A computational model of discourse predictions in sentence processing

Amit Dubey, Frank Keller, & Patrick Sturt (University of Edinburgh)
adubey@inf.ed.ac.uk

Syntax-discourse interactions; Cognitive modeling; English

Recent research in psycholinguistics has seen a growing interest in the role of prediction in sentence processing. Most attempts to computationally model predictive processing have focused on syntactic prediction. Examples include Hale (2001)'s surprisal model, which relates processing effort to the conditional probability of the current word given the previous words in the sentence. Recent work has attempted to integrate semantic and discourse prediction with models of syntactic processing. This includes Mitchell et al. (2010)'s approach, which combines an incremental parser with a vector-space model of semantics. However, this approach only provides a loose integration of the two components, and the notion of semantics is restricted to lexical meaning approximated by word co-occurrences. At the discourse level, Dubey (2010) has proposed a model that combines an incremental parser with a probabilistic logic-based model of co-reference resolution. However, this model does not explicitly model discourse effects in terms of prediction, and again only proposes a loose integration of co-reference and syntax. Furthermore, the Dubey (2010) model has not been evaluated on broad coverage data.

Here, we propose a computational model that captures discourse effects on syntax in terms of prediction. The model posits a strong link between syntax and co-reference: it comprises a co-reference component which maintains a list of previously mentioned NPs, and an hidden Markov model-based syntactic component which predicts facilitation for discourse entities which are discourse-old as opposed to discourse-new. This entails a strong interaction between parsing and reference: referential ambiguities are given probabilities and disambiguated by the parser.

Our evaluation experiments used the Dundee corpus (Kennedy et al., 2003), which has recently become the gold-standard in computational psycholinguistics (e.g. Demberg and Keller, 2008; Frank, 2009; Boston et al., 2008; Mitchell et al., 2010) For each word in the corpus, we computed total reading times, defined as the overall time participants spent looking at a word, including any re-fixations after looking away. We compared three mixed models: (i) a baseline, with only low-level eye movement variables (such as length and frequency) as predictors; (ii) the Syntax-only model, with the baseline factors plus syntactic surprisal scores; and (iii) the Co-reference model, with the baseline factors, syntactic surprisal scores, and surprisal scores based on our new co-reference model (residualized against syntactic surprisal).

We found that both the Syntax and Co-reference models provide a significantly better fit with the reading time data than the Baseline model; all three criteria agree: AIC and BIC lower than for the baseline, and log-likelihood is higher. Moreover, the Co-reference model provides a significantly better fit than the Syntax model, which demonstrates the benefit of co-reference information for modeling reading times. Again, all three measures of model fit provide the same result.

The primary finding of this work is that incorporating discourse information such as co-reference into an incremental probabilistic model of sentence processing has a beneficial effect on the ability of the model to predict broad-coverage human parsing behavior.

References

M. Boston, J. Hale, R. Kliegl, and S. Vasishth. Surprising parser actions and reading difficulty. ACL/HLT, 2008.V. Demberg and F. Keller. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. Cognition, 109:192–210, 2008.

A. Dubey. The influence of discourse on syntax: A psycholinguistic model of sentence processing. ACL 2010.

S. Frank. Surprisal-based comparison between a symbolic and a connectionist model of sentence processing. COGSCI, 2009.

J. T. Hale. A probabilistic Earley parser as a psycholinguistic model. NAACL, 2001.

A. Kennedy, R. Hill, and J. Pynte. The Dundee Corpus. ECEM, 2003.

J. Mitchell, M. Lapata, V. Demberg, and F. Keller. Syntactic and semantic factors in processing difficulty. ACL, 2010.