Automatic Labeling of Semantic Roles

Daniel Gildea & Daniel Jurafsky

Task

- Train using automatically derived parses
- Identifying Frame Element Boundaries
  - Use parse tree to extract features
  - Find Probability of constituent being an argument
- Labeling Elements
  - Argument labels are specific to the frame and the predicate

Frame:	**Statement**
Frame Elements:	Speaker
	Addressee
	Message
	Topic
	Medium

Labeling Frame Elements, Features

- Head
- Target: The predicate in the sentence
- Phrase Type: NP, VP, PP, S etc.
- Grammatical Function: Subject/Object (only applied to NP)
- Voice: Active/Passive
- Position: before/after predicate

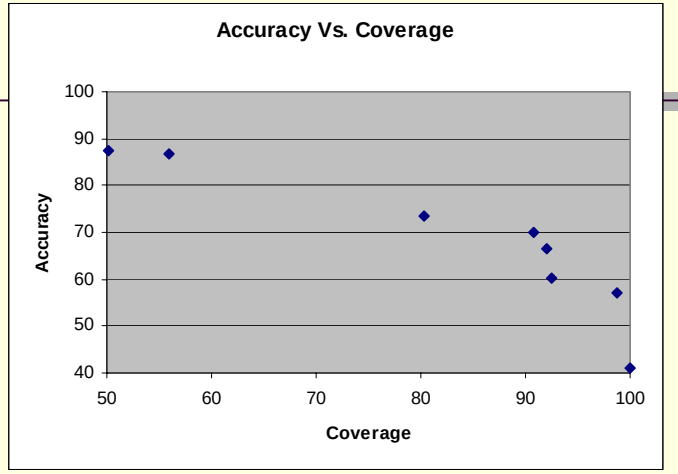
Methodology

- Parse training set using Collins (1997) parser, to extract features
- Find most probable assignment of roles r given features
- P(r | pt, gf, voice, pos, h, t) is sparse, so approximate using reduced feature sets

Reduced Feature Sets

- Coverage: % of cases where feature set is available
- Accuracy: % of correct labelings where feature set is available
- Performance: Product of Coverage and Accuracy

Distribution	*Coverage*	*Accuracy*	*Performance*
$P(r\|t)$	100%	40.9%	40.9%
$P(r\|pt,t)$	92.5	60.1	55.6
$P(r\|pt,gf,t)$	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



No feature set beats any other in both accuracy and coverage.

Combining Probabilities

- Development Set
  - All methods are within the margin of error (1%)
- Test Set
  - Performance is less by 3.5%
  - Is the margin of error really 1% ?

Combining Method	*Correct*
Linear Interpolation	79.5%
Geometric Mean	79.6
Backoff, linear interpolation	80.4
Backoff, geometric mean	79.6
Baseline: Most common role	40.9

	Linear Backoff	*Baseline*
Development Set	80.4%	40.9%
Test Set	76.9	40.6%

Grammatical Function Vs. Position

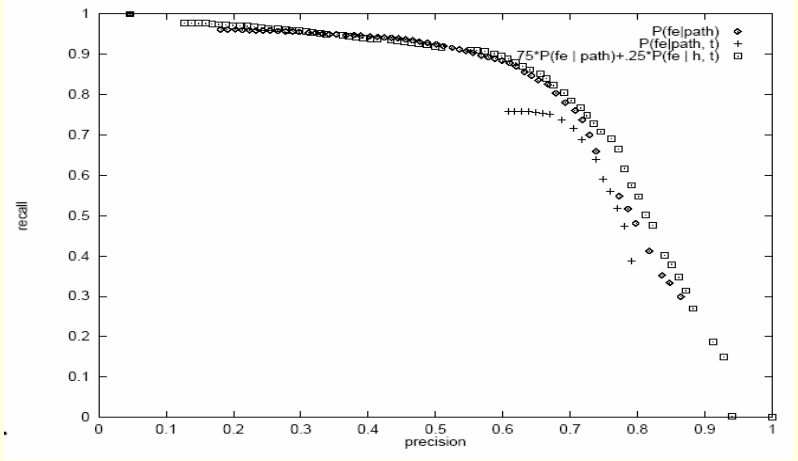
Distribution	*Coverage*	*Accuracy*	*Performance*
$P(r\|pt,gf,t)$	92.0	66.6	61.3
$P(r\|pt,position,voice,t)$	90.8	70.1	63.7
$P(r\|pt,t)$	92.5	60.1	55.6

Feature Set	Performance
gf, h, pt, t	79.2
pos, voice, h, pt, t	80.5
h, pt, t	76.3

Identifying Frame Elements

- Features
  - Head
  - Path
  - Target
- Feature Sets for Obtaining Probabilities
  - Path
  - Path, Target
  - Head, Target

Performance, max F~.73



Partial Overlap

Type of Overlap	*Identified Constituents*	*Number*
Exactly Matching Boundaries	66%	5421
Identified constituent entirely within true frame element	8	663
True frame element entirely within identified constituent	7	599
Partial overlap	0	26
No match to true frame element	13	972

- Allowing one inside the other increases precision by 15%
- Perhaps they should have included the fraction of true frame elements which were not identified as a constituent by the parser.

Shallow Semantic Parsing Using Support Vector Machines

Dan Jurafsky, Smaeer Pradhan, Wayne Ward, Kadri Hacioglu, and James Martin

PropBank Arguments

- Differs from FrameNet
  - Arguments NOT specific to frame
  - More training data/applicability
- Core Arguments:
  - proto-Agent
  - proto-Patient, etc.
- Adjunctive Arguments
  - Location
  - Temporal

Tasks

- Argument Identification:
  Use single SVM to distinguish between nulls and non-nulls
- Argument Classification
  Use an SVM for each argument classification, and pick one with greatest confidence
- Identification & Classification

Old Features (G&J 2002)

- Predicate (or Target)
- Path
- Phrase Type
- Position
- Voice
- Head Word
- Verb Subcatagorization
  - CFG rule used to expand parent of verb

Baseline Performance (old features)

Classes	Task	P (%)	R (%)	F_1	A (%)
ALL ARGs	Id.	90.9	89.8	90.4	
	Classification	-	-	-	87.9
	Id. + Classification	83.3	78.5	80.8	
CORE ARGs	Id.	94.7	90.1	92.3	
	Classification	-	-	-	91.4
	Id. + Classification	88.4	84.1	86.2	

New Features

- Named Entities
- Head Word POS
- Verb Clustering
- Partial Path
- Verb sense information (fine)
- Modified Head of prep. Phrases
- First & Last Word/POS in Constituent
- Ordinal constituent position (concat. w/type)
- Constituent Tree Distance
- Constituent Relative Features

New Features

- Improvement in classification and identification from each feature
- Most features improve at least one task above level of significance

Features	Class Acc.	ARGUMENT ID		
		P	R	F_1
Baseline	87.9	93.7	88.9	91.3
+ Named entities	88.1	-	-	-
+ Head POS	*88.6	94.4	90.1	*92.2
+ Verb cluster	88.1	94.1	89.0	91.5
+ Partial path	88.2	93.3	88.9	91.1
+ Verb sense	88.1	93.7	89.5	91.5
+ Noun head PP (only POS)	*88.6	94.4	90.0	*92.2
+ Noun head PP (only head)	*89.8	94.0	89.4	91.7
+ Noun head PP (both)	*89.9	94.7	90.5	*92.6
+ First word in constituent	*89.0	94.4	91.1	*92.7
+ Last word in constituent	*89.4	93.8	89.4	91.6
+ First POS in constituent	88.4	94.4	90.6	*92.5
+ Last POS in constituent	88.3	93.6	89.1	91.3
+ Ordinal const. pos. concat.	87.7	93.7	89.2	91.4
+ Const. tree distance	88.0	93.7	89.5	91.5
+ Parent constituent	87.9	94.2	90.2	*92.2
+ Parent head	85.8	94.2	90.5	*92.3
+ Parent head POS	*88.5	94.3	90.3	*92.3
+ Right sibling constituent	87.9	94.0	89.9	91.9
+ Right sibling head	87.9	94.4	89.9	*92.1
+ Right sibling head POS	88.1	94.1	89.9	92.0
+ Left sibling constituent	*88.6	93.6	89.6	91.6
+ Left sibling head	86.9	93.9	86.1	89.9
+ Left sibling head POS	*88.8	93.5	89.3	91.4
+ Temporal cue words	*88.6	-	-	-
+ Dynamic class context	88.4	-	-	-

Improvements in F Score, (identification & tagging)

- Disallow Overlaps
  - Choose constituent with greater confidence
  - +0.8% on all arguments
- Argument sequence information + no overlaps
  - Predicate specific trigram model over argument types
  - +2% on core arguments
- Using all new features with significant improvement + no overlaps + sequence info
  - All new features with significant improvement
  - +5.9% on all arguments, +2.7% on core arguments

Best Performance

(all significant improvements)

Classes	Task	Hand-corrected parses			
		P (%)	R (%)	F_1	A (%)
ALL ARGS	Id.	95.2	92.5	93.8	
	Classification	-	-	-	91.0
	Id. + Classification	88.9	84.6	86.7	
CORE ARGS	Id.	96.2	93.0	94.6	
	Classification	-	-	-	93.9
	Id. + Classification	90.5	87.4	88.9	

- Improvement
  - 2-3% in classification
  - 2-4% in identification
  - 3-6% in combined task

Other Comparisons
(all arguments)

- Automatic Parses
  - -7.8% in identification (F score)
  - -1% in tagging (accuracy)
  - -7.3% in id+tagging (F score)
- New Data, hand tagged
  - +2.2% in id
  - +2% in tagging
  - +2.7% in id+tagging
- Different Corpus
  - AQUAINT, NYTimes (similar to WSJ)
  - -20.3% in identification!
  - -7.2% in tagging
  - -23.4% in id+tagging!

Comparison of Coverage

	Propbank		Aquaint	
	Args	Non-args	Args	Non-args
Predicate & Path	87.60	2.91	62.11	4.66
Predicate & Head	48.9	26.55	30.26	17.41

Comparison to Other Systems

- Gildea & Palmer (2002)
  - Probabilities with back-off and interpolation
- Surdeanu et. al. (I, II)
  - Decision Tree
- Gildea & Hockenmaier
- Chen & Rambow (I, II)
  - Decision Tree

Classification

Classifier	Accuracy (%)
SVM	88
Decision Tree (Surdeanu et al., 2003)	79
Gildea and Palmer (2002)	77

Table 11: Argument classification using same features but different classifiers.

Classes	System	Hand Accuracy	Automatic Accuracy
ALL ARGS	SVM	91	90
	G&P	77	74
	Surdeanu System II	84	-
	Surdeanu System I	79	-
CORE ARGS	SVM	93.9	90.5
	C&R System II	93.5	-
	C&R System I	92.4	-

Table 13: Argument classification

Identification

Classes	System	Hand			Automatic		
		P	R	F_1	P	R	F_1
ALL ARGs	SVM	95	92	94	89	83	86
	Surdeanu System II	-	-	89	-	-	-
	Surdeanu System I	85	84	85	-	-	-

Table 12: Argument identification

Both Tasks

Classes	System	Hand			Automatic		
		P	R	F_1	P	R	F_1
ALL ARGs	SVM	89	85	87	84	75	79
	G&H System I	76	68	72	71	63	67
	G&P	71	64	67	58	50	54
CORE ARGs	SVM System	90	87	89	86	78	82
	G&H System I	82	79	80	76	73	75
	C&R System II	-	-	-	65	75	70

Table 14: Identification and classification

Feature Analysis

Features	Accuracy (%)
All	91.0
All except Path	90.8
All except PhraseType	90.8
All except HW and HW-POS	90.7
All except All Phrases	*83.6
All except Predicate	*82.4
All except HW and FW and LW-POS	*75.1
Path, Predicate	74.4
Path, PhraseType	47.2
HeadWord	37.7
Path	28.0

Table 9: Performance of various feature combinations on the task of argument classification.

Conclusion

- Total system performs well
  - Final F score w/automatic parses on combined task: 79.4%
  - SVM's work well
  - Lots of promising new features
  - Beats others
- Still need to work on
  - Analyzing features and feature "families"