Automatic Labeling of Semantic Roles, Gildea and Jurafsky, CL (2002)

Amanda Tollafield-Small

Introduction

The paper presents a system for identifying the semantic roles, filled by constituents of a sentence within a frame.

When given a sentence, target word and frame, the system labels constituents with either abstract roles such as AGENT or PATIENT, or more domain-specific roles such as SPEAKER, MESSAGE, and TOPIC.

*"A frame is a schematic representation of situations involving various participants, props, and other conceptual roles"

Previous systems

- Previous systems were based on domain-specific templates. For example:

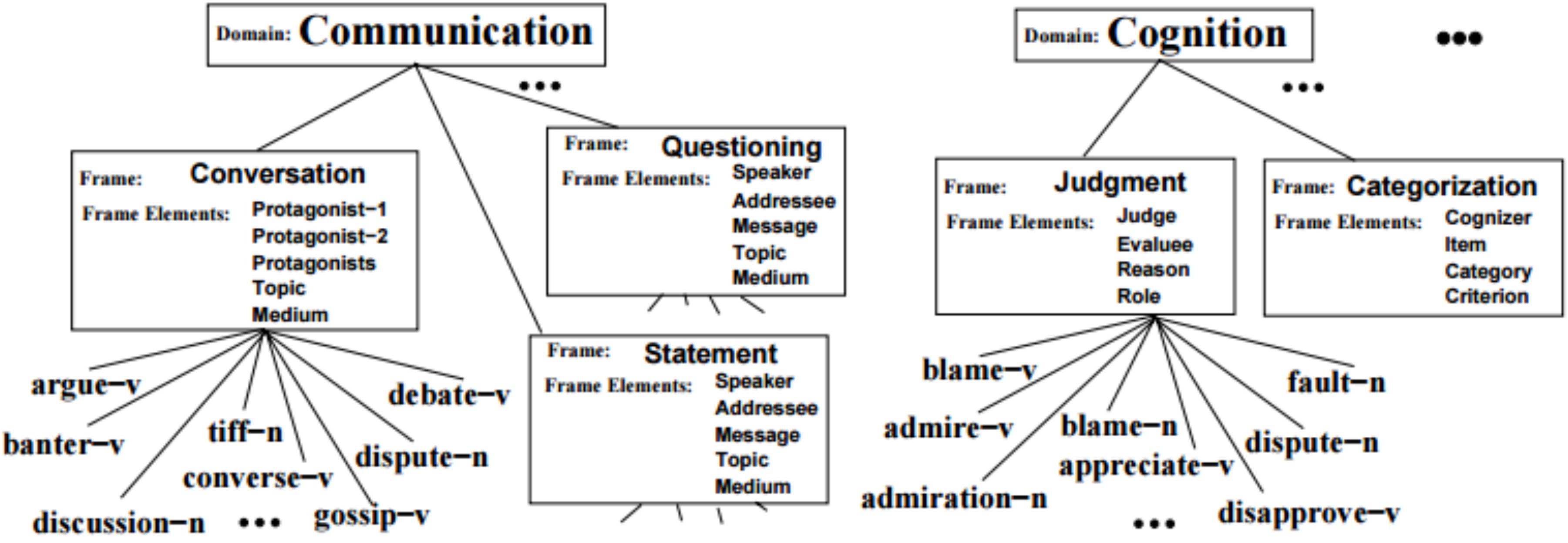
ORIG_CITY, DEST_CITY, DEPART_TIME, PRODUCTS, RELATIONSHIP, JOINT_VENTURE_COMPANY or TO_AIRPORT

- A less specific system, such as the one proposed by Gildea and Jurafsky, is more efficient at generalising information extraction, question answering, semantic dialogue systems, and word-sense disambiguation.

The System

- The statistical algorithms were trained on a hand-labelled dataset: the FrameNet database (Baker, Fillmore, and Lowe, 1998; Johnson et al., 2001).
- The FrameNet database defines a set of semantic roles called frame elements.
- 50,000 sentences from the British National Corpus hand-labelled

FrameNet Example

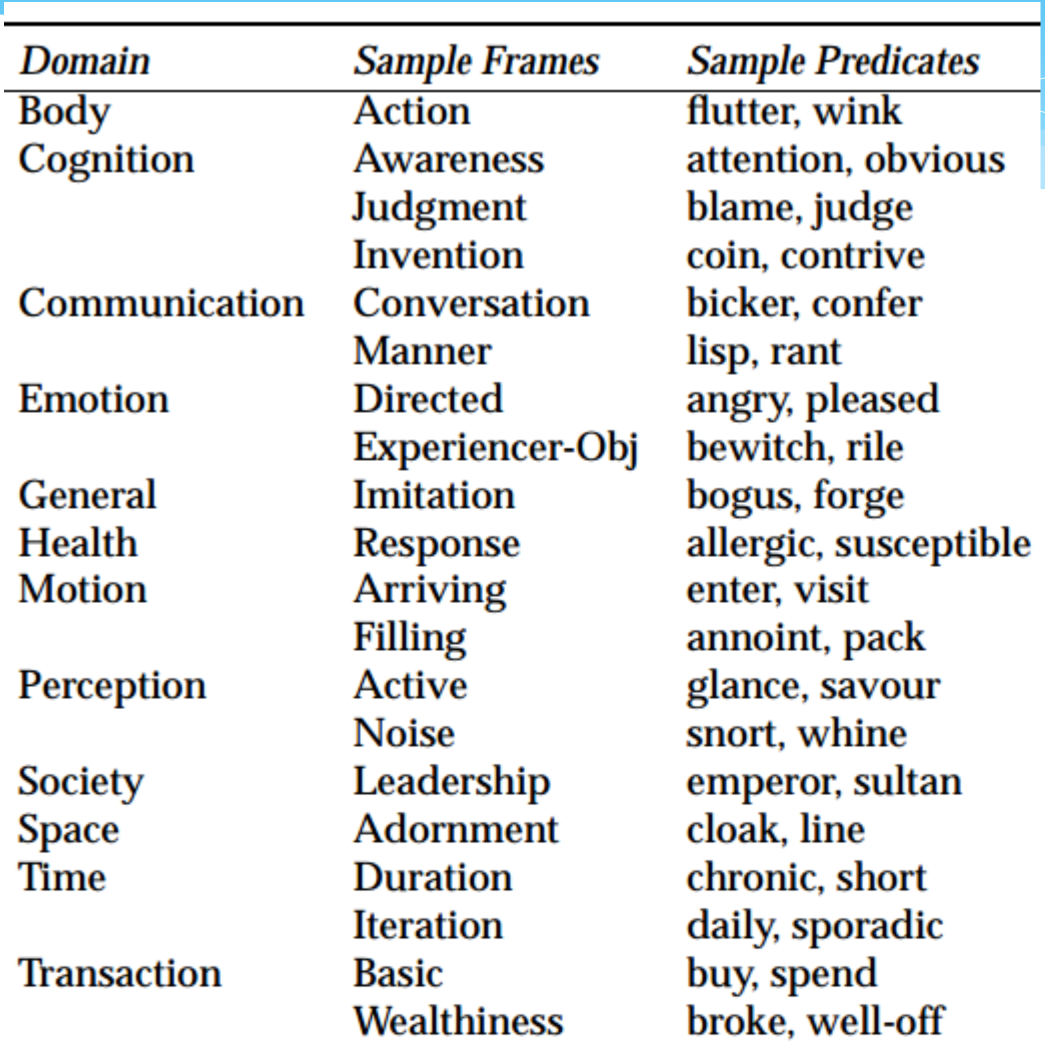


FrameNet Example

Frame: Judgement

- [*Judge* She] **blames** [*Evaluee* the Government] [*Reason* for failing to do enough to help] .
- Holman would characterise this as **blaming** [*Evaluee* the poor] .
- The letter quotes Black as saying that [*Judge* white and Navajo ranchers] misrepresent their livestock losses and **blame** [*Reason* everything] [Evaluee on coyotes] .

Hand-annotation examples



Domain	*Sample Frames*	*Sample Predicates*
Body	Action	flutter, wink
Cognition	Awareness	attention, obvious
	Judgment	blame, judge
	Invention	coin, contrive
Communication	Conversation	bicker, confer
	Manner	lisp, rant
Emotion	Directed	angry, pleased
	Experiencer-Obj	bewitch, rile
General	Imitation	bogus, forge
Health	Response	allergic, susceptible
Motion	Arriving	enter, visit
	Filling	annoint, pack
Perception	Active	glance, savour
	Noise	snort, whine
Society	Leadership	emperor, sultan
Space	Adornment	cloak, line
Time	Duration	chronic, short
	Iteration	daily, sporadic
Transaction	Basic	buy, spend
	Wealthiness	broke, well-off

Performance

- Overall performance was 82.1% compared to 80.4% for frame-specific roles.

Automatic Labelling

- The system is trained by first using an automatic syntactic parser to analyse the training sentences. It matches annotated frame elements to constituents, and extracts various features from the string of words and the parse tree.

The Features

- Phrase Type
- Governing Category
- Parse Tree Path
- Position
- Voice
- Head Word

Probability Estimation

- *r* indicates semantic role, *pt* phrase type, *gov* grammatical function, *h* head word, and *t* target word, or predicate.
- Probability distribution which, given the features, indicates the probability of each semantic role:

$$P(r|h, pt, \mathbf{gov}, position, voice, t)$$

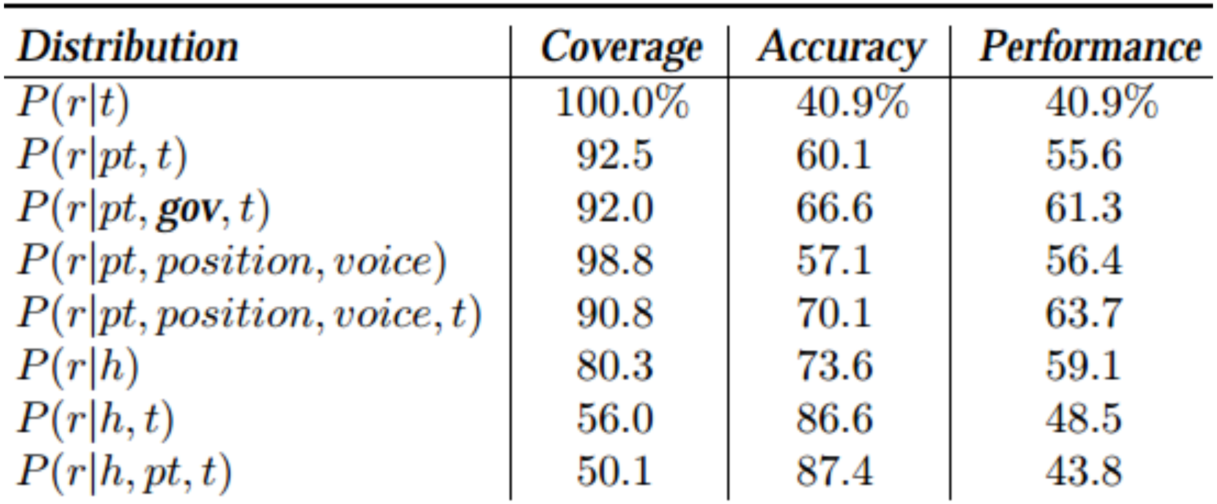
Probability Estimation …

- The distribution can be calculated from the training data using the frequency of the combination of features and the frequency of the combination with a certain role.

$$P(r|h, pt, gov, position, voice, t) = \frac{\#(r, h, pt, \mathbf{gov}, position, voice, t)}{\#(h, pt, \mathbf{gov}, position, voice, t)}$$

Distributions

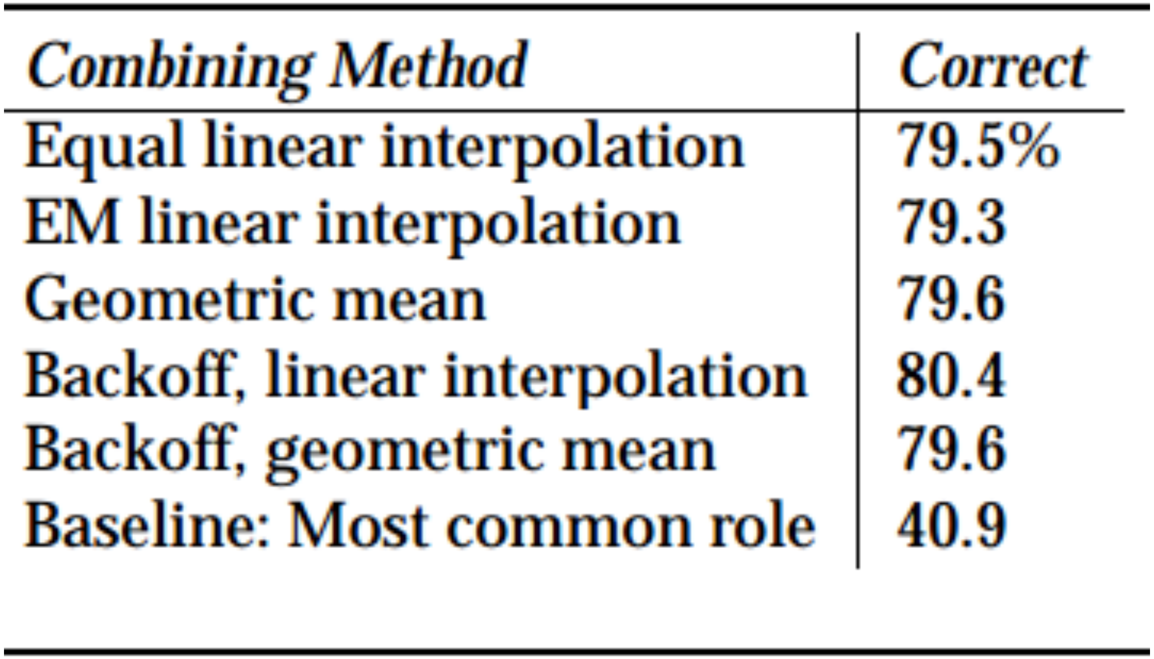
- *r* indicates semantic role, *pt* phrase type, *gov* grammatical function, *h* head word, and *t* target word, or predicate.



Distribution	***Coverage***	***Accuracy***	***Performance***
$P(r\|t)$	100.0%	40.9%	40.9%
$P(r\|pt,t)$	92.5	60.1	55.6
$P(r\|pt,\mathbf{gov},t)$	92.0	66.6	61.3
$P(r\|pt,position,voice)$	98.8	57.1	56.4
$P(r\|pt,position,voice,t)$	90.8	70.1	63.7
$P(r\|h)$	80.3	73.6	59.1
$P(r\|h,t)$	56.0	86.6	48.5
$P(r\|h,pt,t)$	50.1	87.4	43.8

Combining Methods

$$P(r|h, pt, \mathbf{gov}, position, voice, t)$$



Combining Method	*Correct*
Equal linear interpolation	79.5%
EM linear interpolation	79.3
Geometric mean	79.6
Backoff, linear interpolation	80.4
Backoff, geometric mean	79.6
Baseline: Most common role	40.9

Examples: Linear Interpolation & Geometric Mean

$$\begin{aligned} P(r|constituent) &= \lambda_1 P(r|t) + \lambda_2 P(r|pt,t) + \\ &\quad \lambda_3 P(r|pt,\mathbf{gov},t) + \lambda_4 P(r|pt,position,voice) + \\ &\quad \lambda_5 P(r|pt,position,voice,t) + \lambda_6 P(r|h) + \\ &\quad \lambda_7 P(r|h,t) + \lambda_8 P(r|h,pt,t) \end{aligned}$$

where *∑i↑▒λ = 1*

$$\begin{aligned} P(r|constituent) &= \tfrac{1}{Z} exp\{ \quad \lambda_1 logP(r|t) + \lambda_2 logP(r|pt,t) + \\ &\quad \lambda_3 logP(r|pt,\mathbf{gov},t) + \lambda_4 logP(r|pt,position,voice) + \\ &\quad \lambda_5 logP(r|pt,position,voice,t) + \lambda_6 logP(r|h) + \\ &\quad \lambda_7 logP(r|h,t) + \lambda_8 logP(r|h,pt,t) \quad \} \end{aligned}$$

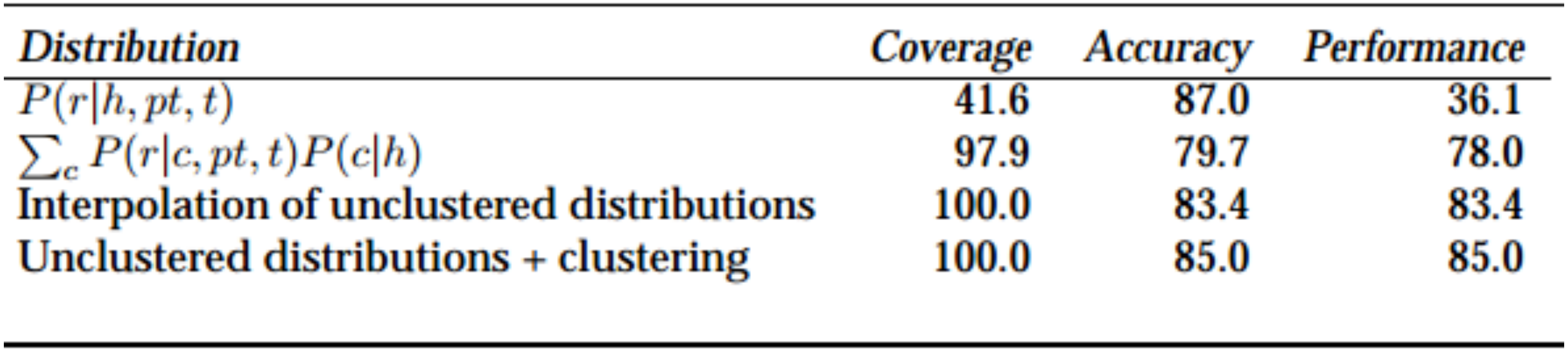
where *Z* is a normalising constant for *∑r↑▒P(r|constituent)* = 1

Generalising Lexical Statistics

- Automatic Clustering
- Semantic Hierarchy (WordNet)
- Bootstrapping

Automatic Clustering

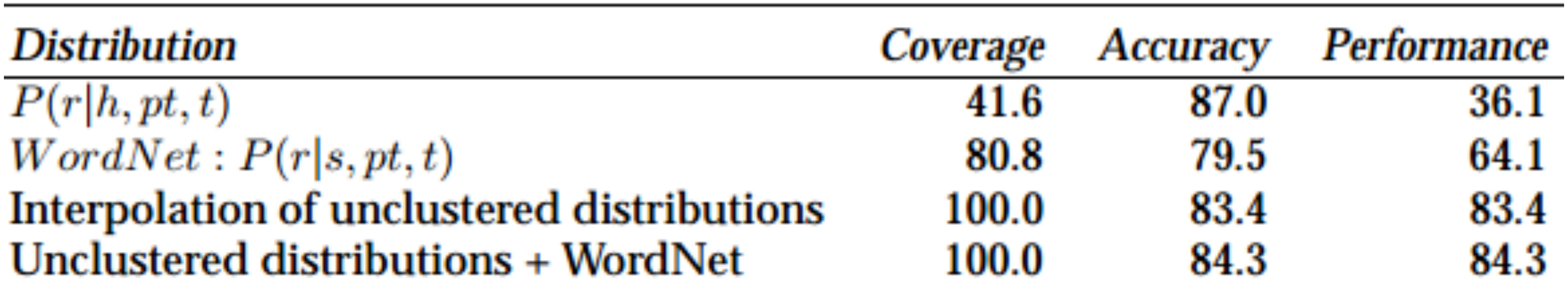
- This technique is based on the expectation that words with similar semantics will tend to be present alongside each other. This expectation was used to as a probabilistic model.



Distribution	*Coverage*	*Accuracy*	*Performance*
$P(r\|h, pt, t)$	41.6	87.0	36.1
$\sum_c P(r\|c, pt, t)P(c\|h)$	97.9	79.7	78.0
Interpolation of unclustered distributions	100.0	83.4	83.4
Unclustered distributions + clustering	100.0	85.0	85.0

Semantic Hierarchy (WordNet)

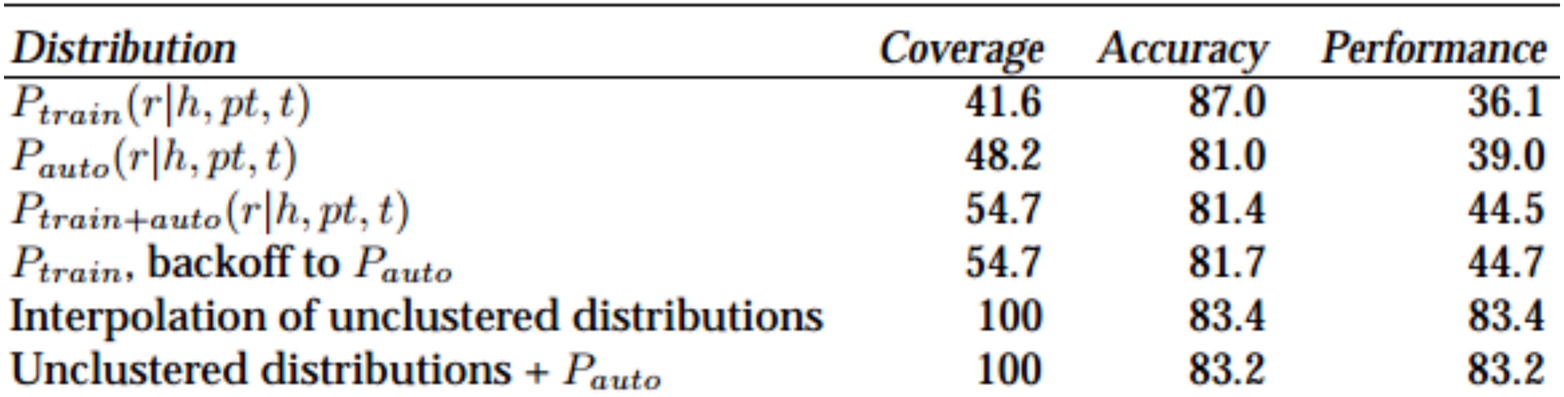
- When a head word that was not seen in the training examples is presented, the hierarchy is ascended until a level with data is found.



Distribution	*Coverage*	*Accuracy*	*Performance*
$P(r\|h, pt, t)$	41.6	87.0	36.1
$WordNet : P(r\|s, pt, t)$	80.8	79.5	64.1
Interpolation of unclustered distributions	100.0	83.4	83.4
Unclustered distributions + WordNet	100.0	84.3	84.3

Bootstrapping

- Use the automatic labelling system to label unannotated data and use the imperfect result as further training data.



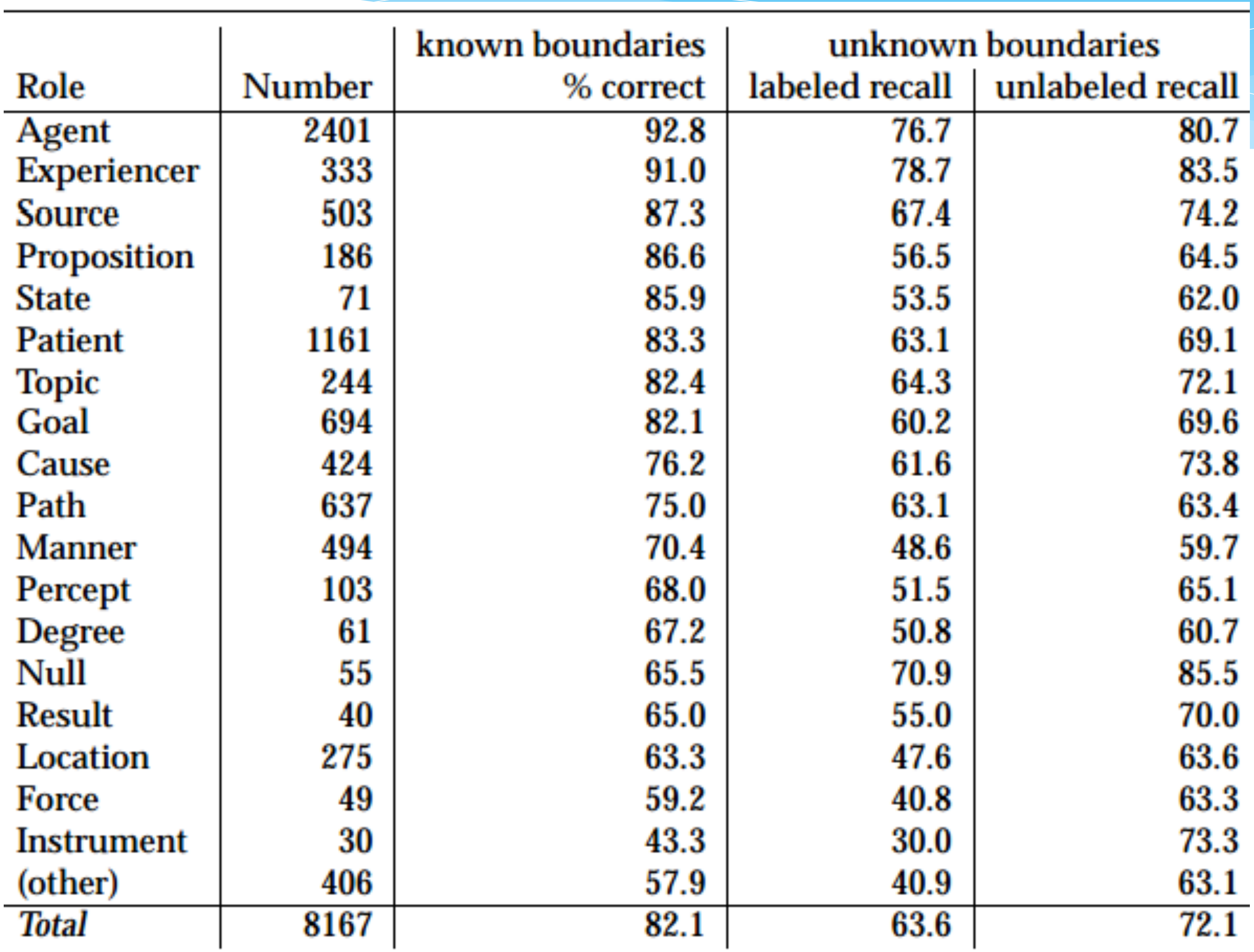
Distribution	*Coverage*	*Accuracy*	*Performance*
$P_{train}(r\|h, pt, t)$	41.6	87.0	36.1
$P_{auto}(r\|h, pt, t)$	48.2	81.0	39.0
$P_{train+auto}(r\|h, pt, t)$	54.7	81.4	44.5
P_{train}, backoff to P_{auto}	54.7	81.7	44.7
Interpolation of unclustered distributions	100	83.4	83.4
Unclustered distributions + P_{auto}	100	83.2	83.2

Generalising Lexical Statistics Comparison

- The differences in the coverage each method provides causes the results.
- The automatic clustering method performed the best.
- The bootstrapping technique made use of much less data than automatic clustering.
- The WordNet shows how difficult it can be to get broad coverage with hand-annotated samples but that they are very useful when they can be applied.

More abstract

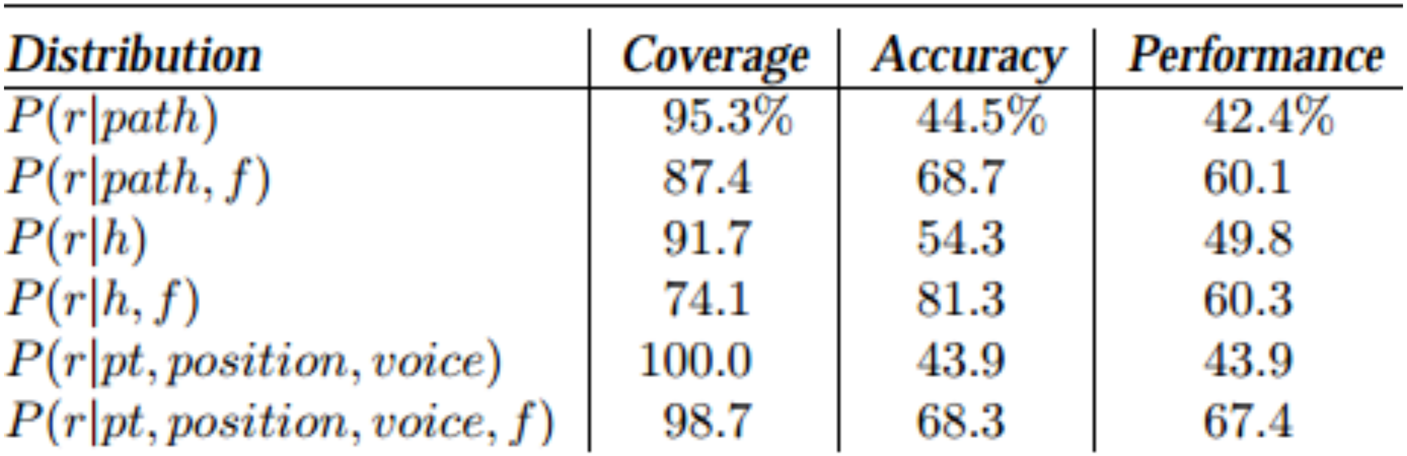
- Performance broken down by abstract role.



		known boundaries	unknown boundaries	
Role	Number	% correct	labeled recall	unlabeled recall
Agent	2401	92.8	76.7	80.7
Experiencer	333	91.0	78.7	83.5
Source	503	87.3	67.4	74.2
Proposition	186	86.6	56.5	64.5
State	71	85.9	53.5	62.0
Patient	1161	83.3	63.1	69.1
Topic	244	82.4	64.3	72.1
Goal	694	82.1	60.2	69.6
Cause	424	76.2	61.6	73.8
Path	637	75.0	63.1	63.4
Manner	494	70.4	48.6	59.7
Percept	103	68.0	51.5	65.1
Degree	61	67.2	50.8	60.7
Null	55	65.5	70.9	85.5
Result	40	65.0	55.0	70.0
Location	275	63.3	47.6	63.6
Force	49	59.2	40.8	63.3
Instrument	30	43.3	30.0	73.3
(other)	406	57.9	40.9	63.1
Total	8167	82.1	63.6	72.1

Cross-frame performance

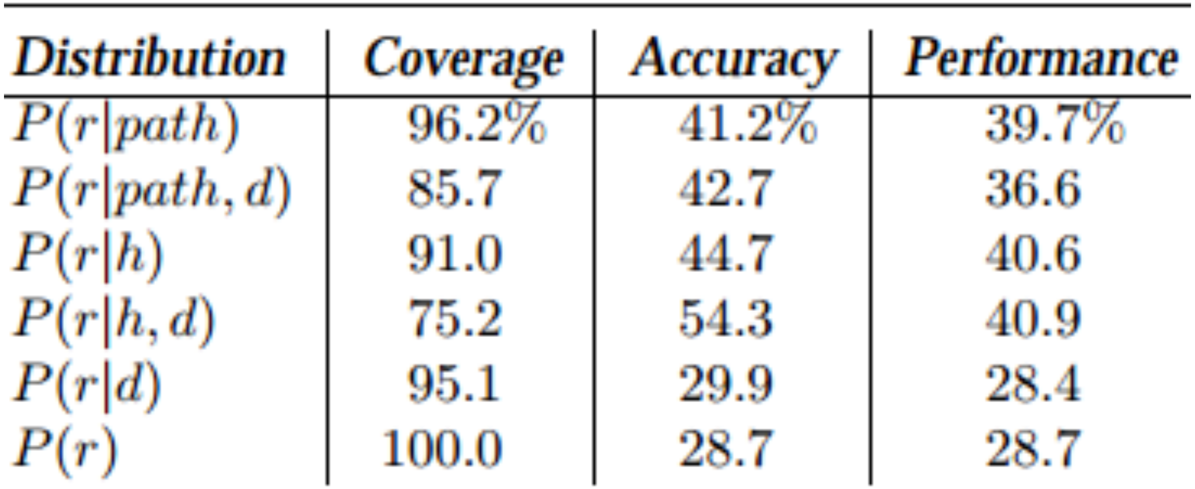
- f represents the FrameNet semantic frame.



Distribution	*Coverage*	*Accuracy*	*Performance*
$P(r\|path)$	95.3%	44.5%	42.4%
$P(r\|path, f)$	87.4	68.7	60.1
$P(r\|h)$	91.7	54.3	49.8
$P(r\|h, f)$	74.1	81.3	60.3
$P(r\|pt, position, voice)$	100.0	43.9	43.9
$P(r\|pt, position, voice, f)$	98.7	68.3	67.4

Cross-frame Performance

- d represents the FrameNet semantic domain.



Distribution	Coverage	Accuracy	Performance
$P(r\|path)$	96.2%	41.2%	39.7%
$P(r\|path, d)$	85.7	42.7	36.6
$P(r\|h)$	91.0	44.7	40.6
$P(r\|h, d)$	75.2	54.3	40.9
$P(r\|d)$	95.1	29.9	28.4
$P(r)$	100.0	28.7	28.7

Conclusion

- The system is able to automatically label semantic roles with reasonably high accuracy.
- After testing different methods to generalise lexical statistics; the coverage of automatic clustering outweighed its imprecision.

References

- Automatic Labeling of Semantic Roles, Gildea and Jurafsky, CL (2002)
- FrameNet database (Baker, Fillmore, and Lowe, 1998; Johnson et al., 2001).
- Marcus, Santorini, and Marcinkiewicz (1993