Semi-supervised Semantic Role Labeling

Hagen Fürstenau

Department of Computational Linguistics
Saarland University

(joint work with Mirella Lapata)

“FEAST” meeting
November 26th, 2008
Outline

1. Frame Semantics
2. A Semi-supervised approach to role labeling
3. Summary
Frame Semantics

Charles J. Fillmore, 1975 & 1981

**Definition**

- A **frame** describes a prototypical situation.
- It is evoked by a **frame evoking element (FEE)**.
- It can have several **frame elements (roles)**.
Frame Semantics

Charles J. Fillmore, 1975 & 1981

Definition

- A **frame** describes a prototypical situation.
- It is evoked by a **frame evoking element (FEE)**.
- It can have several **frame elements (roles)**.

Matilde fried the catfish in a heavy iron skillet.
Frame Semantics

Charles J. Fillmore, 1975 & 1981

Definition

- A **frame** describes a prototypical situation.
- It is evoked by a **frame evoking element (FEE)**.
- It can have several **frame elements (roles)**.

Example:

Apply\_heat

Matilde **fried** the catfish in a heavy iron skillet.
Frame Semantics

Charles J. Fillmore, 1975 & 1981

Definition

- A **frame** describes a prototypical situation.
- It is evoked by a **frame evoking element (FEE)**.
- It can have several **frame elements (roles)**.

```
Apply_heat
FEE
Matilde fried the catfish in a heavy iron skillet.
Roles
Cook
Food
Heating_instrument

Matilde fried the catfish in a heavy iron skillet.
FEE
```
Frame Semantics

- Shallow semantic analysis
- Generalizes well across languages
- Avoids problem of “universal roles”
- Can benefit various NLP tasks (IR, QA, ...)

Google snapped up YouTube for $1.65 billion.
Frame Semantics

- Shallow semantic analysis
- Generalizes well across languages
- Avoids problem of “universal roles”
- Can benefit various NLP tasks (IR, QA, ...)
Frame Semantics

- Shallow semantic analysis
- Generalizes well across languages
- Avoids problem of “universal roles”
- Can benefit various NLP tasks (IR, QA, ...)

Google snapped up YouTube for $1.65 billion.
Frame Semantics

- Shallow semantic analysis
- Generalizes well across languages
- Avoids problem of “universal roles”
- Can benefit various NLP tasks (IR, QA, ...)

How much did Google pay for YouTube?

Commerce_goods-transfer

Google snapped up YouTube for $1.65 billion.
Frame Semantic Parsing

To automatically derive Frame Semantic analyses

1. Take an annotated corpus: FrameNet (135,000 sentences)
2. Train a classifier on this data
3. Use classifier as Frame Semantic parser
Frame Semantic Parsing

To automatically derive Frame Semantic analyses

1. Take an annotated corpus: FrameNet (135,000 sentences)
2. Train a classifier on this data
3. Use classifier as Frame Semantic parser

- Annotation is expensive and time-consuming
- Must be repeated for new languages or domains

Can we reduce this annotation effort?
Semi-supervised learning

Goal: Try to make use of unlabeled data!

Example: Binary classification
Semi-supervised learning

Goal: Try to make use of unlabeled data!

Example: Binary classification
Semi-supervised learning

Goal: Try to make use of unlabeled data!

Example: Binary classification
Semi-supervised learning

Goal: Try to make use of unlabeled data!

Example: Binary classification
Semi-supervised learning

Goal: Try to make use of unlabeled data!

Example: Binary classification
Semi-supervised learning

Goal: Try to make use of unlabeled data!

Example: Binary classification
To expand a Frame Semantic corpus

1. Find unlabeled sentences “similar” to labeled ones
2. Project annotations from labeled sentences
3. Add new labeled examples to annotation pool
4. Hope that the expanded corpus is “better” than the original one
Applied to role labeling

To expand a Frame Semantic corpus

1. Find unlabeled sentences “similar” to labeled ones
2. Project annotations from labeled sentences
3. Add new labeled examples to annotation pool
4. Hope that the expanded corpus is “better” than the original one

- What’s “similar”? Take into account syntactic and semantic measures!
- What’s “better”? A supervised algorithm makes better predictions when trained on the expanded corpus.
The General Framework

FrameNet

training

test

Syntactic parsing

Classifier
The General Framework

FrameNet

training  test

Syntactic parsing

BNC

Syntactic parsing
The General Framework

FrameNet

- Training
- Syntactic parsing
- Annotation of similar sentences

BNC

- Syntactic parsing
The General Framework

FrameNet

training test

Classifier

BNC

Results improved?
Find best alignment between predicate-argument structures:

<table>
<thead>
<tr>
<th>GR</th>
<th>Lemma</th>
<th>Role</th>
</tr>
</thead>
<tbody>
<tr>
<td>subj</td>
<td>Mathilde</td>
<td>Cook</td>
</tr>
<tr>
<td>obj</td>
<td>catfish</td>
<td>Food</td>
</tr>
<tr>
<td>mod_in</td>
<td>skillet</td>
<td>Heating_instrument</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GR</th>
<th>Lemma</th>
<th>Role</th>
</tr>
</thead>
<tbody>
<tr>
<td>subj</td>
<td>we</td>
<td></td>
</tr>
<tr>
<td>obj</td>
<td>egg</td>
<td></td>
</tr>
</tbody>
</table>
Similarity measure

Find best alignment between predicate-argument structures:

Mathilde catfish in the skillet iron Apply_heat

Apply_heat
Cook

Cooking_instrument
Food

GR Lemma Role
subj Mathilde Cook
obj catfish Food
mod_in skillet Heating_instrument

we egg
Apply_heat
Cook
Food
Alignment

- *We can feel the blood coursing through our veins again.*
- *Adrenalin was still coursing through her veins.*
Similarity measure

Consider a partial, injective alignment function

\[ \sigma : \{1, \ldots, m\} \rightarrow \{\varepsilon, 1, \ldots, n\} \]

\[ (\sigma(i) = \sigma(j) \neq \varepsilon \Rightarrow i = j) \]

and define similarity with respect to this alignment:

\[ \text{sim}(\sigma) := \sum_{\sigma(i) \neq \varepsilon}^{m} \left( A \cdot \delta_{GR_i, GR_{\sigma(i)}} + \cos(\vec{v}_i, \vec{v}_{\sigma(i)}) - B \right) \]

The similarity of two predicate-argument structures is

\[ \max_{\sigma} \text{sim}(\sigma) \]
Choosing which sentences to label

We have a large corpus, therefore we can be picky:

- Annotate if similarity is above some threshold? Global threshold value doesn’t work!
- Pick k-NN unlabeled sentences for each labeled one? Neglects some of the global structure.
- Global graph optimization? Computationally expensive.
Choosing which sentences to label

We have a large corpus, therefore we can be picky:

- Annotate if similarity is above some threshold? Global threshold value doesn’t work!
- Pick k-NN unlabeled sentences for each labeled one? Neglects some of the global structure.

- Global graph optimization? Computationally expensive.
Choosing which sentences to label

We have a large corpus, therefore we can be picky:

- Annotate if similarity is above some threshold? Global threshold value doesn’t work!
- Pick k-NN unlabeled sentences for each labeled one? Neglects some of the global structure.
- Global graph optimization? Computationally expensive.
Choosing which sentences to label

We have a large corpus, therefore we can be picky:

- Annotate if similarity is above some threshold? Global threshold value doesn’t work!
- **Pick k-NN unlabeled sentences for each labeled one?** Neglects some of the global structure.
- Global graph optimization? Computationally expensive.
Evaluation

- Compare labeling performance *before* and *after* expansion.
- Independent supervised labeler: *desert* (dependency-based semantic role labeling toolkit)
  - 1. Argument recognition
  - 2. Argument labeling
- Evaluate on 100 verb FrameNet sample (80% training, 20% test)
- Labeled precision: \[ \frac{\text{# correctly labeled roles}}{\text{# all predicted roles}} \]
- Labeled recall: \[ \frac{\text{# correctly labeled roles}}{\text{# all real roles}} \]
- Labeled $F_1$ score: \[ F_1 = \frac{2PR}{P+R} \]
## Results

<table>
<thead>
<tr>
<th>TrainSet</th>
<th>Size</th>
<th>Frame Acc.</th>
<th>Prec (%)</th>
<th>Rec (%)</th>
<th>$F_1$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>≤ 1 seed</td>
<td>95</td>
<td>81.6</td>
<td>24.9</td>
<td>31.3</td>
<td>27.7</td>
</tr>
<tr>
<td>+ 2-NN</td>
<td>170</td>
<td>81.6</td>
<td>26.4</td>
<td>32.6</td>
<td>29.2*</td>
</tr>
<tr>
<td>+ self training</td>
<td>183</td>
<td>81.6</td>
<td>22.2</td>
<td>29.3</td>
<td>25.3</td>
</tr>
<tr>
<td>≤ 5 seeds</td>
<td>450</td>
<td>83.6</td>
<td>29.7</td>
<td>38.4</td>
<td>33.5</td>
</tr>
<tr>
<td>+ 2-NN</td>
<td>844</td>
<td>82.6</td>
<td>31.8</td>
<td>40.4</td>
<td>35.6*</td>
</tr>
<tr>
<td>+ self training</td>
<td>882</td>
<td>83.4</td>
<td>24.3</td>
<td>33.4</td>
<td>28.1</td>
</tr>
<tr>
<td>≤ 10 seeds</td>
<td>849</td>
<td>83.1</td>
<td>35.5</td>
<td>42.0</td>
<td>38.5</td>
</tr>
<tr>
<td>+ 2-NN</td>
<td>1549</td>
<td>82.8</td>
<td>38.1</td>
<td>44.1</td>
<td>40.9*</td>
</tr>
<tr>
<td>+ self training</td>
<td>1609</td>
<td>82.9</td>
<td>34.0</td>
<td>41.0</td>
<td>37.1</td>
</tr>
<tr>
<td>≤ 20 seeds</td>
<td>1414</td>
<td>84.6</td>
<td>38.7</td>
<td>46.1</td>
<td>42.1</td>
</tr>
<tr>
<td>+ 2-NN</td>
<td>2600</td>
<td>85.6</td>
<td>40.5</td>
<td>46.7</td>
<td>43.4</td>
</tr>
<tr>
<td>+ self training</td>
<td>2686</td>
<td>85.3</td>
<td>34.1</td>
<td>42.0</td>
<td>37.6</td>
</tr>
<tr>
<td>all seeds</td>
<td>2323</td>
<td>84.6</td>
<td>38.3</td>
<td>47.0</td>
<td>42.2</td>
</tr>
<tr>
<td>+ 2-NN</td>
<td>4387</td>
<td>84.6</td>
<td>39.5</td>
<td>46.7</td>
<td>42.8</td>
</tr>
<tr>
<td>+ self training</td>
<td>4501</td>
<td>85.0</td>
<td>34.5</td>
<td>44.1</td>
<td>38.7</td>
</tr>
</tbody>
</table>

- Significant improvements for 1, 5 and 10 seeds per verb
- Expanded ”10 seeds set“ almost as good as original ”20 seeds set“!
## Results (contd.)

<table>
<thead>
<tr>
<th>TrainSet</th>
<th>Size</th>
<th>Frame Acc.</th>
<th>Prec (%)</th>
<th>Rec (%)</th>
<th>$F_1$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-NN</td>
<td>849</td>
<td>83.1</td>
<td>35.5</td>
<td>42.0</td>
<td>38.5</td>
</tr>
<tr>
<td>1-NN</td>
<td>1205</td>
<td>83.1</td>
<td>36.4</td>
<td>43.3</td>
<td>39.5</td>
</tr>
<tr>
<td>2-NN</td>
<td>1549</td>
<td>82.8</td>
<td>38.1</td>
<td>44.1</td>
<td>40.9*</td>
</tr>
<tr>
<td>3-NN</td>
<td>1883</td>
<td>83.1</td>
<td>37.9</td>
<td>43.7</td>
<td>40.6*</td>
</tr>
<tr>
<td>4-NN</td>
<td>2204</td>
<td>82.3</td>
<td>38.0</td>
<td>43.9</td>
<td>40.7*</td>
</tr>
<tr>
<td>5-NN</td>
<td>2514</td>
<td>81.9</td>
<td>37.4</td>
<td>43.9</td>
<td>40.4*</td>
</tr>
</tbody>
</table>

- Significant improvements for all but 1-NN
Example


Nearest neighbors

Summary

Our approach
- Only relies on seed corpus and dependency parser
- Uses a general, extensible similarity metric
- Shows encouraging first results

Future Work
- Compare predicate-argument structures of different verbs
- Globally optimize similarity graph partition
- Automatically identify alternation behaviour
- Apply method to other languages (German) and corpora (PropBank)