Modelling Human Semantic Judgments – Adjuncts and Seen Data

Ulrike Padó

12/1/2006

Motivation

- Global goal: Predict human reading times
  - Model incremental processing
  - Model syntactic and semantic preferences
- Standard models: Syntactic parsers
  - Assign structure incrementally
  - Assumption: Best parse at any point is the one humans prefer

Motivation

A parser trained on the WSJ predicts the following sentence structures to be preferred at each word:

The	(S (NP (DT The)))	✔
cop	(S (NP (DT The) (NN cop)))	✔
arrested	(S (NP (DT The) (NN cop)) (VP (V arrested)))	✔
by	(S (NP (NP (DT The) (NN cop)) (VP (V arrested) (PP (IN by))))	✔

Motivation

But: Sentence semantics change human behaviour!

The	(S (NP (DT The)))	✔
crook	(S (NP (DT The) (NN crook)))	✔
arrested	(S (NP (DT The) (NN crook)) (VP (V arrested)))	✘
by	(S (NP (NP (DT The) (NN crook)) (VP (V arrested) (PP (IN by))))	✔

So let‘s evaluate the semantics of each structure, too!

Recap: The Semantic Model

- Model outputs plausibility score for parser's structures to set off syntactic probability
- Use thematic roles to link to semantics of verb-argument relations in a structure
- Estimate plausibility of a verb-role-argument triple as its probability

Recap: The Semantic Model

$$Plausibility_{v,r,a} = P(verb_s, role, arg, gf) = P(verb_s) * P(gf|verb_s) * P(role|verb_s, gf) * P(arg|verb_s, gf, role)$$

$$Plausibility_{v,r,a} = P(arrest_1, crook, suspect, obj) = P(arrest_1) * P(obj|arrest_1) * P(suspect|arrest_1, obj) * P(crook|arrest_1, obj, suspect)$$

Sparse Data

- Training on FN/PB corpora and testing on psycholinguistic items causes extreme sparse data problems
- Two (orthogonal) approaches:
  - Good-Turing Smoothing assigns probs to unseen pairs
  - Class-Based Smoothing adds more counts for estimation: Count “policeman arrest” as well as “cop arrest”

Overview

- Motivation
- Recap: The Model/Smoothing
- Today’s Questions
- Modelling Instruments and Locations
- Modelling Seen Data
- Conclusions

Today's Questions

Our semantic model predicts human plausibility judgements for roles like `Agent` and `Patient.`

- Can the model predict ratings for Instruments and Locations?

Even with smoothing, making good predictions about data is a big problem.

- What happens if the test data is more similar to the training data?

Instruments and Locations

- In the PB corpus
  - Default: ArgM roles (`ArgM-Loc` and `ArgM-Mnr`)
  - Many Instruments have ArgN role (e.g. Arg2)
  - WSJ running text annotates all locations and instruments: Expect many locations/instruments
- In the FN corpus
  - Default: Non-Core roles (e.g. `Place`, `Instrument`)
  - Some instruments receive Core role
  - Non-Core roles are not in the focus of lexicographic interest: Expect fewer data points than in PB

Instr/Loc: Data

- Plausibility judgments on a 7-point scale
- Plausible, implausible and medium items
- Test all items where verb is seen (unless it assigns no role)

eat	in a bedroom	3.5
eat	in a lobby	2.9
eat	in a kitchen	7.0
eat	with a fork	6.7
eat	with a toothpick	2.1
eat	with pliers	1.0

Task

- Correlation task: Reach sig. correlation to human data
- Labelling task: Assign the correct role
- Run model with or without syntactic information
  - (eat, fork, dep-with) vs (eat, fork)

Locations: Results

	Model	Cov.	ρ	Labelling F	Labelling Cov.
FN	Freq. Bsl	--	--	0	100%
	No Syn	65.4%	0.270, **	17.9	100%
	Syn	65.4%	0.209, *	46.8	100%
	Labeller	69.2%	0.190, *	50.0	100%
PB	Freq. Bsl	--	--	0.5	100%
	No Syn	63.3%	0.087, ns	9.2	100%
	Syn	63.3%	0.095, ns	37.8	100%
	Labeller	100%	-0.02, ns	82.5	100%

Locations: Specifying Syn

	Model	Cov.	ρ	Labelling F	Labelling Cov.
FN	Freq. Bsl	--	--	0	100%
	No Syn	65.4%	0.270, **	17.9	100%
	Syn	65.4%	0.209, *	46.8	100%
	Labeller	69.2%	0.190, *	50.0	100%
PB	Freq. Bsl	--	--	0.5	100%
	No Syn	63.3%	0.087, ns	9.2	100%
	Syn	63.3%	0.095, ns	37.8	100%
	Labeller	100%	-0.02, ns	82.5	100%

Locations: Labeller/FN

	Model	Cov.	ρ	Labelling F	Labelling Cov.
FN	Freq. Bsl	--	--	0	100%
	No Syn	65.4%	0.270, **	17.9	100%
	Syn	65.4%	0.209, *	46.8	100%
	Labeller	69.2%	0.190, *	50.0	100%
PB	Freq. Bsl	--	--	0.5	100%
	No Syn	63.3%	0.087, ns	9.2	100%
	Syn	63.3%	0.095, ns	37.8	100%
	Labeller	100%	-0.02, ns	82.5	100%

Locations: Labeller/PB

	Model	Cov.	ρ	Labelling F	Labelling Cov.
FN	Freq. Bsl	--	--	0	100%
	No Syn	65.4%	0.270, **	17.9	100%
	Syn	65.4%	0.209, *	46.8	100%
	Labeller	69.2%	0.190, *	50.0	100%
PB	Freq. Bsl	--	--	0.5	100%
	No Syn	63.3%	0.087, ns	9.2	100%
	Syn	63.3%	0.094, ns	37.8	100%
	Labeller	100%	-0.02, ns	82.5	100%

Coverage Differences

- Our model only assigns roles it has seen with the verb during training: Labeller generalises better
- Skewed distribution of Location roles:
  - Over the tested verbs, FN and PB contain same percentage of default Location roles
  - FN: One verb (3.8% of items) seen without
  - PB: 13 verbs (37.5% of items) seen without

Locations: Observations

- Despite lower frequency baseline, FN does better than PB (both labelling and correlation)
- No significant difference between Labeller and FN model!
- Labeller generalises much better than PB model
  - Many PB verbs seen without `ArgM-Loc` role: Our model restricted to roles seen with verb
  - Running text contains as many Location roles as lexicographic corpus, but distribution is more skewed!

Instruments: Results

	Model	Cov.	ρ	Labelling F	Labelling Cov.
FN	Freq. Bsl	--	--	4.3	100%
	No Syn	45.7%	0.303, **	15.4	100%
	Syn	45.7%	0.157, ns	24.7	100%
	Labeller	70.9%	0.139, ns	25.9	100%
PB	Freq. Bsl	--	--	0	100%
	No Syn	81.4%	0.017, ns	4.1	100%
	Syn	81.4%	0.076, ns	22.4	100%
	Labeller	100%	0.017, ns	47.1	100%

Instruments: Specifying Syn

	Model	Cov.	ρ	Labelling F	Labelling Cov.
FN	Freq. Bsl	--	--	4.3	100%
	No Syn	45.7%	0.303, **	15.4	100%
	Syn	45.7%	0.157, ns	24.7	100%
	Labeller	70.9%	0.139, ns	25.9	100%
PB	Freq. Bsl	--	--	0	100%
	No Syn	81.4%	0.017, ns	4.1	100%
	Syn	81.4%	0.076, ns	22.4	100%
	Labeller	100%	0.017, ns	47.1	100%

Instruments: Labeller/Model

	Model	Cov.	ρ	Labelling F	Labelling Cov.
FN	Freq. Bsl	--	--	4.3	100%
	No Syn	45.7%	0.303, **	15.4	100%
	Syn	45.7%	0.157, ns	24.7	100%
	Labeller	70.9%	0.139, ns	25.9	100%
PB	Freq. Bsl	--	--	0	100%
	No Syn	81.4%	0.017, ns	4.1	100%
	Syn	81.4%	0.076, ns	22.4	100%
	Labeller	100%	0.017, ns	47.1	100%

Instruments: Observations

- FN/PB about equal in terms of labelling/correlations
- No sig. correlations
- Labeller does no better than FN (despite generalising more: Lexicalisation!)
- Labeller does much better for PB: Default role is more predictable
- Instrument roles seem more comparably distributed than Locations

Instr/Loc: Summary

- Prediction task: FN reaches sig. correlations
- Labelling task:
  - Simpler model, yet no sig. difference to role labeller for FN
  - Labeller profits from generalising if default role is predictable (e.g. ArgM-Mnr)
- Locations are sparser for PB than for FN verbs!
- Instruments are harder than locations
  - Fewer (inferable) instruments in corpora?

Overview

- Motivation
- Recap: The Model/Smoothing
- Today's Questions
- Modelling Instruments and Locations
- Modelling Seen Data
- Conclusions

Seen Data

- General problem: Data sparseness because test and training data are very different
- So: Unclear how good the model really is!
- Solution: Run rating study on items from the training corpora

Study Setup

- 18 verbs
  - Covered in PB and FN
  - Assigning `Agt-Pat`, `Agt-Rec`, `Exp-Theme`
- Fillers:
  - From each corpus: 3 most frequent fillers per role
  - No overlap if possible
  - (3+3) * 2: up to 12 fillers per verb
- 207 items; at least 50% seen for each training set

Study

- On-line study using WebExp
- 25 participants rate items on 7-point scale
- Inter-rater correlation (upper bound): 0.68

eliminate job	Agent	3.0	
	Theme	5.0	PB Theme
eliminate law	Agent	4.6	FN Agent
	Theme	3.0	
eliminate policy	Agent	4.2	
	Theme	4.2	PB Agent

Seen Data: Results

	Model	Cov.	ρ	Labelling F	Labelling Cov.
Upper Bnd		100%	0.68	--	--
FN	Freq. Bsl	--	--	28.5	100%
	Bsl Model	96.4%	0.390, ***	48.8	100%
	Model	96.4%	0.476, ***	45.4	100%
PB	Freq. Bsl	--	--	38.6	100%
	Bsl Model	100%	0.323, ***	61.4	100%
	Model	100%	0.297, ***	55.6	100%

Seen Data: FN

	Model	Cov.	ρ	Labelling F	Labelling Cov.
Upper Bnd		100%	0.68	--	--
FN	Freq. Bsl	--	--	28.5	100%
	Bsl Model	96.4%	0.390, ***	48.8	100%
	Model	96.4%	0.476, ***	45.4	100%
PB	Freq. Bsl	--	--	38.6	100%
	Bsl Model	100%	0.323, ***	61.4	100%
	Model	100%	0.297, ***	55.6	100%

Seen Data: PB

	Model	Cov.	ρ	Labelling F	Labelling Cov.
Upper Bnd		100%	0.68	--	--
FN	Freq. Bsl	--	--	28.5	100%
	Bsl Model	96.4%	0.390, ***	48.8	100%
	Model	96.4%	0.476, ***	45.4	100%
PB	Freq. Bsl	--	--	38.6	100%
	Bsl Model	100%	0.323, ***	61.4	100%
	Model	100%	0.297, ***	55.6	100%

Seen Data: Summary

- Correlation results better than for data tested so far (.21 away from upper bound)
- FN model robustly beats baselines
- PB model's correlation sig. worse than model baseline!
  - Using PB clusters makes things worse

Conclusions

- Model manages to predict adjunct roles
  - Sparse data
  - Can't generalise if role unseen
  - No difference to standard role labeller for FN!
- Correlations to seen data are stronger
  - Model does better if test and train are similar
  - Model relatively close to upper bound!

Current/Future Work

- Experiment with parser alone: Does it behave as expected?
- Combine parser and semantic model
- Link combined model predictions to human data (e.g. reading times)