Deep Semantic Role Labeling: What works and what’s next

Luheng He†, Kenton Lee†, Mike Lewis‡ and Luke Zettlemoyer†*

† Paul G. Allen School of Computer Science & Engineering, Univ. of Washington,
‡ Facebook AI Research
* Allen Institute for Artificial Intelligence
Semantic Role Labeling (SRL)

**role label**

**predicate**  
**argument**

who  
what  
when  
where  
why  
...

**Applications**

Question Answering  
Information Extraction  
Machine Translation
Semantic Role Labeling (SRL) - Example

The robot *broke* my mug with a wrench.

My mug *broke* into pieces.
My mug broke into pieces.

The robot *broke* my mug with a wrench.

My mug *broke* into pieces.
My mug broke into pieces.

The robot *broke* my mug with a wrench.

Frame: *break.01*

<table>
<thead>
<tr>
<th>role</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARG0</td>
<td>breaker</td>
</tr>
<tr>
<td>ARG1</td>
<td>thing broken</td>
</tr>
<tr>
<td>ARG2</td>
<td>instrument</td>
</tr>
<tr>
<td>ARG3</td>
<td>pieces (final state)</td>
</tr>
</tbody>
</table>
The Proposition Bank (PropBank)
Paul Kingsbury and Martha Palmer. From Treebank to PropBank. 2002

Core roles:
Verb-specific roles (ARG0-ARG5) defined in frame files

Frame: *break.01*

<table>
<thead>
<tr>
<th>role</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARG0</td>
<td>breaker</td>
</tr>
<tr>
<td>ARG1</td>
<td>thing broken</td>
</tr>
<tr>
<td>ARG2</td>
<td>instrument</td>
</tr>
</tbody>
</table>

Frame: *buy.01*

<table>
<thead>
<tr>
<th>role</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARG0</td>
<td>buyer</td>
</tr>
<tr>
<td>ARG1</td>
<td>thing bough</td>
</tr>
<tr>
<td>ARG2</td>
<td>seller</td>
</tr>
<tr>
<td>ARG3</td>
<td>price paid</td>
</tr>
<tr>
<td>ARG4</td>
<td>benefactive</td>
</tr>
</tbody>
</table>

Adjunct roles:
(ARGM-) shared across verbs

<table>
<thead>
<tr>
<th>role</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>TMP</td>
<td>temporal</td>
</tr>
<tr>
<td>LOC</td>
<td>location</td>
</tr>
<tr>
<td>MNR</td>
<td>manner</td>
</tr>
<tr>
<td>DIR</td>
<td>direction</td>
</tr>
<tr>
<td>CAU</td>
<td>cause</td>
</tr>
<tr>
<td>PRP</td>
<td>purpose</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>
The Proposition Bank (PropBank)

Paul Kingsbury and Martha Palmer. *From Treebank to PropBank*. 2002

Core roles:
Verb-specific roles (ARG0-ARG5) defined in frame files

```
Frame: break.01
role   | description
ARG0   | breaker
ARG1   | thing broken
ARG2   | instrument

Frame: buy.01
role   | description
ARG0   | buyer
ARG1   | thing bough
ARG2   | seller
ARG3   | price paid
ARG4   | benefactive
```

Adjunct roles:
(ARGM-) shared across verbs

```
role   | description
TMP    | temporal
LOC    | location
MNR    | manner
DIR    | direction
CAU    | cause
PRP    | purpose
```

Annotated on top of the Penn Treebank Syntax

PropBank Annotation Guidelines, Bonial et al., 2010
SRL Systems

Pipeline Systems

sentence, predicate
syntactic features

argument id.
candidate argument spans

labeling
labeled arguments

ILP/DP
prediction

Punyakanok et al., 2008
Täckström et al., 2015
FitzGerald et al., 2015
SRL Systems

### Pipeline Systems
- sentence, predicate
- syntactic features
- candidate argument spans
- argument id.
- labeling
- labeled arguments
- ILP/DP
- prediction

Punyakanok et al., 2008
Täckström et al., 2015
FitzGerald et al., 2015

### End-to-end Systems
- sentence, predicate
- context window features
- Deep BiLSTM + CRF layer
- BIO sequence
- Viterbi
- prediction

Collobert et al., 2011
Zhou and Xu, 2015
Wang et al., 2015
SRL Systems

**Pipeline Systems**
- sentence, predicate
- syntactic features
- argument id.
- candidate argument spans
- labeling
- labeled arguments
- ILP/DP
- prediction

- Punyakanok et al., 2008
- Täckström et al., 2015
- FitzGerald et al., 2015

**End-to-end Systems**
- sentence, predicate
- context window features
- Deep BiLSTM + CRF layer
- BIO sequence
- Viterbi
- prediction

- Collobert et al., 2011
- Zhou and Xu, 2015
- Wang et al., 2015

***This work***
- sentence, predicate
- Deep BiLSTM
- BIO sequence
- Hard constraints
- prediction
the cats love hats
(1) Deep BiLSTM tagger
<table>
<thead>
<tr>
<th></th>
<th>B-ARG0 0.4</th>
<th>B-ARG0 0.1</th>
<th>B-ARG0 0.001</th>
<th>B-ARG0 0.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>I-ARG0</td>
<td>0.05</td>
<td>0.5</td>
<td>0.001</td>
<td>0.1</td>
</tr>
<tr>
<td>B-ARG1</td>
<td>0.5</td>
<td>0.1</td>
<td>0.001</td>
<td>0.7</td>
</tr>
<tr>
<td>I-ARG1</td>
<td>0.03</td>
<td>0.2</td>
<td></td>
<td>0.2</td>
</tr>
</tbody>
</table>

(1) Deep BiLSTM tagger

(2) Highway connections

```
the [ ]
cats [ ]
love [V]
hats [ ]
```
(1) Deep BiLSTM tagger

(2) Highway connections

(3) Viterbi decoding with hard constraints
Model - Highway Connections

\[ h_t = F(c_{t-1}, h_{t-1}, x_t) \]

References:
- Deep Residual Networks, Kaiming He, ICML 2016 Tutorial
- Training Very Deep Networks, Srivastava et al., 2015
Model - Highway Connections

\[ h_t = \sigma( \mathcal{F}(c_{t-1}, h_{t-1}, x_t) ) \]

- **recurrent input** \( h_{t-1} \)
- **input from the previous layer** \( x_t \)
- **output to the next layer** \( h_t \)
- **shortcut** \( x_t \)

References:
- Deep Residual Networks, Kaiming He, ICML 2016 Tutorial
- Training Very Deep Networks, Srivastava et al., 2015
Model - Highway Connections

- **Input from the previous layer**
- **Output to the next layer**

Non-linearity:

\[ F(c_{t-1}, h_{t-1}, x_t) \]

**Shortcut**: \[ x_t \]

**Recurrent input**: \[ h_{t-1} \]

**Input from the previous layer**: \[ x_t \]

Gated highway network:

\[ r_t \circ h_t + (1 - r_t) \circ x_t \]

\[ r_t = \sigma(f(h_{t-1}, x_t)) \]

References:

- Deep Residual Networks, Kaiming He, ICML 2016 Tutorial
- Training Very Deep Networks, Srivastava et al., 2015
Model - Viterbi Decoding with Hard Constraints

Greedy Output

Softmax

BiLSTM layers ...

the [ ] cats [ ] love [V] hats [ ]
Model - Viterbi Decoding with Hard Constraints

Greedy Output

BIO inconsistency: invalid transition

Softmax

BiLSTM layers ...

argmax

the []
cats []
love [V]
hats []
Heuristic transition scores

\[ s(B-\text{ARG0} \rightarrow I-\text{ARG0}) = 0 \]
\[ s(B-\text{ARG1} \rightarrow I-\text{ARG0}) = -\infty \]

BIO inconsistency: invalid transition

Greedy Output

Softmax

BiLSTM layers …

Model - Viterbi Decoding with Hard Constraints
Model - Viterbi Decoding with Hard Constraints

Heuristic transition scores

\[ s(B-\text{ARG0} \rightarrow I-\text{ARG0}) = 0 \]
\[ s(B-\text{ARG1} \rightarrow I-\text{ARG0}) = -\infty \]

Viterbi decoding

<table>
<thead>
<tr>
<th></th>
<th>B-ARG0</th>
<th>I-ARG0</th>
<th>B-ARG1</th>
<th>I-ARG1</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.4</td>
<td>0.05</td>
<td>0.5</td>
<td>0.03</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>0.5</td>
<td>0.1</td>
<td>0.2</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.002</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>0.1</td>
<td>0.7</td>
<td>0.2</td>
<td></td>
</tr>
</tbody>
</table>

Softmax

BiLSTM layers …

the [ ]  
cats [ ]  
love [V]  
hats [ ]
Other Implementation Details …

- 8 layer BiLSTMs with 300D hidden layers.
- 100D GloVe embeddings, updated during training.
- **Orthonormal initialization** for LSTM weight matrices (Saxe et al., 2013)
- 0.1 **variational dropout** between layers (Gal and Ghahramani, 2016)
- Trained for 500 epochs.
<table>
<thead>
<tr>
<th>Datasets</th>
<th>CoNLL 2005 (PropBank)</th>
<th>CoNLL 2012 (OntoNotes)</th>
<th>Ablations</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Size</strong></td>
<td>40k sentences</td>
<td>140k sentences</td>
<td></td>
</tr>
<tr>
<td><strong>Domains</strong></td>
<td>• WSJ / newswire</td>
<td>• telephone conversations</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Brown (test-only)</td>
<td>• newswire</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• newsgroups</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• broadcast news</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• broadcast conversation</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• weblogs</td>
<td></td>
</tr>
<tr>
<td><strong>Annotated predicates</strong></td>
<td>Verbs</td>
<td>Added some nominal predicates</td>
<td></td>
</tr>
<tr>
<td>Datasets</td>
<td>CoNLL 2005 Results</td>
<td>CoNLL 2012 (OntoNotes) Results</td>
<td>Ablations</td>
</tr>
<tr>
<td>--------------</td>
<td>---------------------</td>
<td>-------------------------------</td>
<td>-----------</td>
</tr>
<tr>
<td></td>
<td>WSJ Test</td>
<td>Brown (out-domain) Test</td>
<td>*:Ensemble models</td>
</tr>
<tr>
<td>90</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>85</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>80</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>75</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>70</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>65</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>Ours* 2017</td>
<td>Zhou 2015</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ours 2017</td>
<td>FitzGerald* 2015</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Täckström 2015</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Toutanova* 2008</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Punyakanok* 2008</td>
<td></td>
</tr>
</tbody>
</table>
CoNLL 2005 Results

Datasets

- Ours*
  - 2017

- Zhou
  - 2015

CoNLL 2012 (OntoNotes) Results

- Ours
  - 2017

- FitzGerald*
  - 2015

- Täckström
  - 2015

- Toutanova*
  - 2008

- Punyakanok*
  - 2008

*A: Ensemble models

WSJ Test

- Ours*
  - 2017: 84.6

- Ours
  - 2017: 83.1

- Zhou
  - 2015: 82.8

- FitzGerald*
  - 2015: 80.3

- Täckström
  - 2015: 79.9

- Toutanova*
  - 2008: 80.3

- Punyakanok*
  - 2008: 79.4

Brown (out-domain) Test

- Ours*
  - 2017: 82.8

- Ours
  - 2017: 83.1

- Zhou
  - 2015: 84.6

- FitzGerald*
  - 2015: 80.3

- Täckström
  - 2015: 79.9

- Toutanova*
  - 2008: 80.3

- Punyakanok*
  - 2008: 79.4

Ablations
CoNLL 2005 Results

Datasets

CoNLL 2012 (OntoNotes) Results

Ablations

Ours*
2017

Ours
2017

Zhou
2015

FitzGerald*
2015

Täckström
2015

Toutanova*
2008

Punyakanok*
2008

*: Ensemble models

WSJ Test

Brown (out-domain) Test

84.6
83.1
82.8
80.3
79.9
80.3
79.4

73.6
72.1
69.4
72.2
71.3
68.8
67.8

80.8
71.3
72.1
73.6
79.9
80.3
82.8

79.4
80.3
80.8
81.1
84.6
83.1
84.6

71.3
72.1
73.6
80.3
80.8
81.1
82.8

84.6
83.1
82.8
80.3
79.9
80.3
79.4

73.6
72.1
69.4
72.2
71.3
68.8
67.8

WSJ Test

Brown (out-domain) Test

*: Ensemble models
CoNLL 2005 Results

Datasets

CoNLL 2012 (OntoNotes) Results

Ablations

WSJ Test
Brown (out-domain) Test
*: Ensemble models

Ours*
Ours
Zhou
FitzGerald*
Täckström
Toutanova*
Punyakanok*

WSJ Test
Brown (out-domain) Test

BiLSTM models
Pipeline models
CoNLL 2012 (OntoNotes) Results

- **Ours* 2017**: 83.4
- **Ours 2017**: 81.7
- **Zhou 2015**: 81.5
- **FitzGerald* 2015**: 80.2
- **Täckström 2015**: 79.4
- **Pradhan 2013**: 77.5

*: Ensemble models

**Datasets**
- CoNLL 2005

**Results**
- CoNLL 2012 (OntoNotes) Results

**Ablations**

- BiLSTM models
- Pipeline models
Without dropout, model overfits at ~300 epochs.

Without orthonormal initialization, the deep model learns very slowly.

Datasets
- CoNLL 2005
- CoNLL 2012 (OntoNotes)

Results
- CoNLL 2005
- CoNLL 2012 (OntoNotes)

Ablations
(single model, on CoNLL05 Dev)
- Full model
- No highway
- No orthonormal init.
- No dropout

F1 on Dev. Set

Num. Epochs
What can we learn from the results?

1. What’s in the remaining 17%? When does the model still struggle?
What can we learn from the results?

1. What’s in the remaining 17%? When does the model still struggle?

2. BiLSTM-based models are very accurate even without syntax. But can we conclude syntax is no longer useful in SRL?
Question (1): When does the model make mistakes?

Analysis
— Error breakdown with oracle transformation
— E.g. tease apart labeling errors and boundary errors
— Link the error types to known linguistic phenomena, e.g. prepositional phrase (PP) attachment
Error Breakdown
Oracle Transformations

Fix Label:

[We] fly to NYC tomorrow.
Error Breakdown
Oracle Transformations

[We] fly to NYC tomorrow.

Fix Label:

ARG0
ARG1

Labeling error 29%
Error Breakdown
Oracle Transformations

Fix
Label:

[We] fly to NYC tomorrow.

Labeling error 29%

Split/Merge
span:

I eat [pasta with delight].

I eat [pasta] [with broccoli].
**Oracle Transformations (Error Breakdown)**

**Error Breakdown**

**Fix Label:**

[We] *fly* to NYC tomorrow.

**Labeling error 29%**

**Split/Merge span:**

I *eat* [pasta with delight].

Attachment error 25%

I *eat* [pasta] [with broccoli].
### Confusion Matrix for Labeling Errors (Column Normalized)

<table>
<thead>
<tr>
<th></th>
<th>A0</th>
<th>A1</th>
<th>A2</th>
<th>A3</th>
<th>ADV</th>
<th>DIR</th>
<th>LOC</th>
<th>MNR</th>
<th>PNC</th>
<th>TMP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>pred.</strong></td>
<td><strong>gold</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A0</td>
<td>55</td>
<td>11</td>
<td>13</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A1</td>
<td>76</td>
<td>46</td>
<td>0</td>
<td>0</td>
<td>22</td>
<td>11</td>
<td>10</td>
<td>25</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>A2</td>
<td>11</td>
<td>23</td>
<td>48</td>
<td>15</td>
<td>56</td>
<td>33</td>
<td>41</td>
<td>25</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>A3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>25</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>ADV</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>-</td>
<td>15</td>
<td>29</td>
<td>25</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td>DIR</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>4</td>
<td>0</td>
<td>11</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>LOC</td>
<td>5</td>
<td>9</td>
<td>12</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>-</td>
<td>10</td>
<td>0</td>
<td>14</td>
</tr>
<tr>
<td>MNR</td>
<td>3</td>
<td>0</td>
<td>12</td>
<td>26</td>
<td>33</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>PNC</td>
<td>0</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>0</td>
<td>11</td>
<td>4</td>
<td>2</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>TMP</td>
<td>0</td>
<td>8</td>
<td>5</td>
<td>0</td>
<td>41</td>
<td>11</td>
<td>26</td>
<td>6</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
### Labeling Errors

Confusion matrix for labeling errors (column normalized)

<table>
<thead>
<tr>
<th></th>
<th>gold A0</th>
<th>A1</th>
<th>A2</th>
<th>A3</th>
<th>ADV</th>
<th>DIR</th>
<th>LOC</th>
<th>MNR</th>
<th>PNC</th>
<th>TMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>pred. A0</td>
<td>55</td>
<td>11</td>
<td>13</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A1</td>
<td>78</td>
<td>46</td>
<td>0</td>
<td>0</td>
<td>22</td>
<td>11</td>
<td>10</td>
<td>25</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>A2</td>
<td>11</td>
<td>23</td>
<td>48</td>
<td>15</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>25</td>
<td>14</td>
</tr>
<tr>
<td>A3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>25</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>ADV</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>15</td>
<td>29</td>
<td>25</td>
<td>36</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DIR</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>4</td>
<td>0</td>
<td>11</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>LOC</td>
<td>5</td>
<td>9</td>
<td>12</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>MNR</td>
<td>3</td>
<td>0</td>
<td>12</td>
<td>26</td>
<td>33</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>PNC</td>
<td>0</td>
<td>3</td>
<td>4</td>
<td>0</td>
<td>11</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TMP</td>
<td>0</td>
<td>8</td>
<td>5</td>
<td>0</td>
<td>41</td>
<td>11</td>
<td>26</td>
<td>6</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

- ARG2 is often confused with certain adjuncts (DIR, LOC, MNR), why?
Labeling Errors

Confusion matrix for labeling errors (column normalized)

- ARG2 is often confused with certain adjuncts (DIR, LOC, MNR), why?

Predicate: **move**
- Arg0-PAG: mover
- Arg1-PPT: moved
- Arg2-GOL: destination
- Arg3-VSP: aspect, domain in which arg1 moving

Predicate: **cut**
- Arg0-PAG: intentional cutter
- Arg1-PPT: thing cut
- Arg2-DIR: medium, source
- Arg3-MNR: instrument, unintentional cutter
- Arg4-GOL: beneficiary

Predicate: **strike**
- Arg0-PAG: Agent
- Arg1-PPT: Theme(-Creation)
- Arg2-MNR: Instrument
Confusion matrix for labeling errors (column normalized)

- ARG2 is often confused with certain adjuncts (DIR, LOC, MNR), why?

**Predicate:** move
- Arg0-PAG: mover
- Arg1-PPT: moved
- Arg2-GOL: destination
- Arg3-VSP: aspect, domain in which arg1 moving

**Predicate:** cut
- Arg0-PAG: intentional cutter
- Arg1-PPT: thing cut
- Arg2-DIR: medium, source
- Arg3-MNR: instrument, unintentional cutter
- Arg4-GOL: beneficiary

**Predicate:** strike
- Arg0-PAG: Agent
- Arg1-PPT: Theme(-Creation)
- Arg2-MNR: Instrument

**Argument-adjunct distinctions** are difficult even for expert annotators!
Sumimoto *financed* the acquisition from Sears
Wrong PP attachment (attach high)

Sumimoto *financed* the acquisition from Sears

Correct PP attachment (attach low)
PP Attachment

Wrong PP attachment (attach high)

Sumimoto *financed* the acquisition from Sears

Correct PP attachment (attach low)

Wrong SRL spans

Can Syntax Still Help?

Correct SRL spans

Wrong SRL spans

merge
Sumimoto *financed* the acquisition from Sears

Wrong PP attachment
(attach high)

Correct PP attachment
(attach low)

**Attachment mistakes: 25%**.

Categorize the Y spans in:

- \([XY]\rightarrow[X][Y]\) and
- \([X][Y]\rightarrow[XY]\) operations

by gold syntactic labels
PP Attachment

Wrong PP attachment (attach high)

Sumimoto financed the acquisition from Sears

Correct PP attachment (attach low)

Attachment mistakes: 25%.

Categorize the Y spans in:

- [XY]—->[X][Y] and
- [X][Y]—->[XY] operations

by gold syntactic labels

Can Syntax Still Help?

Wrong SRL spans

merge

Correct SRL spans

Error Breakdown

Labeling Errors

Can Syntax Still Help?

Attachment mistakes: 25%.

Categorize the Y spans in:

- [XY]—->[X][Y] and
- [X][Y]—->[XY] operations

by gold syntactic labels

Can Syntax Still Help?
PP Attachment

Wrong PP attachment
(attach high)

Correct PP attachment
(attach low)

Wrong SRL spans
merge
Correct SRL spans

Takeaway

— Traditionally hard tasks, such as argument-adjunct distinction and PP attachment decisions are still challenging!

— Use external information/PropBank frame inventory.
Question (2): Can syntax still help SRL?

Recap
— PropBank SRL is annotated on top of the PTB syntax.
— More than 98% of the gold SRL spans are syntactic constituents.

Analysis
— At decoding time, make predicted argument spans agree with given syntactic structure (unlabeled).
— See if SRL performance increases.
Can Syntax Still Help?
Constrained Decoding with Syntax

[The cats] ∈ Syntax Tree
[hats and the dogs] ∉ Syntax Tree

[The cats] love [hats and the dogs] love bananas.
ARG0  ARG1
Penalize sequence score
Can Syntax Still Help?
Constrained Decoding with Syntax

\[ \text{Sequence score: } \sum_{i=1}^{t} \log p(\text{tag}_t | \text{sentence}) - C \times \sum_{\text{span}} \mathbf{1}(\text{span} \notin \text{Syntax Tree}) \]

- **ARG0**: [The cats]  \( \in \) Syntax Tree
- **ARG1**: [hats and the dogs]  \( \notin \) Syntax Tree

[The cats] *love* [hats and the dogs] *love* bananas.

**Penalize sequence score**

**Penalty strength**

**Num. arguments disagree w\ syntax**
Can Syntax Still Help?
Constrained Decoding with Syntax

[The cats] ∈ Syntax Tree
[hats and the dogs] ∉ Syntax Tree

[The cats] love [hats and the dogs] love bananas.

Sequence score: \[
\sum_{i=1}^{t} \log p(\text{tag}_i \mid \text{sentence}) - C \times \sum_{\text{span}} \mathbf{1}(\text{span} \not\in \text{Syntax Tree})
\]

- Constraints are not locally decomposable.
- A* search (Lewis and Steedman 2014) for a sequence with highest score.
Can Syntax Still Help?
Syntax Decoding Results

Gold: Penn Treebank constituents.
Choe: Parsing as language modeling, Choe and Charniak, 2016 (SOTA)
Charniak: A maximum-entropy-inspired parser, Charniak, 2000
Can Syntax Still Help?
Syntax Decoding Results

Takeaway
— Modest gain when using accurate syntax.
— More improvement: Joint training, use syntactic labels, etc.

Gold: Penn Treebank constituents.
Choe: Parsing as language modeling, Choe and Charniak, 2016 (SOTA)
Charniak: A maximum-entropy-inspired parser, Charniak, 2000
Thank You!

- New state-of-the-art deep network for end-to-end SRL.
- Code and models are publicly available at: https://github.com/luheng/deep_srl
- In-depth error analysis indicating where the models work well and where they still struggle.
- Syntax-based experiments pointing towards directions for future improvements.