Modelling Semantic Role Plausibility in Human Sentence Processing

Comparing FrameNet and PropBank

Ulrike Padó

Computational Linguistics
Saarland University
ulrike@coli.uni-sb.de

Multilingual Semantic Annotation: Theory and Applications
Workshop, Saarbrücken July 26-27

Models of Human Sentence Processing

Goal

- Model processes and mechanisms of human sentence processing
- Mirrored in observable behaviour: Human reading of isolated sentences (e.g., duration)
  - Investigate behaviour during difficulty: (Syntactic) ambiguity

Models of Human Sentence Processing

Goal

- Model processes and mechanisms of human sentence processing
- Mirrored in observable behaviour: Human reading of isolated sentences (e.g., duration)
  - Investigate behaviour during difficulty: (Syntactic) ambiguity

Existing Models

- Rely on probability of syntactic analyses, for example: The Frequency Hypothesis (Jurafsky 1996)
  - Readers prefer the syntactically most likely analysis at each word
  - Show difficulty if their preferred analysis has to be revised
- Models are implemented using PCFG-based parser
- Make correct predictions for many well-studied ambiguities
- But: Cannot capture effects of semantic plausibility!

Semantics Influence Processing of Ambiguity

Semantics Influence Processing of Ambiguity

The hunter

Semantics Influence Processing of Ambiguity

The hunter shot

Semantics Influence Processing of Ambiguity

The hunter shot by the teenager

Semantics Influence Processing of Ambiguity

The hunter shot by the teenager was quite young.

Semantics Influence Processing of Ambiguity

Semantics Influence Processing of Ambiguity

The deer

Semantics Influence Processing of Ambiguity

The deer shot

Semantics Influence Processing of Ambiguity

The deer shot by the hunter

Semantics Influence Processing of Ambiguity

The deer shot by the hunter was truly impressive.

Semantics Influence Processing of Ambiguity

The hunter shot by the teenager was quite young.
The deer shot by the hunter was truly impressive.

Semantics Influence Processing of Ambiguity

The hunter shot by the teenager was quite young.
The deer shot by the hunter was truly impressive.

Main Clause/Reduced Relative Ambiguity

- At *shot*, main clause or reduced relative are possible
- Semantics of the first NP influences the preferred interpretation (McRae et al. 1998)
- But: Syntactic structure is identical, so parser makes identical predictions
- Semantics influence many ambiguities: PP Attachment, Direct object/Sentential complement, Direct Object/Main Clause, ...

Overall Goal: Include Semantics

Strategy 1: Lexicalisation

- Existing smoothing approaches aren't enough: Lexical material differs widely between training and test data
  - Training: Wall Street Journal
  - Test: Psycholinguistic experimental items
- No semantic representation: Word co-occurrence, possibly with structural links

Overall Goal: Include Semantics

Strategy 2: Independent Semantic Module

- Assumption: Frequency Hypothesis from syntax can be extended to semantics
  - Most likely interpretation given the training data is preferred
- Construct and evaluate shallow semantic representation
- Better control of smoothing methods

Outline

Outline

Outline

Outline

Outline

Outline

Modelling Sentence Plausibility

Step 1: Using Role Semantics

- Extract predicate-argument pairs from parser
- Model plausibility of predicate-argument structure: Plausibility of thematic role assignment to predicate-argument pairs

Step 2: Applying the Frequency Hypothesis

- Model plausibility as probability: Predict most likely role to be preferred

A Probabilistic Model of Semantic Plausibility

- Plausibility of verb-role-argument triple: Joint probability of verb sense, argument, role and grammatical function

$$Plausibility(shoot, deer, patient) = P(shoot_1, deer, patient, obj)$$

A Probabilistic Model of Semantic Plausibility

- Plausibility of verb-role-argument triple: Joint probability of verb sense, argument, role and grammatical function

$$\begin{aligned} Plausibility(shoot, deer, patient) =& P(shoot_1, deer, patient, obj) \\ =& P(shoot_1) \cdot P(obj|shoot_1) \cdot \\ & P(patient|shoot_1, obj) \cdot \\ & P(deer|shoot_1, obj, patient) \end{aligned}$$

Sparse data:

- Decompose into less sparse subterms
- Extra smoothing effort for crucial most specific term

Smoothing Methods

Good-Turing (GT) Smoothing

Smooth estimated distribution: Assign probability mass to unseens

Smoothing Methods

Good-Turing (GT) Smoothing

Smooth estimated distribution: Assign probability mass to unseens

Class-Based (CB) Smoothing

Modify estimation by generalising from word tokens to word classes

- Assume {deer, elk, moose}
- Count <shoot, elk, patient> as well as <shoot, deer, patient>

Use WN SynSets for nouns, derive verb classes from training data

Smoothing Methods

Good-Turing (GT) Smoothing

Smooth estimated distribution: Assign probability mass to unseens

Class-Based (CB) Smoothing

Modify estimation by generalising from word tokens to word classes

- Assume {deer, elk, moose}
- Count <shoot, elk, patient> as well as <shoot, deer, patient>

Use WN SynSets for nouns, derive verb classes from training data

Combining the Methods

Use CB smoothing for $P(arg\ head|verb, gf, role)$ term, GT smoothing for all other terms and if CB term is still sparse

Outline

Task

Question: Do the model predictions mirror human intuitions?

- Predict human semantic judgements for verb-argument-role triples (independent of parser)
- Human ratings are on a 1-7 scale
  - Rating for <shoot, hunter, agent>: 6.7
  - Rating for <shoot, deer, agent>: 1.0
- Correlate predictions and human data: Spearman's ρ
  - Ranges between 0 and 1
  - Human rater inter-correlation is generally lower: around 0.7

Training Data

FrameNet (FN)
Killing: [The hunter $_{Killer}$] shot [the deer $_{Victim}$].

- 57,000 propositions, 2,000 verbs
- Verbs are grouped into frames (prototypical situations)
- Frames have role inventory that describes typical participants: Abstract semantic information
- Caveat: FrameNet corpus built to lexicographic criteria, frequencies may not be representative of English

PropBank (PB)
shoot.02: [The hunter $_{Arg0}$] shot [the deer $_{Arg1}$].

- 120,000 propositions, 3,000 verbs
- Annotation of running text, roles are verb-specific and not semantically defined

Test Data

shoot	hunter	agent	6.9
shoot	hunter	patient	2.8
shoot	deer	agent	1.0
shoot	deer	patient	6.4

- McRae et al. 1998. Balanced: half good agents, half good patients. PB verb coverage: 92%, FN verb coverage 64%, 100 items in total (only covered items tested).
- Padó et al. 2006. All verbs and c. 50% of fillers seen for each training resource. PB+FN verb coverage: 100%, 414 items in total.

Test Data

shoot	hunter	agent	6.9
shoot	hunter	patient	2.8
shoot	deer	agent	1.0
shoot	deer	patient	6.4

- McRae et al. 1998. Balanced: half good agents, half good patients. PB verb coverage: 92%, FN verb coverage 64%, 100 items in total (only covered items tested).
- Padó et al. 2006. All verbs and c. 50% of fillers seen for each training resource. PB+FN verb coverage: 100%, 414 items in total.

Results: Correlations with Human Judgements

Train	Test	Unsmoothed			Smoothed		
		Cov.	ρ		Cov.	ρ	
PB	McRae	2%	–		93.5%	0.107,	ns
	Padó	34%	0.400,	**	100%	0.272,	***
FN	McRae	6%	–		87.5%	0.414,	*
	Padó	27.1%	0.365,	*	98.5%	0.522,	***

ns: not significant, *: p<0.05, **: p<0.01, ***: p<0.001

Results: Correlations with Human Judgements

Train	Test	Unsmoothed			Smoothed		
		Cov.	ρ		Cov.	ρ	
PB	McRae	2%	–		93.5%	0.107,	ns
	Padó	34%	0.400,	**	100%	0.272,	***
FN	McRae	6%	–		87.5%	0.414,	*
	Padó	27.1%	0.365,	*	98.5%	0.522,	***

ns: not significant, *: p<0.05, **: p<0.01, ***: p<0.001

Observations

- Combining Good-Turing smoothing and Class-Based smoothing results in good coverage

Results: Correlations with Human Judgements

Train	Test	Unsmoothed			Smoothed		
		Cov.	ρ		Cov.	ρ	
PB	McRae	2%	–		93.5%	0.107,	ns
	Padó	34%	0.400,	**	100%	0.272,	***
FN	McRae	6%	–		87.5%	0.414,	*
	Padó	27.1%	0.365,	*	98.5%	0.522,	***

ns: not significant, *: p<0.05, **: p<0.01, ***: p<0.001

Observations

- Smoothing methods are appropriate: Significant correlations to human judgements

Results: Correlations with Human Judgements

Train	Test	Unsmoothed			Smoothed		
		Cov.	ρ		Cov.	ρ	
PB	McRae	2%	–		93.5%	0.107,	ns
	Padó	34%	0.400,	**	100%	0.272,	***
FN	McRae	6%	–		87.5%	0.414,	*
	Padó	27.1%	0.365,	*	98.5%	0.522,	***

ns: not significant, *: p<0.05, **: p<0.01, ***: p<0.001

Observations

- Smoothing methods are appropriate: Significant correlations to human judgements
- But: PB-Padó ρ better unsmoothed than smoothed?

Results: Correlations with Human Judgements

Train	Test	Unsmoothed			Smoothed		
		Cov.	ρ		Cov.	ρ	
PB	McRae	2%	–		93.5%	0.107,	ns
	Padó	34%	0.400,	**	100%	0.272,	***
FN	McRae	6%	–		87.5%	0.414,	*
	Padó	27.1%	0.365,	*	98.5%	0.522,	***

ns: not significant, *: p<0.05, **: p<0.01, ***: p<0.001

Training Data

- FrameNet training data yields better results than PropBank data
- Using FrameNet, we reliably predict human intuitions

Outline

The MC/RR Ambiguity Revisited

The hunter shot by the teenager was quite young.

Experimental Results

- McRae et al. 1998: Thematic fit of *hunter* and *shot* influences processing

The MC/RR Ambiguity Revisited

The hunter [shot by] the teenager was quite young.

Experimental Results

- McRae et al. 1998: Thematic fit of *hunter* and *shot* influences processing
- Reading speed measured for [verb + by],

The MC/RR Ambiguity Revisited

The hunter [shot by] [the teenager] was quite young.

Experimental Results

- McRae et al. 1998: Thematic fit of *hunter* and *shot* influences processing
- Reading speed measured for [verb + by], [agent NP],

The MC/RR Ambiguity Revisited

The hunter [shot by] [the teenager] [was] quite young.

Experimental Results

- McRae et al. 1998: Thematic fit of *hunter* and *shot* influences processing
- Reading speed measured for [verb + by], [agent NP], [main verb]

The MC/RR Ambiguity Revisited

The hunter [shot by] [the teenager] [was] quite young.

Experimental Results

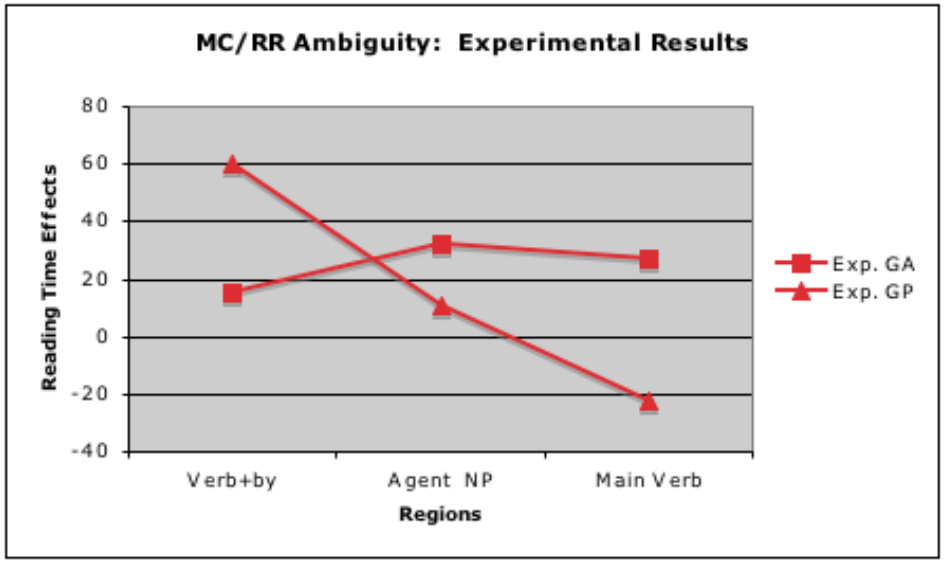
- McRae et al. 1998: Thematic fit of *hunter* and *shot* influences processing
- Reading speed measured for [verb + by], [agent NP], [main verb]
- Longer reading times indicate processing difficulty

The MC/RR Ambiguity Revisited

The hunter [shot by] [the teenager] [was] quite young.

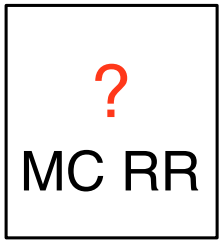
Findings

- Verb+by: *The deer shot by* (good patient, GP) is harder than *The hunter shot by* (good agent, GA)
- Agent NP, Main Verb: Harder after reading good agents

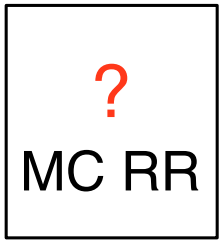


Predicting Difficulty

The deer shot by the teenager ...



The hunter shot by the teenager ...

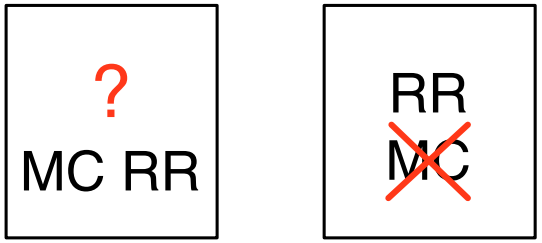


When is difficulty observed?

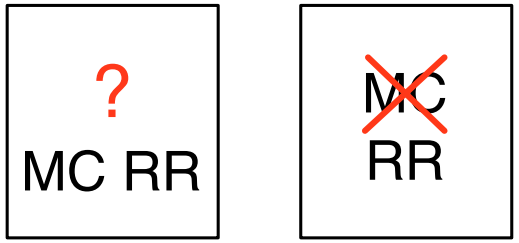
- During ambiguities: Several interpretations, possibly conflicting evidence for which to prefer: **Conflict**

Predicting Difficulty

The deer shot by the teenager ...



The hunter shot by the teenager ...

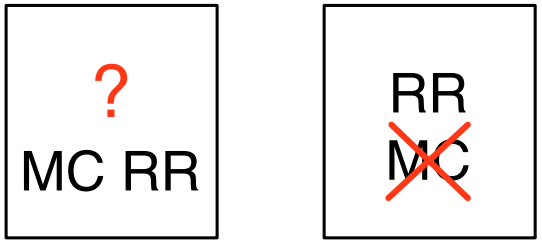


When is difficulty observed?

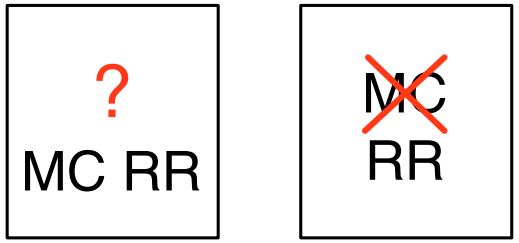
- During ambiguities: Several interpretations, possibly conflicting evidence for which to prefer: **Conflict**
- At point of disambiguation: One (possibly unexpected) interpretation is forced: **Revision**

Predicting Difficulty

The deer shot by the teenager ...



The hunter shot by the teenager ...

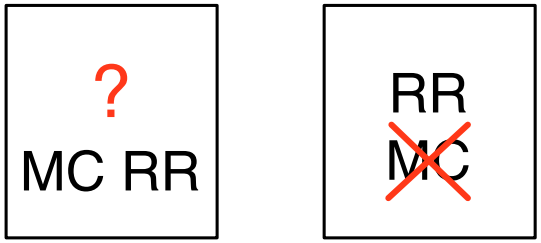


How do we predict difficulty?

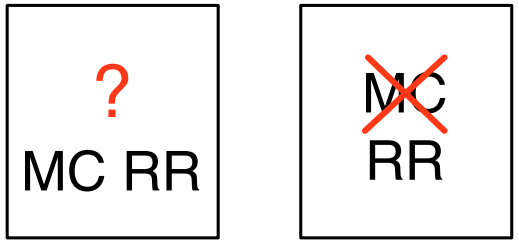
- Propose combined model with two components
  - Syntactic model: Accounts for syntactic preferences
  - Semantic model: Evaluates semantic plausibility of parses

Predicting Difficulty

The deer shot by the teenager ...



The hunter shot by the teenager ...



How do we predict difficulty?

- Propose combined model with two components
  - Syntactic model: Accounts for syntactic preferences
  - Semantic model: Evaluates semantic plausibility of parses
- Predict difficulty
  - If syntax and semantics disagree on preferred analysis → Conflict
  - If semantic interpretation has to be revised from one step to the next and becomes less likely → Revision

The Sentence Processing Model

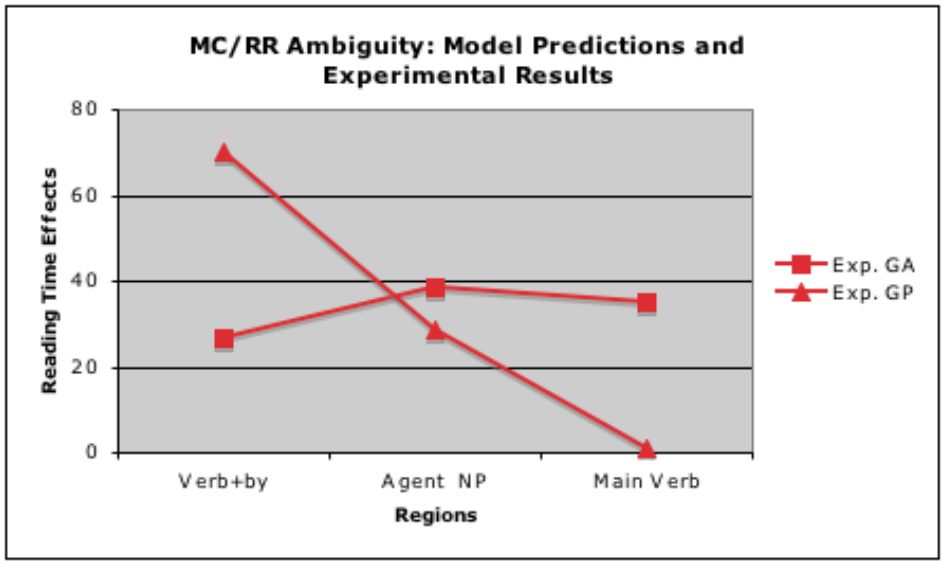
Components

- Syntax
  - Incremental probabilistic parser as model of syntactic preferences: Roark (2001)
  - Trained on Penn Treebank, sections 2-21+24
- Semantics
  - Semantic model trained on FrameNet 1.2 data
  - Evaluates predicate-argument pairs from parser analyses

Predicting Difficulty

- Find overall preferred analysis: Interpolate syntax and semantics
- Assign fixed cost if Conflict occurs
- Assign fixed cost if Revision is necessary
- Average cost over items: High cost predicts difficulty

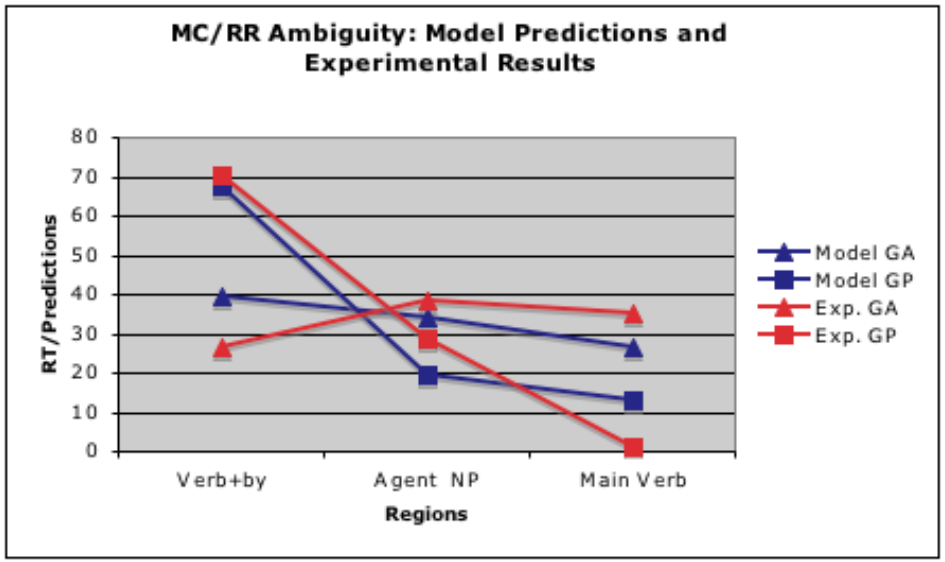
Predictions and Observed Data



Predictions and Observations

- Predictions and observations scaled to 1: Which percentage of (predicted) difficulty falls to the region?

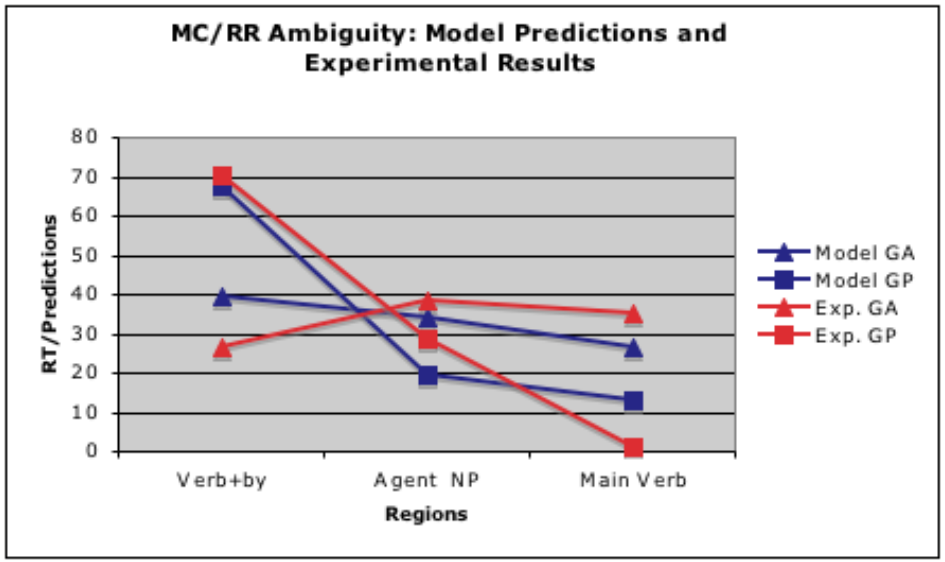
Predictions and Observed Data



Predictions and Observations

- Model's predictions match observed pattern well
  - GP harder than GA at Verb+by
  - GA harder than GP everywhere else

Predictions and Observed Data



Predictions and Observations

- Model's predictions match observed pattern well
  - GP harder than GA at Verb+by
  - GA harder than GP everywhere else
- Predictions significantly correlated to observations: Pearson's $r=0.896$, $p<0.01$

Outline

FrameNet works better – why?

Observations

- FN predictions are more reliable than PB predictions
- On FN, smoothing makes coverage and correlation ρ go up to reliable amounts
- On PB (Padó data) correlation ρ goes down with smoothing

Reasons

- Genre/vocabulary: BNC closer to test items
- Smoothing power of induced verb classes: FN semantic annotation yields better clusters

Genre and Vocabulary

Semantic Model assumes that plausible events are frequently talked about in corpora
→ Corpus data effectively models human experience of the world

Genre Effect

- BNC (FN) vs WSJ newswire (PB): PB models more restricted area of human experience
  - Better fit of BNC vocabulary with vocabulary in experimental items
  - Easier to model events like <shoot, deer, patient> if <shoot, elk, patient> and <shoot, moose, patient> have been seen

Genre and Vocabulary

Semantic Model assumes that plausible events are frequently talked about in corpora
→ Corpus data effectively models human experience of the world

Genre Effect

- BNC (FN) vs WSJ newswire (PB): PB models more restricted area of human experience
  - Better fit of BNC vocabulary with vocabulary in experimental items
  - Easier to model events like <shoot, deer, patient> if <shoot, elk, patient> and <shoot, moose, patient> have been seen
- But: This is not the whole story
  - Half of Padó et al. data is from PB and unseen in FN
  - FN model also generalises well to PB data

Verb Classes

Clustering Verbs into Classes

- Aim: Induce classes of semantically similar verbs
- Features for each verb-argument occurrence:
  - Argument head
  - Argument role label
  - Verb sense (FN frame, PB sense)
  - Syntactic path between verb and argument
  - Path-Role: Explicit linking information
- Soft clustering
  - Verbs are assigned to classes according to feature similarities
  - Verbs can occur in more than one class (probabilistic membership)
- FN: 13 clusters (relatively homogeneous size, clear semantic similarity),
  PB: 4 clusters (one huge, three small; loose semantic similarity)

Example Clusters

FN		PB	
“Move”	“Communicate”	“?”	“?”
cycle	tell	pay	turn
follow	advise	bring	begin
travel	confide	support	tell
lead	inform	receive	follow
chase	urge	develop	go
accompany	confess	announce	affect
escort	write	yield	join
commute	address	sell	trade
trail	ask	reduce	operate
usher	request	consider	introduce

Importance of Features for Induced Classes

	FN		PB	
Features	Correlation	Coverage	Correlation	Coverage
All	0.790,**	21%	0.136, ns	41%
No label	0.745, *	19%	0.074, ns	45%
No sense	0.657, *	21%	0.136, ns	41%
No arg head	0.790,**	21%	0.136, ns	42%
No path	0.414, ns	21%	–	2%
No path-label	0.711,*	19%	0.074, ns	45%

ns: not significant, *: $p < 0.05$, **: $p < 0.01$

Observations

- Importance of features: Leaving out one feature at a time for FN and PB data and estimating smoothing power of the resulting verb clusters

Importance of Features for Induced Classes

	FN		PB	
Features	Correlation	Coverage	Correlation	Coverage
All	0.790,**	21%	0.136, ns	41%
No label	0.745, *	19%	0.074, ns	45%
No sense	0.657, *	21%	0.136, ns	41%
No arg head	0.790,**	21%	0.136, ns	42%
No path	0.414, ns	21%	–	2%
No path-label	0.711,*	19%	0.074, ns	45%

ns: not significant, *: $p < 0.05$, **: $p < 0.01$

Observations

- FN classes have relatively low coverage, but significant correlations

Importance of Features for Induced Classes

	FN		PB	
Features	Correlation	Coverage	Correlation	Coverage
All	0.790,**	21%	0.136, ns	41%
No label	0.745, *	19%	0.074, ns	45%
No sense	0.657, *	21%	0.136, ns	41%
No arg head	0.790,**	21%	0.136, ns	42%
No path	0.414, ns	21%	–	2%
No path-label	0.711,*	19%	0.074, ns	45%

ns: not significant, *: $p < 0.05$, **: $p < 0.01$

Observations

- Syntactic information (path feature): Strong influence for both sets, absence is catastrophic for PB (eqivalent to no classes)

Importance of Features for Induced Classes

	FN		PB	
Features	Correlation	Coverage	Correlation	Coverage
All	0.790,**	21%	0.136, ns	41%
No label	0.745, *	19%	0.074, ns	45%
No sense	0.657, *	21%	0.136, ns	41%
No arg head	0.790,**	21%	0.136, ns	42%
No path	0.414, ns	21%	–	2%
No path-label	0.711,*	19%	0.074, ns	45%

ns: not significant, *: $p < 0.05$, **: $p < 0.01$

Observations

- Filler information (argument head feature): Meaningless (too sparse?) for both sets

Importance of Features for Induced Classes

	FN		PB	
Features	Correlation	Coverage	Correlation	Coverage
All	0.790,**	21%	0.136, ns	41%
No label	0.745, *	19%	0.074, ns	45%
No sense	0.657, *	21%	0.136, ns	41%
No arg head	0.790,**	21%	0.136, ns	42%
No path	0.414, ns	21%	–	2%
No path-label	0.711,*	19%	0.074, ns	45%

ns: not significant, *: $p < 0.05$, **: $p < 0.01$

Observations

- Semantic information: Label, path-label, verb sense important for FN

Importance of Features for Induced Classes

	FN		PB	
Features	Correlation	Coverage	Correlation	Coverage
All	0.790,**	21%	0.136, ns	41%
No label	0.745, *	19%	0.074, ns	45%
No sense	0.657, *	21%	0.136, ns	41%
No arg head	0.790,**	21%	0.136, ns	42%
No path	0.414, ns	21%	–	2%
No path-label	0.711,*	19%	0.074, ns	45%

ns: not significant, *: $p < 0.05$, **: $p < 0.01$

Observations

- Semantic information: Label, path-label, verb sense important for FN
- Label, path-label important for PB – but not informative enough to allow class formation if path information is missing

Importance of Features for Induced Classes

	FN		PB	
Features	Correlation	Coverage	Correlation	Coverage
All	0.790,**	21%	0.136, ns	41%
No label	0.745, *	19%	0.074, ns	45%
No sense	0.657, *	21%	0.136, ns	41%
No arg head	0.790,**	21%	0.136, ns	42%
No path	0.414, ns	21%	–	2%
No path-label	0.711,*	19%	0.074, ns	45%

ns: not significant, *: $p < 0.05$, **: $p < 0.01$

Observations

- FN clustering profits from semantic annotation
- PB annotation yields clusters with much lower smoothing power, clustering relies heavily on syntactic feature

Outline

Conclusions

- Introduced a model of human semantic intuitions and demonstrated the reliability of its predictions
- Gave example of how semantic effects in human sentence processing can be modelled using the semantic model
- Investigated why FrameNet is better training data for us than PropBank:
  - General reason: Verb clusters are more useful due to semantic content of annotation
  - Specific to task: Vocabulary/events in BNC are closer to test data