1996. *Tree-bank grammars*. Brown University, Department of Computer Science, report CS-96-02.

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

de Marcken, C. 1990. Parsing the LOB corpus. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, New York. 243–251.

DeRose, S. 1988. Grammatical category disambiguation by statistical optimization. *Computational Linguistics* 14.1: 31–39.

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

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

Gazdar, G. & Mellish, C. 1989. *Natural language processing in Lisp*. Addison Wesley.

Goodman, J. 1996. Efficient algorithms for parsing the DOP model. In *Proceedings of the ACL SIGDAT Conference on Empirical Methods in Natural Language Processing*, University of Pensylvania, Philadelphia, PA. 143–152.

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*, Trento, Italy. 156–161.

Hindle, D. 1983a. Deterministic parsing of syntactic nonfluencies. In *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics*, . 123–128.

Hindle, D. 1983b. *User manual for Fidditch, a deterministic parser*. Naval Research Laboratory Technical Memorandum 7590-142.

Hindle, D. 1989. Acquiring disambiguation rules from text. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, Vancouver, Canada. 118–125.

Jensen, K. 1991. A broad-coverage natural language analysis system. In Tomita, M. eds. *Current Issues in Parsing Technology*. Dordrecht: Kluwer.

Jensen, K., Heidorn, G., Miller, L. & Ravin, Y. 1983. Parse fitting and prose fixing: getting a hold on ill-formedness. *Computational Linguistics* 9: 147–153.

Jones, B. 1994. Can punctuation help parsing?. In *Proceedings of the International Conference on Computational Linguistics, COLING-94*, Kyoto, Japan. 421–425.

Karlsson, F., Voutilainen, A., Heikkilä, J. & Anttila, A. 1995. *Constraint grammar. A language-independent system for parsing unrestricted text*. Berlin / New York: Mouton de Gruyter.

Kwasny, S. & Sondheimer, N. 1981. Relaxation techniques for parsing ill-formed input. *American Journal of Computational Linguistics* 7.2: 99–108.

Lari, K. & Young, S. 1990. The estimation of stochastic context-free grammars using the inside-outside algorithm. *Computer Speech and Language Processing* 4: 35–56.

Leech, G. & Garside, R. 1991. Running a grammar factory: the production of syntactically analysed corpora or 'treebanks'. In Johansson, S. & Stenstrom, A. eds. *English Computer Corpora: Selected Papers and Bibliography*. Berlin: Mouton de Gruyter.

Magerman, D. 1995. Statistical decision-tree models for parsing. In *Proceedings of the 33rd Meeting of the Association for Computational Linguistics*, Boston, MA. 276–283.

Magerman, D. & Weir, C. 1992. Efficiency, robustness and accuracy in Picky chart parsing. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, Newark, Delaware. 40–47.

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.

Marcus, M. 1980. *A theory of syntactic recognition for natural language*. Cambridge, MA: MIT Press.

Marcus, M., Hindle, D. & Fleck, M. 1983. D-theory: talking about talking about trees. In *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics*, Cambridge, MA.

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

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

Pereira, F. & Schabes, Y. 1992. Inside-outside re-estimation for partially bracketed corpora. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, Newark, Delaware. 128–135.

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

Schabes, Y., Roth, M. & Osborne, R. 1993. Parsing the Wall Street Journal with the inside-outside algorithm. In *Proceedings of the 6th Conference of the European Chapter of the Association for Computational Linguistics*, Utrecht, The Netherlands. 341–346.

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

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

Appendix

The latest round of talks in the series which began on July 26 is slated to take place tomorrow.

At least 105 people were killed this morning in east Sri Lanka in the biggest massacre this year by the Liberation Tigers of Tamil Eelam (LTTE).

The Tamil Tigers raided a predominantly Muslim cluster of four villages in Polonnaruwa district, near the border with the eastern province district of Trincomalee, around 4.45 am.

The joint rescue operation launched by the Indian Air Force and paracommandos came to a successful conclusion at 10 am today when Major Ivan Crasto was winched up by a M-17 helicopter from a stranded cable car 42 hours after tragedy struck.

These will be supplied to the VHP and the AIBMAC the following day and, by October 29, both the sides will submit their replies to the government, the official spokesman told newspersons.

The government will announce the date of talks after receiving the replies, the spokesman clarified.

The Guatemalan Indian leader and human rights campaigner, Ms Rigoberta Menchu, won the 1992 Nobel Peace Prize today, the Norwegian Nobel committee said, reports Reuter.

They said the killers on Saturday night opened fire with machine guns on Mohammad Houedi, a local commander of PLO chief Yasser Arafat's mainstream Fatah group, in the busy Palestinian refugee camp east of Tyre.

The President, Mr Yang Shangkun, the National People's Congress chairman, Mr Wan Li, and the Vice Premiers, Mr Yao Yilin and Mr Wu Xueqian, fail to figure in the Central Committee list passed unanimously by the 14th Party Congress that ended its week long session today.

He must come out with a definite statement, said the VHP leader.

"Ganga gaye to Gangadas, Jamuna gaye to Jamunadas" he yields both ways depending on the situation, quipped Mr Singhal.

A fierce anti nuclear activist, Ms Kelly who once said that the only answer to the arms race was to go back to the principles of Mahatma Gandhi, and Mr Bastian, a former army General who turned pacifist, had died either in a case of suicide or murder, investigation authorities said.

It was Ms Kelly who had brought Mr Bastian, a former member of NATO's nuclear planning group, to the Greens.

The list of demands was later referred to the labour commissioner here.

India and Nepal today formalised a series of measures to expand bilateral cooperation, enhance Nepalese exports to India on liberalised terms and harness the immense water resource potential for the mutual benefit of the two countries, PTI reports.

The government's subsequent climb down and the new plan which would phase out closure did not satisfy at least nine Tory peers who voted for a Labour motion.

Henna, directed by Randhir Kapoor, was given the Raj Kapoor special award by the festival organisers.

Giving details, the city senior superintendent of police, Mr Hardeep Dhillon, said Mr Gargach and one Waryam Singh had a long drinking bout and then had a wordy duel over some payments.

In panic, Mr Gargach ran to the roof top of his house while Waryam Singh tried to escape from the main gate.

The security guard has since been arrested.

Mr Dhillon said Waryam Singh was a frequent visitor to Mr Gargach's house.

The SSP did not rule out the possibility of Waryam Singh having links with militants.

Indian Airlines has cancelled the leave sanctioned to its employees and recalled them to duties at all stations on the network except in most exceptional cases, as the agitation by the Indian Commercial Pilots Association (ICPA) entered the eleventh day on Friday without a solution in sight, reports PTI.

The three new polls also show independent, Mr Ross Perot, nearly doubling his support since the three Presidential debates earlier this month.

That's all we can broadcast at this hour, a radio announcer said at 2.43 pm (12.43 pm GMT).

The third round of talks between the Indian Airline pilots and the management remained inconclusive on Wednesday with both sides adhering to their stands on the issue of safety.

The talks will continue on Thursday before the deputy chief labour commissioner (central) here, reports TOINS.

Top religious leaders of various Hindu sects and communities have congregated in the capital for the two day event.

Deposing before the Joint Parliamentary Committee, Mr Vyas said if RBI had objected to any of those transactions, he would have come to know of the irregularity and taken action.

In a major political development, the state tribal development minister, Mr Karam Chand Bhagat, today formally submitted his resignation from the ministry to the Chief Minister, Mr Laloo Prasad Yadav, to press for the creation of a separate Jharkhand state.