Semantic Role Labeling
CS 224N, Spring 2009
Christopher Manning

Slides mainly from a tutorial from Scott Wen-tau Yih and Kristina Toutanova (Microsoft Research), with additional slides from Sameer Pradhan (BSN) as well as Dan Jurafsky and myself.

Syntactic Variations versus Semantic Roles

Yesterday, Kristina hit Scott with a baseball
Scott was hit by Kristina yesterday with a baseball
Yesterday, Scott was hit with a baseball by Kristina
With a baseball, Kristina hit Scott yesterday
Yesterday Scott was hit by Kristina with a baseball
The baseball with which Kristina hit Scott yesterday was hard

Kristina hit Scott with a baseball yesterday

Agent, hitter
Patient, Thing hit
Instrument
Temporal adjunct

Syntactic Variations (as trees)

Semantic Role Labeling – Giving Semantic Labels to Phrases

AGENT John] broke [THEME the window]
[THEME The window] broke
[AGENT Sotheby’s] offered [RECIPIENT the Dorrance heirs] [THEME a money-back guarantee]
[AGENT Sotheby’s] offered [THEME a money-back guarantee] to [RECIPIENT the Dorrance heirs]
[THEME a money-back guarantee] offered by [AGENT Sotheby’s]
[RECIPIENT the Dorrance heirs] will [ARM-NEG not] be offered [THEME a money-back guarantee]

Some typical semantic roles

<table>
<thead>
<tr>
<th>Thematic Role</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGENT</td>
<td>The volitional cause of an event</td>
</tr>
<tr>
<td>EXPERIENCER</td>
<td>The experiencer of an event</td>
</tr>
<tr>
<td>FORCE</td>
<td>The non-volitional cause of the event</td>
</tr>
<tr>
<td>THEME</td>
<td>The participant most directly affected by an event</td>
</tr>
<tr>
<td>RESULT</td>
<td>The end product of an event</td>
</tr>
<tr>
<td>CONTENT</td>
<td>The proposition or content of a propositional event</td>
</tr>
<tr>
<td>INSTRUMENT</td>
<td>An instrument used in an event</td>
</tr>
<tr>
<td>BENEFICIARY</td>
<td>The beneficiary of an event</td>
</tr>
<tr>
<td>SOURCE</td>
<td>The origin of the object of a transfer event</td>
</tr>
<tr>
<td>GOAL</td>
<td>The destination of the object of a transfer event</td>
</tr>
</tbody>
</table>

Some typical semantic roles

<table>
<thead>
<tr>
<th>Thematic Role</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGENT</td>
<td>The water spilled the soup.</td>
</tr>
<tr>
<td>EXPERIENCER</td>
<td>John has a headache.</td>
</tr>
<tr>
<td>FORCE</td>
<td>The wind blows debris from the mall into our yards.</td>
</tr>
<tr>
<td>THEME</td>
<td>Only after Benjamin Franklin broke the ice...</td>
</tr>
<tr>
<td>RESULT</td>
<td>The French government has built a regulation size baseball diamond...</td>
</tr>
<tr>
<td>CONTENT</td>
<td>Mona asked “You not Mary-song at a supermarket?”</td>
</tr>
<tr>
<td>INSTRUMENT</td>
<td>He turned to poaching turkeys, stunning them with a shocking device...</td>
</tr>
<tr>
<td>BENEFICIARY</td>
<td>Whoever Ann Callahan makes hotel reservations for her boss...</td>
</tr>
<tr>
<td>SOURCE</td>
<td>I flew to from Rome.</td>
</tr>
<tr>
<td>GOAL</td>
<td>I drove to Portland.</td>
</tr>
</tbody>
</table>
**What is SRL good for?**

**Question Answering**

Q: What was the name of the first computer system that defeated Kasparov?
A: **PATIENT Kasparov** was defeated by **AGENT Deep Blue** **TIME in 1997**.

Q: When was Napoleon defeated?
Look for: **PATIENT Napoleon** [defeat-synset] [TIME]

More generally:

- Who hit Scott with a baseball?
- Whom did Kristina hit with a baseball?
- What did Kristina hit Scott with?
- When did Kristina hit Scott with a baseball?

---

**Applications as a simple meaning rep’n**

- **Machine Translation**
  - English (SVO)
  - Farsi (SOV)
  - English (SVO)
  - Farsi (SOV)
  - **AGENT The little boy**
  - **THEME the red ball**
  - **AGENT pesar koocholo**
  - **THEME toop germezi**

- **Document Summarization**
  - Predicates and Heads of Roles summarize content

- **Information Extraction**
  - SRL can be used to construct useful rules for IE

---

**Application: Semantically precise search**

**Query: afghans destroying opium poppies**

**Diathesis alternations**

- **John** broke the window.
  - **AGENT**
  - **THEME**
  - **INSTRUMENT**

- **John** broke the window with a rock.
  - **AGENT**
  - **THEME**
  - **INSTRUMENT**

- **The rock** broke the window.
  - **INSTRUMENT**
  - **THEME**

- **The window** broke.
  - **THEME**

- **Doris** gave the book to Cary.
  - **AGENT**
  - **THEME**
  - **GOAL**

- **Doris** gave Cary the book.
  - **AGENT**
  - **GOAL**

---

**Problems with semantic roles**

- It’s very hard to produce a formal definition of a role
- There are all sorts of arbitrary role splits
- Intermediary instruments (1-2) vs. enabling instruments (3-4):
  1. The cook opened the jar with the new gadget
  2. The new gadget opened the jar
  3. Sally ate the sliced banana with a fork
  4. *The fork ate the sliced banana*

---

**Solutions to the difficulty of defining semantic roles**

- Ignore semantic role labels, and just mark arguments of individual verbs as 0, 1, 2
  - PropBank
- Define semantic role labels for a particular semantic domain
  - FrameNet
PropBank

- A corpus of labeled sentences (Penn Treebank WSJ)
- The arguments of each verb are labeled with numbers rather than names, though there are verb frame files:

19.20: agree.vb
A0: Agree
A1: Proposes
A2: Other entity agreeing
Ex1: [A0: The group] agreed [A2: it wouldn't make an offer unless it had been given the right]
Ex2: [A0: I] agreed [A2: him] agreed [A2: with Mary]

19.40: fail.vb
A0: Logical subject, patient, thing falling
A1: Fall
A2: Start point
A3: End point and state of end

Proposition Bank (PropBank)
Add a Semantic Layer

S V P S N P V P N P

Kristina hit Scott with a baseball yesterday

[AO Kristina] hit [A1 Scott] [A2 with a baseball] [AM-TMP yesterday].

Proposition Bank (PropBank)
Add a Semantic Layer – Continued

S V P S N P V P C-A1

The worst thing about him said Kristina is his laziness

[AO The worst thing about him] said [AO Kristina] is his laziness.

Proposition Bank (PropBank)
Final Notes

- Current release (Mar 4, 2005): Proposition Bank I
  - Verb Lexicon: 3,324 frame files
  - Annotation: ~113,000 propositions
    http://www.nlp.cis.upenn.edu/project_pages/ACE.htm
  - Alternative format: CoNLL-04, 05 shared task
    - Represented in table format
    - Has been used as standard data set for the shared tasks on semantic role labeling
    http://www.lsi.upc.es/~srlconll/soft.html

Other Corpora

- Chinese PropBank http://www.cis.upenn.edu/~chinese/prop/
  - Similar to PropBank, it has a semantic layer onto Chinese Treebank
  - Label arguments that co-occur with nouns in PropBank
  - [AO Her] [REL gift] of [A1 a book] [A2 to John]
CoNLL format

1. lie("he", "...")
2. leak("he", "..."
3. obtain("he", "..."
4. supervise("he", "..."

FrameNet [Fillmore et al. 01]

FrameNet

- A frame is a semantic structure based on
  a set of participants and events
- Consider the “change_position_on_scale” frame

Examples / Quiz question

- [ITEM Oil] rose [ATTRIBUTE in price] [DIFFERENCE by 2%]
- [ITEM Oil] has increased [VALUE_RANGE of having them 1 day a month]
- [ITEM Microsoft shares] fell [VALUE_RANGE to 7 5/8]
- a steady [INCREASE] [from 9.5] to [14.3] in dividends
- a [5%] dividend [INCREASE]

Give the roles for the 3 items with blanks
- Core: Attribute, Difference, Final_State, Final_Value, Initial_State, Initial_Value, Item, Value_Range
- An Item has a scalar Attribute which moves in the Value_Range portion of a scale
- Some Non-Core: Duration, Speed, Group
- The Item changes its Attribute within a Group for a certain Duration or at a Speed

Problems with FrameNet

- Example sentences are chosen by hand
- Not randomly selected
- Complete sentences not labeled
- Since TreeBank wasn’t used
- No perfect parses for each sentence
- Still ongoing (that’s good and bad)
Some History

- Fillmore 1968: The case for case
  - Proposed semantic roles as a shallow semantic representation
- Simmons 1973:
  - Built first automatic semantic role labeler
  - Based on first parsing the sentence

FrameNet vs PropBank -1

FrameNet vs PropBank -2

Information Extraction versus Semantic Role Labeling

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>IE</th>
<th>SRL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coverage</td>
<td>narrow</td>
<td>broad</td>
</tr>
<tr>
<td>Depth of semantics</td>
<td>shallow</td>
<td>shallow</td>
</tr>
<tr>
<td>Directly connected to application</td>
<td>sometimes</td>
<td>no</td>
</tr>
</tbody>
</table>

Overview of SRL Systems

- Definition of the SRL task
- Evaluation measures
- General system architectures
- Machine learning models
  - Features & models
  - Performance gains from different techniques

Subtasks

- Identification: $\binom{1,2,\ldots,n}{l} \rightarrow \{\text{NONE, ARG}\}$
  - Very hard task: to separate the argument substrings from the rest in this exponentially sized set
  - Usually only 1 to 9 (avg. 2.7) substrings have labels ARG and the rest have NONE for a predicate
- Classification: $\binom{1,2,\ldots,n}{l} \rightarrow L \setminus \{\text{NONE}\}$
  - Given the set of substrings that have an ARG label, decide the exact semantic label
- Core argument semantic role labeling: (easier)
  - Label phrases with core argument labels only. The modifier arguments are assumed to have label NONE.
Evaluation Measures

- Correct: \[\text{The queen} \text{ broke} \] [\text{the window}]
- Guess: \[\text{The queen} \text{ broke} \] [\text{the window}]

<table>
<thead>
<tr>
<th>Correct</th>
<th>Guess</th>
</tr>
</thead>
</table>
| \[\text{The queen} \text{ broke} \] [\text{the window}] | \[\text{The queen} \text{ broke} \] [\text{the window}]
| \[\text{Yesterday} \] | \[\text{AM-TMP} \text{ yesterday} \]

- Precision, Recall, F-Measure (tp=1, fp=2, fn=2) \( pr=1/3 \)
- Measures for subtasks
  - Identification (Precision, Recall, F-measure) (tp=2, fp=1, fn=1) \( pr=1/2 \)
  - Classification (Accuracy) \( acc = 0.5 \) (labeling of correctly identified phrases)
  - Core arguments (Precision, Recall, F-measure) (tp=1, fp=1, fn=1) \( pr=1/2 \)

Why this parse-tree architecture?

- Semantic role chunks tend to correspond to syntactic constituents
- Propbank:
  - 96% of arguments = 1 (gold) parse tree constituent
  - 90% of arguments = 1 (Charniak) parse tree constituent
  - Simple rules can recover missing 4-10%
- FrameNet:
  - 87% of arguments = 1 (Collins) parse tree constituent
  - Why?
  - they were labeled from parse trees by humans trained in syntax

What’s the problem with these evaluations?

- Approximating human evaluations is dangerous
- Humans don’t always agree
- Not clear if it’s good for anything
- Sometimes called the “match-a-linguist” task
- What’s a better evaluation?

SRL architecture:
Walk the tree, labeling each parse tree node

- Given a parse tree \( t \), label the nodes (phrases) in the tree with semantic labels

Parsing Algorithm

- Use a syntactic parser to parse the sentence
- For each predicate (non-copula verb)
  - For each node in the syntax tree
    - Extract a feature vector relative to the predicate
    - Classify the node
  - Do second-pass informed by global info
### Baseline Features

- Predicate (verb)
- Path from constituent to predicate
- Phrase type (syntactic)
- Position (before/after)
- Voice (active/passive)
- Head Word
- Sub-categorization

### Pradhan et al. (2004) Features

- Predicate cluster
- Noun head and POS of PP constituent
- Verb sense
- Partial path
- Named entities in constituent (7) [Surdeanu et al., 2003]
- Head word POS [Surdeanu et al., 2003]
- First and last word in constituent and their POS
- Parent and sibling features
- Constituent tree distance
- Ordinal constituent position
- Temporal cue words in constituent
- Previous 2 classifications

### Predicate cluster, automatic or WordNet

- Spoke
- Lectured
- Chatted
- Explained

### Noun Head and POS of PP

- She
- It
- They
- Half an hour
- 60 seconds

### Partial Path

### Named Entities and Head Word POS

- The
- They
- Half an hour
- 60 seconds
First and Last Word and POS

Parent and Sibling features

Constituent tree distance

Ordinal constituent position

Temporal Cue Words (~50)

time years;ago
recently night
days hour
end decade
period late

Previous 2 classifications
Combining Identification and Classification Models

- Constraint 1: Argument phrases do not overlap
  - By [i., working [i.], hard] he said, you can achieve a lot.
  - Pradhan et al. (04) – greedy search for a best set of non-overlapping arguments
  - Toutanova et al. (05) – exact search for the best set of non-overlapping arguments (dynamic programming, linear in the size of the tree)
  - Punyakanok et al. (05) – exact search for best non-overlapping arguments using integer linear programming
- Other constraints (Punyakanok et al. 04, 05)
  - no repeated core arguments (good heuristic)
  - phrases do not overlap the predicate

Joint Scoring: Enforcing Hard Constraints

- These models have scores for a whole labeling of a tree (not just individual labels)
  - Encode some dependencies among the labels of different nodes
  \[ P_{\text{JOIN}}(l_1, \ldots, l_n|n, t, p) = \prod_i P(l_i|n_i, t, p) \]

Joint Scoring Models

- Combining Local and Joint Scoring Models
  - Tight integration of local and joint scoring in a single probabilistic model and exact search [Cohn&Blabzom 05] [Marquez et al. 05] [Thompson et al. 03]
  - When the joint model makes strong independence assumptions
  - Re-ranking or approximate search to find the labeling which maximizes a combination of local and a joint score [Glisak & Jurafsky 02] [Pradhan et al. 04] [Toutanova et al. 05]
  - Usually exponential search required to find the exact maximizer
- Exact search for best assignment by local model satisfying hard joint constraints
  - Using integer Linear Programming [Punyakanok et al. 04,05] (worst case NP-hard)

Joint Scoring: Integrating Soft Preferences

- There are many statistical tendencies for the sequence of roles and their syntactic realizations
  - When both are before the verb, AM-TMP is usually before A0
  - Usually, there aren’t multiple temporal modifiers
  - Many others which can be learned automatically
Joint Scoring: Integrating Soft Preferences

- Gildea and Jurafsky (02) – a smoothed relative frequency estimate of the probability of frame element multi-sets: 
  \( P(A_0, AM_{TFP}, A_1, AM_{TFP}) \) [hit]
  - Gains relative to local model: 59.2 → 62.9
  - FrameNet automatic parses

- Pradhan et al. (04) – a language model on argument label sequences (with the predicate included):
  \( P(A_0, AM_{TFP}, hit, A_1, AM_{TFP}) \)
  - Small gains relative to local model for a baseline system: 88.0 → 88.9 on core arguments. PropBank correct parses

- Toutanova et al. (05) – a joint model based on CRFs with a rich set of joint features of the sequence of labeled arguments:
  - Gains relative to local model on PropBank correct parses: 88.4 → 91.2 (24% error reduction); gains on automatic parses 78.2 → 80.0.
  - Also tree CRFs [Cohn & Brunson] have been used.

Semantic roles: joint models boost results [Toutanova et al. 2005]

Accuracies of local and joint models on core arguments

![Graph showing accuracies](image)

Error reduction from best published result:
- 44.6% on Integrated
- 52% on Classification

System Properties

- Features
  - Most modern systems use the standard set of Gildea, Pradhan, and Surdeanu features listed above
  - Lots of features important for building a good system

- Learning Methods
  - SNoW, MaxEnt, AdaBoost, SVM, CRFs, etc.
  - The choice of learning algorithms is less important.

System Properties – Continued

- Syntactic Information
  - Charniak’s parser, Collins’ parser, cluser, chunker, etc.
  - Top systems use Charniak’s parser or some mixture
  - Quality of syntactic information is important

- System/Information Combination
  - Greedy, Re-ranking, Stacking, ILP inference
  - Combination of systems or syntactic information is a good strategy to reduce the influence of incorrect syntactic information!

Per Argument Performance

CoNLL-05 Results on WSJ-Test

<table>
<thead>
<tr>
<th>Core Arguments (Freq. ~70%)</th>
<th>Adjuncts (Freq. ~30%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Best F</strong>, <strong>Freq.</strong></td>
<td><strong>Best F</strong>, <strong>Freq.</strong></td>
</tr>
<tr>
<td>A0 88.31 25.58%</td>
<td>TMP 78.21 8.86%</td>
</tr>
<tr>
<td>A1 79.91 35.36%</td>
<td>ADV 59.73 3.46%</td>
</tr>
<tr>
<td>A2 72.26 8.26%</td>
<td>DIS 50.45 2.05%</td>
</tr>
<tr>
<td>A3 65.26 1.39%</td>
<td>MNR 59.22 2.67%</td>
</tr>
<tr>
<td>A4 77.25 1.39%</td>
<td>LOC 60.99 2.48%</td>
</tr>
<tr>
<td></td>
<td>MOD 98.47 3.83%</td>
</tr>
<tr>
<td></td>
<td>CAU 64.62 0.50%</td>
</tr>
<tr>
<td></td>
<td>NEG 98.91 1.36%</td>
</tr>
</tbody>
</table>

Arguments that need to be improved

Summary

- Semantic role labeling
  - An important attempt at a general approach to shallow semantic extraction
  - Relatively successful in terms of approximating
    - Human FrameNet labels
    - Human PropBank labels
- Are these good for anything?
  - We don’t know for sure yet