Abstract

We explore the extent to which high-resource manual annotations such as tree-banks are necessary for the task of semantic role labeling (SRL). We examine how performance changes without syntactic supervision, comparing both joint and pipelined methods to induce latent syntax. This work highlights a new application of unsupervised grammar induction and demonstrates several approaches to SRL in the absence of supervised syntax. Our best models obtain competitive results in the high-resource setting and state-of-the-art results in the low resource setting, reaching 72.48% F1 averaged across languages. We release our code for this work along with a larger toolkit for specifying arbitrary graphical structure.

1 Introduction

The goal of semantic role labeling (SRL) is to identify predicates and arguments and label their semantic contribution in a sentence. Such labeling defines who did what to whom, when, where and how. For example, in the sentence “The kids ran the marathon”, ran assigns a role to kids to denote that they are the runners; and a role to marathon to denote that it is the race course.

Models for SRL have increasingly come to rely on an array of NLP tools (e.g., parsers, lemmatizers) in order to obtain state-of-the-art results (Björk et al., 2009; Zhao et al., 2009). Each tool is typically trained on hand-annotated data, thus placing SRL at the end of a very high-resource NLP pipeline. However, richly annotated data such as that provided in parsing treebanks is expensive to produce, and may be tied to specific domains (e.g., newswire). Many languages do not have such supervised resources (low-resource languages), which makes exploring SRL cross-linguistically difficult.

The problem of SRL for low-resource languages is an important one to solve, as solutions pave the way for a wide range of applications: Accurate identification of the semantic roles of entities is a critical step for any application sensitive to semantics, from information retrieval to machine translation to question answering.

In this work, we explore models that minimize the need for high-resource supervision. We examine approaches in a joint setting where we marginalize over latent syntax to find the optimal semantic role assignment; and a pipeline setting where we first induce an unsupervised grammar. We find that the joint approach is a viable alternative for making reasonable semantic role predictions, outperforming the pipeline models. These models can be effectively trained with access to only SRL annotations, and mark a state-of-the-art contribution for low-resource SRL.

To better understand the effect of the low-resource grammars and features used in these models, we further include comparisons with (1) models that use higher-resource versions of the same features; (2) state-of-the-art high resource models; and (3) previous work on low-resource grammar induction. In sum, this paper makes several experimental and modeling contributions, summarized below.

Experimental contributions:
• Comparison of pipeline and joint models for SRL.
• Subtractive experiments that consider the removal of supervised data.
• Analysis of the induced grammars in unsupervised, distantly-supervised, and joint training settings.

See our paper from ACL ‘14
Shallow Semantics

Our End Task:

- Representation: **Semantic Role Labeling** (SRL)
  - Intuitively captures *who did what to whom, when and where*
  - Similar to open-domain relation extraction
- Languages: Catalan, Czech, German, English, Spanish, Chinese
Syntax

Intermediate Tasks:
- Dependency parsing
  - Captures the structure of the sentence
  - Diverges from the shallow semantics representation
- Part-of-speech (POS) tagging
Background:
The Supervised SRL Pipeline

**Pipelined Training:** Train each component of the pipeline independently using the predictions of the previous stage(s) as features.

<table>
<thead>
<tr>
<th>Tool</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Semantic role labeler</td>
<td></td>
</tr>
<tr>
<td>Dependency parser</td>
<td></td>
</tr>
<tr>
<td>Part-of-speech tagger</td>
<td></td>
</tr>
<tr>
<td>Morphological feature extractor</td>
<td></td>
</tr>
<tr>
<td>Lemmatizer</td>
<td></td>
</tr>
</tbody>
</table>

- **Costly**
- Does not work well on informal text
- Resources not available across languages

- President Morsi creates unrest
- president morsi create unrest
- tense=p
Background: The Supervised SRL Pipeline

Pipelined Training: Train each component of the pipeline independently using the predictions of the previous stage(s) as features.

<table>
<thead>
<tr>
<th>Tool</th>
<th>Our Emphasis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semantic role labeler</td>
<td></td>
</tr>
<tr>
<td>Dependency parser</td>
<td></td>
</tr>
<tr>
<td>Part-of-speech tagger</td>
<td></td>
</tr>
<tr>
<td>Morphological feature extractor</td>
<td></td>
</tr>
<tr>
<td>Lemmatizer</td>
<td></td>
</tr>
</tbody>
</table>

- **Role**
  - NN
  - num=s
- **Holder**
  - NNP
  - num=s
  - per=3s
- **Agent**
  - VBZ
  - tense=p
- **Theme**
  - NN
  - num=s

```
[president morsi create unrest]
President Morsi creates unrest
```
This Talk in a Nutshell

• We want to do SRL in a new language.
• Syntax helps, but is very expensive to annotate.

<table>
<thead>
<tr>
<th>Supervised Annotation</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semantic roles</td>
<td>$$$</td>
</tr>
<tr>
<td>Dependency parses</td>
<td>$$$$$$$</td>
</tr>
<tr>
<td>Part-of-speech (POS) tags</td>
<td>$$</td>
</tr>
<tr>
<td>Morphology</td>
<td>$$$</td>
</tr>
<tr>
<td>Lemmas</td>
<td>$</td>
</tr>
</tbody>
</table>
This Talk in a Nutshell

• We want to do SRL in a new language.
• Syntax helps, but is very expensive to annotate.

• Having annotated syntax would be nice, but we can make progress without it!
Supervised SRL as a Factor Graph

- Jointly *identify* and *classify* semantic roles
- One variable per possible labeled edge
- $O(n^2)$ independent logistic regressions

Juan_Carlos abdica reino

• Jointly *identify* and *classify* semantic roles
• One variable per possible labeled edge
• $O(n^2)$ independent logistic regressions
Supervised SRL as a Factor Graph

- Jointly identify and classify semantic roles
- One variable per possible labeled edge
- $O(n^2)$ independent logistic regressions
Supervised SRL as a Factor Graph

- Jointly identify and classify semantic roles
- One variable per possible labeled edge
- $O(n^2)$ independent logistic regressions
High Resource
Pro: high accuracy parsers
Con: expensive

Pipeline
Pro: easy to throw in lots of features
Con: propagation of errors

Low Resource
Pro: cheap, easily deployable
Con: low accuracy latent syntax

Joint
Pro: confidence flows between levels of the model
Con: features must permit efficient inference

Marginalized
DMV+C
DMV
Contributions

• Experimental contributions:
  – Comparison of pipeline and joint models for SRL.
  – **Subtractive experiments** that consider the removal of supervised data.
  – Analysis of the induced grammars in (1) unsupervised, (2) distantly-supervised, and (3) joint training settings.

• **Modeling contributions:**
  – **Simpler joint CRF** for syntactic and semantic dependency parsing than previously reported.
  – **New application** of unsupervised **grammar induction**: low-resource SRL.
  – **Constrained grammar induction** using SRL for distant-supervision.
  – Use of **Brown clusters** in place of POS tags for low-resource SRL.
Three Training Settings for Latent Syntax

1. Fully Unsupervised (DMV)
2. Distantly Supervised (DMV+C)
3. Jointly Learned with SRL (Marginalized)
1. Fully Unsupervised

A. Brown clusters (Brown et al., 1992) in place of POS tags
   • Clusters formed by hierarchical clustering; maximizes likelihood under latent-class bigram model

B. Syntax from Dependency Model with Valence (DMV) (Klein & Manning, 2004)
   • Children generated recursively
   • Viterbi EM training (Spitkovsky et al., 2010)
2. Distantly Supervised

• Viterbi EM training of DMV
• Observes semantic graph during training
• Constrain a CKY parser in E-step to respect SRL

Algorithm:
1. Define DMV as a PCFG (Cohn et al., 2010)
2. CKY parse (Younger, 1967; Aho and Ullman, 1972)
3. Populate cells with SRL-compatible non-terminals
2. Distantly Supervised

- Viterbi EM training of DMV
- Observes semantic graph during training
- Constrain a CKY parser in E-step to respect SRL

Algorithm:
1. Define DMV as a PCFG (Cohn et al., 2010)
2. CKY parse (Younger, 1967; Aho and Ullman, 1972)
3. Populate cells with SRL-compatible non-terminals
2. Distantly Supervised

- Viterbi EM training of DMV
- Observes semantic graph during training
- Constrain a CKY parser in E-step to respect SRL

Algorithm:
1. Define DMV as a PCFG (Cohn et al., 2010)
2. CKY parse (Younger, 1967; Aho and Ullman, 1972)
3. Populate cells with SRL-compatible non-terminals
3. Joint Model

Marginalize over latent syntax to find the optimal semantic role assignment

• **Model:** Slight simplification of Naradowsky et al. (2012). Jointly *identify* and *classify* semantic roles.

• **Inference:** Belief propagation with inside-outside algorithm embedded in global factor (Smith & Eisner, 2008)

• **Brown clusters** in place of **POS** tags
How do we encode a syntactic dependency tree with binary variables?
3. Joint Model

How do we encode a syntactic dependency tree with binary variables?
How do we encode a syntactic dependency tree with binary variables?
How do we encode a syntactic dependency tree with binary variables?
3. Joint Model

Now we jointly encode the **semantic roles** and the **syntactic dependency tree**.
3. Joint Model

Now we jointly encode the semantic roles and the syntactic dependency tree.
Features and Feature Selection

• Define millions of features using 100+ feature templates
• Incorporate feature ideas from:
  – Koo et al. (2008)
  – Björkelund et al. (2009)
  – Zhao et al. (2009)

What about pairs of unigram templates?

Unigram Templates:

- word(p)
- lemma(p)
- pos(p)
- bc0(p)
- bc1(p)
- morpho(p)
- deprel(p)
- lc(p)
- chpre5(p)
- capitalized(p)
- wordTopN(p)
- morpho1(p)
- morpho2(p)
- morpho3(p)
- eachmorpho(p)
Features and Feature Selection

Use Information Gain (IG) to find top unigram templates (Martins et al., 2011)

\[
IG_{a,m} = \sum_{f \in T_m} \sum_{x_a} p(f, x_a) \log_2 \frac{p(f, x_a)}{p(f)p(x_a)}
\]

Then combine top unigram templates to find top **bigram** templates.
Experiments

Datasets:

– Semantic Roles:
  • CoNLL-2009 Shared Task
  • Languages: Catalan, Czech, German, English, Spanish, Chinese

– Grammar Induction:
  • Additional experiments on WSJ portion of Penn Treebank for comparability

– Brown Clusters:
  • Wikipedia

<table>
<thead>
<tr>
<th>CoNLL-2009 Supervised Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semantic roles</td>
</tr>
<tr>
<td>Dependency parses</td>
</tr>
<tr>
<td>Part-of-speech tags</td>
</tr>
<tr>
<td>Morphological features</td>
</tr>
<tr>
<td>Lemmas</td>
</tr>
</tbody>
</table>
Experiments

• Abbreviations for Latent Syntax:
  1. Fully Unsupervised (DMV)
  2. Distantly Supervised (DMV+C)
  3. Jointly Learned with SRL (Marginalized)

• Abbreviations for Tag Types:
  1. Part-of-speech Tags (pos)
  2. Brown Clusters (bc)
Grammar Induction Analysis

Does the latent syntax look any good?

<table>
<thead>
<tr>
<th>Dependency Parser</th>
<th>Avg UAS</th>
<th>Catalan</th>
<th>Czech</th>
<th>German</th>
<th>English</th>
<th>Spanish</th>
<th>Chinese</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised</td>
<td>87.1</td>
<td>89.4</td>
<td>85.3</td>
<td>89.6</td>
<td>88.4</td>
<td>89.2</td>
<td>80.7</td>
</tr>
</tbody>
</table>
# Grammar Induction Analysis

Does the latent syntax look any good?

<table>
<thead>
<tr>
<th>Dependency Parser</th>
<th>Avg UAS</th>
<th>Catalan</th>
<th>Czech</th>
<th>German</th>
<th>English</th>
<th>Spanish</th>
<th>Chinese</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised</td>
<td>87.1</td>
<td>89.4</td>
<td>85.3</td>
<td>89.6</td>
<td>88.4</td>
<td>89.2</td>
<td>80.7</td>
</tr>
<tr>
<td>Marginalized, IG(_B)</td>
<td>50.2</td>
<td>52.4</td>
<td>43.4</td>
<td>41.3</td>
<td>52.6</td>
<td>55.2</td>
<td>56.2</td>
</tr>
<tr>
<td>Marginalized, IG(_C)</td>
<td>43.8</td>
<td>50.3</td>
<td>45.8</td>
<td>27.2</td>
<td>44.2</td>
<td>46.3</td>
<td>48.5</td>
</tr>
<tr>
<td>DMV+C (bc)</td>
<td>40.2</td>
<td>46.3</td>
<td>37.5</td>
<td>28.7</td>
<td>40.6</td>
<td>50.4</td>
<td>37.5</td>
</tr>
<tr>
<td>DMV+C (pos)</td>
<td>37.5</td>
<td>50.2</td>
<td>34.9</td>
<td>21.5</td>
<td>36.9</td>
<td>49.8</td>
<td>32</td>
</tr>
<tr>
<td>DMV (pos)</td>
<td>30.2</td>
<td>45.3</td>
<td>22.7</td>
<td>20.9</td>
<td>32.9</td>
<td>41.9</td>
<td>17.2</td>
</tr>
<tr>
<td>DMV (bc)</td>
<td>22.1</td>
<td>18.8</td>
<td>32.8</td>
<td>19.6</td>
<td>22.4</td>
<td>20.5</td>
<td>18.6</td>
</tr>
</tbody>
</table>

Gap between supervised and not-so-supervised parser is very large
Subtractive Experiments

Effectiveness of our joint models as the available supervision is decreased

CoNLL-2009 Supervised Data
- Semantic roles
- Dependency parses
- Morphology
- Part-of-speech tags
- Lemmas

SRL F1

Supervision Removed
(# Feat Templates)

Catalan
Spanish
German
Subtractive Experiments

Effectiveness of our joint models as the available supervision is decreased

![Graph showing SRL F1 scores for Catalan, Spanish, and German languages with different amounts of supervision.](image-url)
Subtractive Experiments

Effectiveness of our joint models as the available supervision is decreased

CoNLL-2009 Supervised Data
- Semantic roles
- Dependency parses
- Morphology
- Part-of-speech tags
- Lemmas

SRL F1

Supervision Removed
(# Feat Templates)

- (127+32)
- Dep (40+32)
- Mor (30+32)
- POS (23+32)
- Lem (21+32)

Catalan
Spanish
German

(127+32)
(40+32)
(30+32)
(23+32)
(21+32)
Subtractive Experiments

Effectiveness of our joint models as the available supervision is decreased
Subtractive Experiments

Effectiveness of our joint models as the available supervision is decreased
Learning Curves

Lowest resource setting: joint training yields higher SRL F1 than distant-supervision.
## SRL Main Results

<table>
<thead>
<tr>
<th>Parse</th>
<th>SRL Approach</th>
<th>Feat</th>
<th>Avg SRL F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gold</td>
<td>Pipeline</td>
<td>$I_{G_c}$</td>
<td>84.98</td>
</tr>
<tr>
<td></td>
<td>Pipeline</td>
<td>$I_{G_B}$</td>
<td>84.74</td>
</tr>
<tr>
<td></td>
<td>Naradowsky et al. (2012)</td>
<td></td>
<td>72.73</td>
</tr>
</tbody>
</table>

$I_{G_c}$ Features from Information Gain template selection, Coarse-Grained properties

$I_{G_B}$ Features from Information Gain template selection, Björkelund et al. (2009) properties
## SRL Main Results

<table>
<thead>
<tr>
<th>Parse</th>
<th>SRL Approach</th>
<th>Feat</th>
<th>Avg SRL F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gold</td>
<td>Pipeline</td>
<td>IG&lt;sub&gt;c&lt;/sub&gt;</td>
<td>84.98</td>
</tr>
<tr>
<td></td>
<td>Pipeline</td>
<td>IG&lt;sub&gt;B&lt;/sub&gt;</td>
<td>84.74</td>
</tr>
<tr>
<td></td>
<td>Naradowsky et al. (2012)</td>
<td></td>
<td>72.73</td>
</tr>
<tr>
<td></td>
<td>Björkelund et al. (2009)</td>
<td></td>
<td>81.55</td>
</tr>
<tr>
<td></td>
<td>Zhao et al. (2009)</td>
<td></td>
<td>80.85</td>
</tr>
<tr>
<td></td>
<td>Pipeline</td>
<td>IG&lt;sub&gt;c&lt;/sub&gt;</td>
<td>78.03</td>
</tr>
<tr>
<td></td>
<td>Pipeline</td>
<td>IG&lt;sub&gt;B&lt;/sub&gt;</td>
<td>75.68</td>
</tr>
</tbody>
</table>

**IG<sub>c** Features from Information Gain template selection, Coarse-Grained properties

**IG<sub>B** Features from Information Gain template selection, Björkelund et al. (2009) properties
### SRL Main Results

<table>
<thead>
<tr>
<th>Parse</th>
<th>SRL Approach</th>
<th>Feat</th>
<th>Avg SRL F1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SRL</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Features**

- $\text{IG}_c$: Features from Information Gain template selection, Coarse-Grained properties
- $\text{IG}_b$: Features from Information Gain template selection, Björkelund et al. (2009) properties
# SRL Main Results

<table>
<thead>
<tr>
<th>Parse</th>
<th>SRL Approach</th>
<th>Feat</th>
<th>Avg SRL F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gold</td>
<td>Pipeline</td>
<td>IGₐ</td>
<td>84.98</td>
</tr>
<tr>
<td></td>
<td>Pipeline</td>
<td>IGₐ</td>
<td>84.74</td>
</tr>
<tr>
<td></td>
<td>Naradowsky et al. (2012)</td>
<td></td>
<td>72.73</td>
</tr>
<tr>
<td></td>
<td>Björkelund et al. (2009)</td>
<td></td>
<td>81.55</td>
</tr>
<tr>
<td></td>
<td>Zhao et al. (2009)</td>
<td></td>
<td>80.85</td>
</tr>
<tr>
<td></td>
<td>Pipeline</td>
<td>IGₐ</td>
<td>78.03</td>
</tr>
<tr>
<td></td>
<td>Pipeline</td>
<td>IGₐ</td>
<td>75.68</td>
</tr>
<tr>
<td></td>
<td>Joint</td>
<td>IGₐ</td>
<td>72.48</td>
</tr>
<tr>
<td></td>
<td>Joint</td>
<td>IGₐ</td>
<td>72.40</td>
</tr>
<tr>
<td></td>
<td>Naradowsky et al. (2012)</td>
<td></td>
<td>71.27</td>
</tr>
<tr>
<td></td>
<td>Distant (DMV+C)</td>
<td></td>
<td>70.08</td>
</tr>
<tr>
<td></td>
<td>Pipeline</td>
<td>IGₐ</td>
<td>65.61</td>
</tr>
<tr>
<td></td>
<td>Distant (DMV+C)</td>
<td></td>
<td>69.26</td>
</tr>
<tr>
<td></td>
<td>Pipeline</td>
<td>IGₐ</td>
<td>66.81</td>
</tr>
</tbody>
</table>

**IGₐ** Features from Information Gain template selection, Coarse-Grained properties

**IGₐ** Features from Information Gain template selection, Björkelund et al. (2009) properties
Comparison with Work in Grammar Induction in Low-Resource Setting

WSJ portion of Penn Treebank:

<table>
<thead>
<tr>
<th>Approach</th>
<th>Distant Supervision</th>
<th>Unlabeled Syntactic Dependency Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spitkovsky et al (2010)</td>
<td>None</td>
<td>44.8</td>
</tr>
<tr>
<td>Spitkovsky et al (2013)</td>
<td>None</td>
<td>64.4</td>
</tr>
<tr>
<td>Spitkovsky et al (2010)</td>
<td>HTML</td>
<td>50.4</td>
</tr>
<tr>
<td>Naseem and Barzilay (2011)</td>
<td>ACE05</td>
<td>59.4</td>
</tr>
<tr>
<td>DMV</td>
<td>None</td>
<td>24.8</td>
</tr>
<tr>
<td>DMV+C</td>
<td>SRL</td>
<td>44.8</td>
</tr>
<tr>
<td>Marginalized, IG(_c)</td>
<td>SRL</td>
<td>48.8</td>
</tr>
<tr>
<td>Marginalized, IG(_B)</td>
<td>SRL</td>
<td>58.9</td>
</tr>
</tbody>
</table>

- MBR decoding of marginalized grammars best DMV method
- May get gains with better search to break out of local optima
Is dependency accuracy the right evaluation metric?

In the joint model, higher dependency accuracy (UAS) does not *always* correlate with higher Labeled F1 on SRL.

<table>
<thead>
<tr>
<th>Parse</th>
<th>SRL Approach</th>
<th>Feat</th>
<th>English UAS</th>
<th>English SRL F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joint (Marginalized)</td>
<td>Joint</td>
<td>IG_c</td>
<td>44.2</td>
<td>76.16</td>
</tr>
<tr>
<td>Joint</td>
<td>Joint</td>
<td>IG_B</td>
<td>52.6</td>
<td>75.57</td>
</tr>
</tbody>
</table>
Conclusions

• Semantic role labeling doesn’t necessarily require a long costly pipeline of NLP tools (cf. Boxwell et al. (2011); Naradowsky et al. (2012))
• “Quality” of the latent syntax has a big effect (especially with limited end-task training data)
• Joint models seem to outperform the pipeline models in the low-resource setting
Questions?