Robust Models of Human Parsing

Frank Keller
School of Informatics
University of Edinburgh
2 Buccleuch Place
Edinburgh EH8 9LW, UK
keller@inf.ed.ac.uk

1 Robustness and Human Parsing

A striking property of the human parser is its efficiency and robustness. For the vast majority of sentences, the parser will effortlessly and rapidly deliver the correct analysis. In doing so, it is robust to noise, i.e., it can provide an analysis even if the input is distorted, e.g., by ungrammaticalities. Furthermore, the human parser achieves broad coverage: it deals with a wide variety of syntactic constructions, and is not restricted by the domain, genre, or modality of the input.

Current research on human parsing rarely investigates the issues of efficiency, robustness, and broad coverage, as pointed out by Crocker and Brants (2000). Instead, most researchers have focussed on the *difficulties* that the human parser has with certain types of sentences. Based on the study of garden path sentences (which involve a local ambiguity that makes the sentence hard to process), theories have been developed that successfully explain how the human parser deals with ambiguities in the input. However, garden path sentences are arguably a pathological case for the parser; garden paths are not representative of naturally occurring text. This means that the corresponding processing theories face a scaling problem: it is not clear how they can explain the normal behavior of the human parser, where sentence processing is highly efficient and very robust (see Crocker and Brants 2000 for details on this scalability argument).

This criticism applies to most existing theories of human parsing, including the classical garden path model advanced by Frazier and Rayner (1982) and Frazier (1989), and more recent lexicalist parsing frameworks, of which MacDonald et al. (1994) and MacDonald (1994) are representative examples. Both the garden path model and the lexicalist model are designed to deal with idealized input, i.e., with input that is (locally) ambiguous, but fully well-formed. A real life parser, however, has to cope with a large amount of noise, which often renders the input ungrammatical or fragmentary, due to errors such as typographical mistakes in the case of text, or slips of the tongue, disfluencies, or repairs in the case of speech. A quick search in the Penn Treebank (Marcus et al., 1993) shows that about 17% of all sentences contain parentheticals or other sentence fragments, interjections, or unbracketable constituents. Note that this figure holds for carefully edited newspaper text; the figure is likely to be much higher for speech. The human parser is robust to such noise, i.e., it is able to assign an (approximate) analysis to a sentence even if it is ungrammatical or fragmentary.

2 Probabilistic Parsing Models

In computational linguistics, probabilistic approaches to language processing play a central role. Significant advances toward robust, broad-coverage parsing models have been made based on probabilistic techniques such as maximum likelihood estimation or expectation maximization (for an overview, see Manning and Schütze, 1999).

An example of a simple probabilistic parsing model are probabilistic context-free grammars (PCFGs), which extend the formalism of context-free grammars (CFGs) by annotating each rule with a probability. PCFGs constitute an efficient, well-understood technique for assigning probabilities to the analyses produced by a context-free grammar. They are commonly used for broad-coverage grammars, as CFGs large enough to parse unrestricted text are typically highly ambiguous, i.e., a single sentence will receive a large number of parses. The probabilistic component of the grammar can then be used to rank the analyses a sentence might receive, and improbable ones can be eliminated.

In the computational linguistics literature, a number of highly successful extensions to the basic PCFG model have been proposed. Of particular interest are lexicalized parsing models such as the ones developed by Collins (1996, 1997) and Carroll and Rooth (1998).

In the human parsing literature, a PCFG-based model has been proposed by Jurafsky (1996) and

Narayanan and Jurafsky (1998). This model shows how different sources of probabilistic information (such as subcategorization information and rule frequencies) can be combined using Bayesian inference. The model accounts for a range of disambiguation phenomena in linguistic processing. However, the model is only small scale, and it is not clear if it can be extended to provide robustness and coverage of unrestricted text.

This problem is addressed by Brants and Crocker (2000) and Crocker and Brants (2000), who propose a broad-coverage model of human parsing based on PCFGs. This model is incremental, i.e., it makes word-by-word predictions, thus mimicking the behavior of the human parser. Also, Brants and Crocker's (2000) model imposes memory restrictions on the parser that are inspired by findings from the human sentence processing literature.

3 Robust Models of Human Parsing

The main weakness of both the Narayanan/Jurafsky and the Crocker/Brants model (discussed in the previous section) is that they have not been evaluated systematically. The authors only describe the performance of their models on a small set of hand-picked example sentences. No attempts are made to test the models against a full set of experimental materials and the corresponding reading times, even though a large amount of suitable data are available in the literature. This makes it very hard to obtain a realistic estimate of how well these models achieve the aim of providing robust, broad coverage models of human parsing. This can only be assessed by testing the models against realistic samples of unrestricted text or speech obtained from corpora.

In this talk, we will present work that aims to perform such an evaluation. We train a series of increasingly sophisticated probabilistic parsing models on an identical training set (the Penn Treebank). These models include a standard unlexicalized PCFG parser, a head-lexicalized parser (Collins, 1997), and a maximum-entropy inspired parser (Charniak, 2000). We test all three models on the Embra corpus, a corpus of newspaper texts annotated with eye-tracking data from 23 subjects (McDonald and Shillcock, 2003). A series of regression analyses are conducted to determine if per-sentence reading time measures correlate with sentence probabilities predicted by the parsing models. Three baseline models are also included in the evaluation: word frequency, bigram and trigram probability (as predicted by a language model), and part of speech (POS) probability (as predicted by a POS tagger). Models based on *n*-grams have already been used successfully to model eye-tracking data, both on a word-by-word basis (McDonald and Shillcock, 2003) and for whole sentences (Keller, 2004).

Our results show that for all three parsing models, sentence probability is significantly correlated with reading times measures. However, the models differ as to whether they predict early or late measures: the PCFG and the Collins model significantly predict late reading time measures (total time and gaze duration), but not early measures (first fixation time and skipping rate). The Charniak model is able to significantly predict both early and late measures.

An analysis of the baseline models shows that word frequency and POS probability only predict early measures, while bigram and trigram probability only predict late measures. This indicates that the Charniak model is able to predict both early and late measures because it successfully combines lexical information (word frequencies and POS probabilities) with phrasal information (as modeled by a PCFG). This finding is in line with Charniak's own analysis, which shows that the high performance of his model is due to the fact that it combines a third-order Markov grammar with sophisticated phrasal and lexical features (Charniak, 2000).

4 Implications

The results reported in the previous section have interesting theoretical implications. Firstly, there is a methodological lesson here: simple baseline models based on *n*-gram or POS probabilities perform surprisingly well as robust, broad coverage models of human language processing. This is an important point that has not been recognized in the literature, as previous models have not been tested on realistic corpus samples, and have not been compared to plausible baselines.

A second point concerns the role of lexical information in human parsing. We found that the best performing model was Charniak's maximum entropy-inspired parser, which combines lexical and phrasal information, and manages to predict both early and late eye-tracking measures. A number of existing theories of human parsing incorporate lexical information (MacDonald et al., 1994; MacDonald, 1994), but have so far failed to demonstrate how the use of such information can be scaled up to yield robust, broad coverage parsing models that can be tested on realistic data such as the Embra eye-tracking corpus.

Finally, a major challenge that remains is the crosslinguistic aspect of human parsing. Virtually all existing computational models have only been implemented and tested for English data. However, a wide range of interesting problems arise for other languages. An examples are head-final languages, in which the probabilistic information associated with the head becomes available only at the end of the phrase, which poses a potential problem for incremental parsing models. Some initial results on a limited dataset have been obtained by Baldewein and Keller (2004) for head-final constructions in German.

References

Baldewein, Ulrike and Frank Keller. 2004. Modeling attachment decisions with a probabilistic parser: The case of head final structures. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society*. Chicago.

Brants, Thorsten and Matthew W. Crocker. 2000. Probabilistic parsing and psychological plausibility. In *Proceedings of the 18th International Conference on Computational Linguistics*. Saarbrücken/Luxembourg/Nancy.

Carroll, Glenn and Mats Rooth. 1998. Valence induction with a head-lexicalized PCFG. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Granada, pages 36–45.

Charniak, Eugene. 2000. A maximum-entropy-inspired parser. In *Proceedings of the 1st Conference of the North American Chapter of the Association for Computational Linguistics*. Seattle, WA, pages 132–139.

Collins, Michael. 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, pages 184–191.

Collins, Michael. 1997. Three generative, lexicalised models for statistical parsing. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European Chapter of the Association for Computational Linguistics*. Madrid, pages 16–23.

Crocker, Matthew W. and Thorsten Brants. 2000. Wide-coverage probabilistic sentence processing. *Journal of Psycholinguistic Research* 29(6):647–669.

Frazier, Lynn. 1989. Against lexical generation of syntax. In William D. Marslen-Wilson, editor, *Lexical Representation and Process*, MIT Press, Cambridge, Mass., pages 505–528.

Frazier, Lynn and Keith Rayner. 1982. Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology* 14:178–210.

Jurafsky, Daniel. 1996. A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science* 20(2):137–194.

Keller, Frank. 2004. The entropy rate principle as a predictor of processing effort: An evaluation against eye-tracking data. In Dekang Lin and Dekai Wu, editors, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Barcelona.

MacDonald, Maryellen C. 1994. Probabilistic constraints and syntactic ambiguity resolution. *Language and Cognitive Processes* 9:157–201.

MacDonald, Maryellen C., Neal J. Pearlmutter, and Mark S. Seidenberg. 1994. Lexical nature of syntactic ambiguity resolution. *Psychological Review* 101:676–703.

Manning, Christopher D. and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19(2):313–330.

McDonald, Scott A. and Richard C. Shillcock. 2003. Low-level predictive inference in reading: The influence of transitional probabilities on eye movements. *Vision Research* 43:1735–1751.

Narayanan, Srini and Daniel Jurafsky. 1998. Bayesian models of human sentence processing. In Morton A. Gernsbacher and Sharon J. Derry, editors, *Proceedings of the 20th Annual Conference of the Cognitive Science Society*. Lawrence Erlbaum Associates, Mahwah, NJ.