Semantic Role Labeling
CS 224N
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 an 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 blew 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 are Mary Ann at a supermarket?”</td>
</tr>
<tr>
<td>INSTRUMENT</td>
<td>He turned to reaching out fish, stunning them with a shocking device...</td>
</tr>
<tr>
<td>BENEFICIARY</td>
<td>Whenever Ann Callahan makes hotel reservations for her boss...</td>
</tr>
<tr>
<td>SOURCE</td>
<td>I flew to From Russia.</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] [PRED: defeat-synset] [ARGM-TMP: "ANS"]

More generally:

```
WHO

WHAT

WHEN
```

- 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)**
    ```
    [AGENT: The little boy] [PRED: kicked] [THEME: the red ball]
    [ARGM-MNR: hard]
    [ARGM-COMP: zaad-e]
    ```
  - **Persian (SOV)**
    ```
    [AGENT: pesar koocholo] [AGENT: The little boy] [PRED: kicked] [THEME: the red ball]
    [ARGM-MNR: hard]
    [ARGM-COMP: zaad-e]
    ```
- 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

**Example:**

**English:**
```
John broke the window.
AGENT: John
THEME: The window
```

**Persian:**
```
Doris gave the book to Cary.
AGENT: Doris
THEME: Cary
```

---

**Diathesis alternations**

John broke the window.

<table>
<thead>
<tr>
<th>Agent</th>
<th>Theme</th>
</tr>
</thead>
<tbody>
<tr>
<td>John</td>
<td>Theme</td>
</tr>
<tr>
<td>Theme</td>
<td>Agent</td>
</tr>
</tbody>
</table>

---

**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:

(19.29) strike.01
A0: agent, hitter; A1: thing hit; A2: instrument, thing hit by or with
Kristina hit [A1 Scott] [A2 with a baseball] yesterday.

(20.80) seemed.02
A0: seemer; A1: seemed like; A2: seemed to
It looked to her like [A0 he] deserved [A1 this].

deserve.01
A0: deserving entity; A1: thing deserved; A2: in-exchange-for
It looked to her like [A0 he] deserved [A1 this].

AM-TMP
Time
Proposition Bank (PropBank)

1. lie("he",…)
2. leak("he", "information obtained from … he supervised")
3. obtain(X, "information", "from a wiretap he supervised")
4. supervise("he", "a wiretap")

CoNLL format
Other Corpora

- **Chinese PropBank**
  
  http://www.cis.upenn.edu/~chinese/chp/
  
  Similar to PropBank, it adds a semantic layer onto Chinese Treebank

- **NomBank**

  http://nlp.cs.nyu.edu/meyers/NomBank.html

  Label arguments that co-occur with nouns in PropBank

  [as Her [nel gift] of [as a book]] [as to John]

FrameNet

- **FrameNet** [Fillmore et al. 01]

  A frame is a semantic structure based on a set of participants and events

  Consider the “change_position_on_scale” frame

  | VERBS: | advance | climb | edge | flood | move | move | roam | soar | swing | tumble | turn | walk | wave |
  |--------|---------|-------|------|-------|------|------|------|------|-------|-------|------|------|
  | ADJ:   | down    | low   | slow | still | slow | slow | slow | slow | slow  | slow  | slow | slow | slow |
  | ADV:   | adjoin  | away  | back | below | beyond | below | below | below | below | below | below | below |
  | NOUN:  | arena   | area  | beach | deck  | deck  | deck  | deck  | deck  | deck  | deck  | deck  | deck  |
  | SYL:   | above   | ahead | after | away  | away  | away  | away  | away  | away  | away  | away  | away  |

Examples / Quiz question

- [ITEM Oil] rose [ATTRIBUTE in price] [DIFFERENCE by 2%].
- [ITEM Oil] has increased [FINAL_VALUE to having them 1 day a month].
- [ITEM Microsoft shares] fell [FINAL_VALUE to 7 5/8].

- Colon cancer incidence fell by 50% among men.
- 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

Some Non-Core: Duration, Speed, Group

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>

- The goal of SRL is to provide a general capability for semantic relation and argument identification, not custom ones for particular applications

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: \[ \text{ARGS} \rightarrow \{ \text{NONE}, \text{ARG} \} \]
  - 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: \[ \{ 1, \ldots, m \} \rightarrow L \setminus \{ \text{NONE} \} \]
  - 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: [A0: The queen] broke [A1: the window] [AM-TMP: yesterday]
Guess: [A0: The queen] broke the [A1: window] [AM-LOC: yesterday]

<table>
<thead>
<tr>
<th>Correct</th>
<th>Guess</th>
</tr>
</thead>
<tbody>
<tr>
<td>(the queen)</td>
<td>(the queen)</td>
</tr>
<tr>
<td>(the window)</td>
<td>(window)</td>
</tr>
<tr>
<td>(yesterday)</td>
<td>(yesterday)</td>
</tr>
<tr>
<td>all other</td>
<td>all other</td>
</tr>
<tr>
<td>→ NONE</td>
<td>→ NONE</td>
</tr>
</tbody>
</table>

- Precision, Recall, F-Measure \(tp=1, fp=2, fn=2\) \(p=r=f=1/3\)
- Measures for subtasks
  - Identification (Precision, Recall, F-measure) \(tp=2, fp=1, fn=1\) \(p=r=f=2/3\)
  - Classification (Accuracy) \(acc = 0.5\)
  - Core arguments (Precision, Recall, F-measure) \(tp=1, fp=1, fn=1\) \(p=r=f=1/2\)

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?

Basic Architecture of a Generic SRL System

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

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

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 [Gildea & Jurafsky, 2000]

- 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

Noun Head and POS of PP

Partial Path

Named Entities and Head Word POS [Surdeanu et al., 2003]
First and Last Word and POS

Parent and Sibling features

Constituent tree distance

Ordinal constituent position

Temporal Cue Words (~50)

<table>
<thead>
<tr>
<th>time</th>
<th>years;ago</th>
</tr>
</thead>
<tbody>
<tr>
<td>recently</td>
<td>night</td>
</tr>
<tr>
<td>days</td>
<td>hour</td>
</tr>
<tr>
<td>end</td>
<td>decade</td>
</tr>
<tr>
<td>period</td>
<td>late</td>
</tr>
</tbody>
</table>

Previous 2 classifications
Combining Identification and Classification Models

Step 1. Pruning: Using a hand-specified filter.

Step 2. Identification: Identification model assigns one of the argument labels to selected nodes (or sometimes possibly NONE).

Step 3. Classification: Classification model assigns one of the argument labels to selected nodes (or sometimes possibly NONE).

Combining Identification and Classification Models – Continued

One Step: Simultaneously identify and classify using \( P(l|c, t, p) \).

Combining Local and Joint Scoring Models

- Tight integration of local and joint scoring in a single probabilistic model and exact search [Cohn & Blunsom 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 [Staal & 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 Models

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

Joint Scoring: Enforcing Hard Constraints

- Constraint 1: Argument phrases do not overlap
  - By [working [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: 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:
  \[ P(A0, AM_{TMP}, A1, AM_{TMP}) | \text{hit} \]
  - Gains relative to local model 59.2 \(\rightarrow\) 62.9 on FrameNet automatic parses
- Pradhan et al. (04) – a language model on argument label sequences (with the predicate included):
  \[ P(A0, AM_{TMP}, \text{hit}, A1, AM_{TMP}) \]
  - Small gains relative to local model for a baseline system 88.0 \(\rightarrow\) 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 \(\rightarrow\) 91.2 (24% error reduction); gains on automatic parses 78.2 \(\rightarrow\) 80.0
- Also tree CRFs [Cohn & Brunson] have been used

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.

Per Argument Performance

CoNLL-05 Results on WSJ-Test

- Core Arguments (Freq. \(~70\%\))

<table>
<thead>
<tr>
<th></th>
<th>Best F</th>
<th>Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>A0</td>
<td>88.31</td>
<td>25.58%</td>
</tr>
<tr>
<td>A1</td>
<td>79.91</td>
<td>35.36%</td>
</tr>
<tr>
<td>A2</td>
<td>70.26</td>
<td>8.26%</td>
</tr>
<tr>
<td>A3</td>
<td>65.26</td>
<td>1.39%</td>
</tr>
<tr>
<td>A4</td>
<td>77.25</td>
<td>1.09%</td>
</tr>
</tbody>
</table>

- Adjuncts (Freq. \(~30\%\))

<table>
<thead>
<tr>
<th></th>
<th>Best F</th>
<th>Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>TMP</td>
<td>78.21</td>
<td>8.66%</td>
</tr>
<tr>
<td>ADV</td>
<td>59.73</td>
<td>3.46%</td>
</tr>
<tr>
<td>DIS</td>
<td>60.45</td>
<td>2.05%</td>
</tr>
<tr>
<td>MNR</td>
<td>59.22</td>
<td>2.67%</td>
</tr>
<tr>
<td>LOC</td>
<td>60.99</td>
<td>2.48%</td>
</tr>
<tr>
<td>MOD</td>
<td>98.47</td>
<td>3.83%</td>
</tr>
<tr>
<td>CAU</td>
<td>64.62</td>
<td>0.50%</td>
</tr>
<tr>
<td>NEG</td>
<td>98.91</td>
<td>1.36%</td>
</tr>
</tbody>
</table>

Arguments that need to be improved

Data from Carreras&Màrkquez's slides (CoNLL 2006)

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