The Integration of Syntax and Semantic Plausibility in a Wide-Coverage Model of Human Sentence Processing

ulrike@coli.uni-sb.de

Ulrike Padó[1], Matthew W. Crocker[1] and Frank Keller[2]
[1]Saarland University, [2]University of Edinburgh

Abstract

We present the SynSem-Integration model of difficulty in human sentence processing. It integrates a probabilistic wide-coverage grammar-based model with a separate model for verb-argument thematic role assignment and thematic fit prediction which accounts for semantic plausibility effects.

Motivation

Important properties of the human sentence processor are

- Sensitivity to prior linguistic experience
- Immediate incremental interpretation
- Robust and accurate processing of unseen input (wide coverage)
- Influence of semantic plausibility

Different existing models account for different properties; none covers all.

Grammar-Based Models Crocker and Brants, 2000; Levy, 2005
- Rely on frequency information induced from large corpora
- Generate and rank syntactic analyses
- Do not integrate plausibility

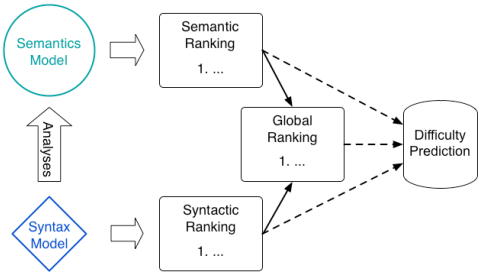
Constraint-Based Models McRae et al., 1998; Narayanan and Jurafsky, 2002
- Combine information from different sources
- Constraints support pre-defined analyses, most active one wins
- Do not show broad coverage (due to hand-selection of constraints)

The SynSem-Integration Model

The Syntax-Semantics Integration model combines a probabilistic model of **syntactic processing** with a general model of **semantic plausibility**. Both models have wide coverage and are automatically induced.

Both models rank the proposed syntactic analyses:
The syntax model by syntactic probability estimates, the semantics model by plausibility estimates for the verb-argument pairs in each structure.

A globally preferred analysis is determined by interpolating the two models' predictions. This analysis is assumed to be adopted by readers.



Predicting Difficulty

We associate difficulty in human sentence processing with two events:

- A **conflict** in syntactic and semantic preferences for the highest-ranked structure, e.g., during an *ambiguous region*
- A **revision** in the interpretation of the globally preferred analysis, e.g., at *disambiguation*

Conflict	Revision	
The patient cured	*by the therapy had*	*invented it himself.*

Comparison to Existing Models

Shared with constraint-based architectures:
- Combination of preferences from different sources
- Prediction of difficulty if preferences conflict
- But: Wide-coverage model, no need to hand-select constraints

Shared with grammar-based architectures:
- Probabilistic ranking of generated analyses
- Models automatically induced from large corpora
- But: Integration of plausibility

Evaluation

The SynSem-Integration model's predictions were tested against patterns of processing difficulty found for the main clause/reduced relative (MC/RR), NP/S, NP/0 and PP attachment phenomena (two studies per phenomenon).

The model significantly predicts the observed patterns of human difficulty, while a syntax-only baseline (equivalent to a lexicalized grammar-based model) fails.

Phenomenon	N	Model	Spearman's ρ
All	36	**SynSem**	**0.700,*****
		Baseline	-0.223,ns
MC/RR	14	**SynSem**	**0.792,*****
		Baseline	0.199,ns
NP/S	12	**SynSem**	**0.688, ***
		Baseline	-0.165,ns

MC/RR	McRae et al., JML 98 MacDonald, LCP 94
NP/S	Garnsey et al., JML 97 Pickering & Traxler, JEP:LMC 98
NP/0	Pickering & Traxler, JEP:LMC 98 Pickering et al., JML 00
PP	Rayner et al., J.V Learn V Beh 83 Taraban & McClelland, JML 88

The Semantic Model

The semantic model approximates world knowledge by exploiting the link between plausibility and word co-occurrence in a corpus annotated with thematic roles. Padó et al., 2006

Given an arbitrary verb-argument pair, it predicts a preferred role relation and its plausibility. Plausibility is equated to probability of encountering the pair in the respective relation in the FrameNet corpus.

Smoothing sparse data to achieve wide coverage:
- Semantic generalization: Pooling observations of words from the same semantic class
  Nouns: WordNet synsets
  Verbs: classes automatically induced from FrameNet
- Re-estimation smoothing: Good-Turing smoothing

The predictions of this wide-coverage semantic model are significantly correlated to human plausibility judgments

Data Set	N	Spearman's ρ
McRae et al., 98	64	0.415, **
Pado et al., 06	414	0.522, ***

The Syntax Model

The syntax model incrementally constructs structural analyses of the input and ranks them by their syntactic probability.

We use an incremental top-down parser (Roark 2001).
Its grammar and lexicon are derived from the syntactically-annotated Penn TreeBank corpus.

Performance on the standard parser test set shows that the model assigns accurate analyses to unseen input.

Test Set	Coverage	F-Score
PTB 23	100%	86.29

References

M. Crocker and Th. Brants, Wide-coverage probabilistic sentence processing. JPR, 29(6), 647-669. 2000.
R. Levy. Probabilistic models of word order and syntactic discontinuity. PhD Thesis, Stanford University. 2005.
K. McRae, M. Spivey-Knowlton and M. Tanenhaus, Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. JML, 38, pp. 283-312, 1998.
S. Narayanan and D. Jurafsky, A Bayesian model predicts human parse preference and reading time in sentence processing. In: Dietterich, Becker and Ghahramani, eds., Advances in Neural Information Processing Systems 14, MIT Press, 2002.
U. Padó, M. Crocker and F. Keller, Modeling semantic role plausibility in human sentence processing. In: Proceedings of the Meeting of the European Chapter of the Association for Computational Linguistics, 2006.
B. Roark, Robust probabilistic predictive syntactic processing: Motivations, models, and applications. PhD Thesis, Brown University. 2001.