Automatic Semantic Role Labeling

- Semantic Role Labeling (SRL) systems automatically assign semantic roles to phrases.
- Both FrameNet and PropBank data are frequently used for training and evaluation.
- SRL has also been a shared task for CoNLL 2004 & 2005, Senseval-3, and SemEval-2007.
- There has been a lot of activity in this area, though not much use of SRL systems for downstream applications yet. In principle, many applications should be able to benefit from it though.

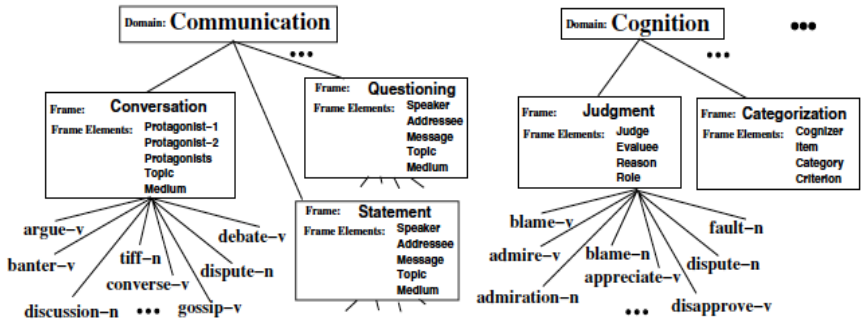
Common Architecture

1. **Argument Identification:** identify a set of candidate argument phrases. This is primarily a syntactic task.
   - nearly any sequence of words can be an argument, so heuristics may be used to identify the most likely candidates.
2. **Argument Classification:** for each candidate argument, label it with a semantic role or with a No Argument label.
3. **Global Scoring** may be done to make sure the argument assignments are globally plausible.
   - For example, check that arguments don't overlap, each core argument does not repeat, etc.

Statistical SRL

- [Gildea & Jurafsky, 2002] created an early influential system for semantic role labeling with statistical modeling.
- G&J used FrameNet data for training and testing.
- Syntactic features were generated from a parser.
- Statistical models were then used to learn how to assign semantic roles.
- They systematically explored subtasks and issues related to accurate and generalizable semantic role labeling.

Sample FrameNet Domains & Frames



Sample Domains, Frames, & Predicates

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

FrameNet Data

G&J used FrameNet data:

- 49,013 annotated sentences from the British National Corpus
- 67 frames from 12 general semantic "domains" such as motion, cognition, and communication
- 1,462 target words (predicates): 927 verbs, 339 nouns, 175 adj
- 99,232 annotated frame elements
- FrameNet's sentences were not chosen randomly, but selected based on target words to illustrate typical uses. So the annotated data is not necessarily statistically representative!

(*Now*, FrameNet has > 170,000 annotated sentences.)

General Approach

- Task 1: identify boundaries of frame elements (argument identification)
- Task 2: label each FE with its semantic role
- First, they focus on Task 2 and evaluate results using gold annotated FEs.
- Then, they automate Task 1 and evaluate the accuracy of system-generated FEs.
- They use statistical models based on probabilities estimated from the training examples.

Features

- **Phrase Type:** the constituent type (non-terminal) of the FE in the parse tree, such as NP, VP, S.

  In their training data, the breakdown of FEs was:

  NPs: 47% PPs: 22% ADVPs: 4% S/SBARs: 4% PRTs: 2%

  For training, FEs that didn't match a constituent were discarded.

  For testing, the largest constituent matching the FE's left boundary and lying entirely within the FE was used.

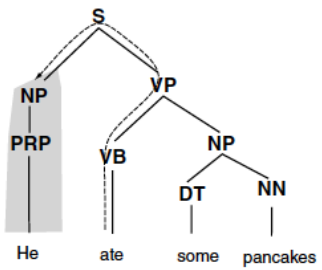
- **Grammatical Function / Governing Category:** the syntactic role of the FE's constituent in the sentence (only applies to NPs).

  gov=S → Subect *gov*=VP → Object

Features

- **Parse Tree Path:** the path from the target word to the candidate constituent through the parse tree.

Example:



The path from "ate" (target) to "He" (candidate) would be:

VB↑ VP↑ S↓NP

Most Frequent Paths

Frequency	Path	Description
14.2%	VB↑VP↓PP	PP argument/adjunct
11.8	VB↑VP↑S↓NP	Subject
10.1	VB↑VP↓NP	Object
7.9	VB↑VP↑VP↑S↓NP	Subject (embedded VP)
4.1	VB↑VP↓ADVP	Adverbial adjunct
3.0	NN↑NP↑NP↓PP	Prepositional complement of noun
1.7	VB↑VP↓PRT	Adverbial particle
1.6	VB↑VP↑VP↑VP↑S↓NP	Subject (embedded VP)
14.2		No matching parse constituent
31.4	Other	

Features

- **Position:** does the FE occur before or after the target.
  - Subjects usually occur before, Objects after, so this heuristically serves as a secondary check to identify grammatical function.
- **Voice:** does the verb occur in active or passive voice?
  Ten patterns were used to identify passive voice.
  - Only ~5% of the examples were identified as passive voice!
- **Head Word:** the lexical head word of the FE phrase, as determined by the parser.
  - the head of a PP is the preposition
  - complementizers are heads, such as "to" for infinitive VPs and "that" for subordinate clauses.

Evaluation

- For each target word, 10% of the annotated sentences were reserved for testing, another 10% set aside for tuning, and the remaining 80% used for training.
- Target words with < 10 instances were discarded.
- The average number of sentences for each target = 34.
  The average number of sentences for each frame = 732.
- Due to the small amount of data, probabilities were generated for small sets of features, rather than all of them combined.

The Statistical Model

- Ideally, the goal is to estimate the probability that a constituent fills a semantic role based on the features:

$$P(role \mid head, phrase, gov, position, voice, target) = \frac{\#(role, head, phrase, gov, position, voice, target)}{\#(head, phrase, gov, position, voice, target)}$$

- But … the data set is relatively small, so most combinations of these features will have been seen few times, if at all. The target word and head feature in particular are very specific.
- Solution: combine probabilities for subsets of features.

Probability Distributions in Corpus

Table 3
Distributions calculated for semantic role identification: r indicates semantic role, pt phrase type, gov grammatical function, h head word, and t target word, or predicate.

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

Coverage is % of test data for which conditioned event has been seen.
Accuracy is proportion of covered test data for which highest probability role is correct.
Performance is % of data for which predicted role is correct (Coverage * Accuracy)

Sample Counts & Probabilities

Sample probabilities for $P(r \mid pt, gov, t)$ calculated from training data for the verb *abduct*. The variable gov is defined only for noun phrases. The roles defined for the *removing* frame in the *motion* domain are AGENT (AGT), THEME (THM), COTHEME (COTHM) ("… had been abducted *with him*"), and MANNER (MANR).

$P(r \mid pt, gov, t)$	Count in training data
$P(r = \text{AGT} \mid pt = \text{NP}, gov = \text{S}, t = \text{abduct}) = .46$	6
$P(r = \text{THM} \mid pt = \text{NP}, gov = \text{S}, t = \text{abduct}) = .54$	7
$P(r = \text{THM} \mid pt = \text{NP}, gov = \text{VP}, t = \text{abduct}) = 1$	9
$P(r = \text{AGT} \mid pt = \text{PP}, t = \text{abduct}) = .33$	1
$P(r = \text{THM} \mid pt = \text{PP}, t = \text{abduct}) = .33$	1
$P(r = \text{COTHM} \mid pt = \text{PP}, t = \text{abduct}) = .33$	1
$P(r = \text{MANR} \mid pt = \text{ADVP}, t = \text{abduct}) = 1$	1

Combining Probabilities

- They merged the feature subset probabilities to obtain estimates of the full distribution in several ways.
- Linear Interpolation averaging the probabilities:

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

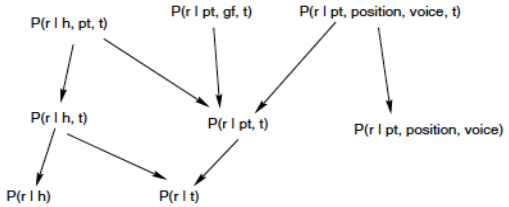
- Geometric Mean in the log domain:

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

where Z is a normalizing constant ensuring that $\sum_r P(r \mid constituent) = 1$.

Back-off Models

- **Back-off models** are often used in statistical modeling to help deal with sparse data. When a conditioned event has not been seen or has been seen rarely, the probability estimate is unreliable.
- The probability estimate is replaced with a probability estimate for a more general event that (hopefully) occurs more. For example, in language models, bigrams may be used as a back-off for a trigrams.
- G&J defined a lattice of feature combinations as a back-off model.



Results for Probability Combinations

- G&J used two methods for assigning weights:
  - using equal weights
  - using the expectation-maximization (EM) algorithm to learn weights
- The baseline always assigns the most common semantic role for the target word to all of its constituents: P(*r*|*t*).

Results on development set, 8,167 observations.

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

Results on test set, 7,900 observations.

Combining Method	Correct
EM linear interpolation	78.5%
Backoff, linear interpolation	76.9
Baseline: Most common role	40.6

Automatically Identifying FE Boundaries

- The previous results all assume that the frame element boundaries are given to the system as input.
- But automated SRL systems must also find the FE boundaries. A more realistic scenario:

  Input: a target word and the frame to which it belongs.

  Output: for each constituent in the parse tree, predict whether it is a frame element or not.

- The path, target word, and head word features were used, separately and in linear combination:

$$P(fe \mid p,h,t) = \lambda_1 P(fe \mid p) + \lambda_2 P(fe \mid p,t) + \lambda_3 P(fe \mid h,t)$$

Sample FE Probabilities

Sample probabilities of a constituent's being a frame element.

Distribution	Sample Probability	Count in training data
$P(fe \mid path)$	$P(fe \mid path = \text{VB}\uparrow\text{VP}\downarrow\text{ADJP}\downarrow\text{ADVP}) = 1$	1
	$P(fe \mid path = \text{VB}\uparrow\text{VP}\downarrow\text{NP}) = .73$	3,963
	$P(fe \mid path = \text{VB}\uparrow\text{VP}\downarrow\text{NP}\downarrow\text{PP}\downarrow\text{S}) = 0$	22
$P(fe \mid path, t)$	$P(fe \mid path = \text{JJ}\uparrow\text{ADJP}\downarrow\text{PP}, t = \text{apparent}) = 1$	10
	$P(fe \mid path = \text{NN}\uparrow\text{NP}\uparrow\text{PP}\uparrow\text{VP}\downarrow\text{PP}, t = \text{departure}) = .4$	5
$P(fe \mid h, t)$	$P(fe \mid h = \text{sudden}, t = \text{apparent}) = 0$	2
	$P(fe \mid h = \text{to}, t = \text{apparent}) = .11$	93
	$P(fe \mid h = \text{that}, t = \text{apparent}) = .21$	81

Recall/Precision Tradeoff

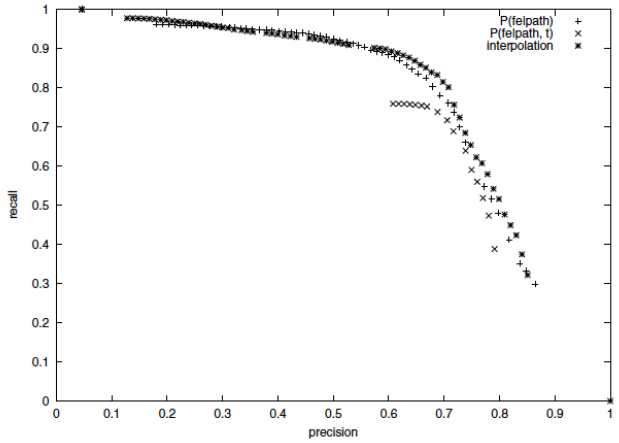


Figure 10
Plot of precision/recall curve for various methods of identifying frame elements. Recall is calculated over only frame elements with matching parse constituents.

Generalizing over Lexical Heads

- The probabilities that condition on a head are the most accurate, but new sentences often have unseen heads.
- G&J experiment with 3 approaches:
  - cluster nouns and estimate P(cluster | head)
  - use the WordNet hierarchy for semantic generalization
  - apply the SRL system to unannotated texts and use the automatic labels as additional training data.
- These approaches all improved coverage, but with lower accuracy than conditioning just on heads.

Generalizing over Unseen Predicates

- Most of the probabilities are conditioned on the target word. So what if new sentences contain unseen targets?
- Approach: train an SRL system on more general semantic roles (akin to thematic roles), which can apply to most target words.
- They mapped the frame-specific roles to more abstract roles and re-trained the system, yielding comparable levels of performance.

Role	Example
AGENT	**Henry** *pushed* the door open and went in.
CAUSE	Jeez, **that** *amazes* me as well as riles me.
DEGREE	I **rather** *deplore* the recent manifestation of Pop; it doesn't seem to me to have the intellectual force of the art of the Sixties.
EXPERIENCER	It may even have been that **John** *anticipating* his imminent doom ratified some such arrangement perhaps in the ceremony at the Jordan.
FORCE	If this is the case can it be *substantiated* **by evidence from the history of developed societies**?
GOAL	Distant across the river the towers of the castle rose against the sky straddling the only land *approach* **into Shrewsbury**.
INSTRUMENT	In the children with colonic contractions **fasting motility** did not *differentiate* children with and without constipation.
LOCATION	These fleshy appendages are used to detect and *taste* food **amongst the weed and debris on the bottom of a river**.
MANNER	His brow *arched* **delicately**.
NULL	Yet while she had no intention of surrendering her home, **it** would be *foolish* to let the atmosphere between them become too acrimonious.
PATH	The dung-collector *ambled* slowly **over**, one eye on Sir John.
PATIENT	As soon as a character lays a hand on this item, the skeletal Cleric *grips* **it** more tightly.
PERCEPT	What is *apparent* **is that this manual is aimed at the non-specialist technician, possibly an embalmer who has good knowledge of some medical procedures**.
PROPOSITION	It says that rotation of partners does not *demonstrate* **independence**.
RESULT	All the arrangements for stay-behind agents in north-west Europe collapsed, but Dansey was able to *charm* most of the governments in exile in London **into recruiting spies**.
SOURCE	He heard the sound of liquid slurping in a metal container as Farrell *approached* him **from behind**.
STATE	Rex *spied* out Sam Maggott **hollering at all and sundry and making good use of his over-sized red gingham handkerchief**.
TOPIC	He said, "We would urge people to be aware and be *alert* **with fireworks** because your fun might be someone else's tragedy."

Results for Abstract Semantic Roles

Role	Number	Known Boundaries % Correct	Unknown Boundaries Labeled Recall	Unknown Boundaries 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

Summary

- Many types of SRL systems have been created, e.g.
  - [Thompson, Levy, and Manning, 2003] defined a probabilistic generative model.
  - [Pradhan et al., 2005] created sequence tagging classifiers.
  - Johansson and Nugues, 2007] used a collection of SVMs.
  - [Das et al., 2010] defined *frame-semantic parsing* as a *structured prediction* task. They use two discriminative log-linear probabilistic models to infer frames and semantic roles.
- The best systems obtain F scores of ~80%.
- Automatic argument identification often accounts for many errors.