Neural Semantic Role Labeling: What works and what’s next
or: What else can we do other than using 1000 LSTM layers :)

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)

**Motivation**

who did what to whom, when and where

- Find out “who did what to whom” in text.
- Given predicate, identify arguments and label them.
Semantic Role Labeling (SRL)

Motivation

who did what to whom, when and where

Applications

Question Answering
Information Extraction
Machine Translation

predicate

role label

argument

who
what
when
where
why...
Semantic Role Labeling (SRL)

The robot *broke* my favorite mug with a wrench.

My mug *broke* into pieces immediately.
The robot *broke* my favorite mug with a wrench.

My mug *broke* into pieces immediately.
The robot *broke* my favorite mug with a wrench.

My mug *broke* into pieces immediately.
The robot *broke* my favorite mug with a wrench.

My mug *broke* into pieces immediately.
The robot *broke* my favorite mug with a wrench.

My mug *broke* into pieces immediately.

**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</td>
</tr>
<tr>
<td>ARG4</td>
<td>broken away from what?</td>
</tr>
</tbody>
</table>
The robot *broke* my favorite mug with a wrench.

My mug *broke* into pieces immediately.

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>
<tr>
<td>ARG4</td>
<td>broken away from what?</td>
</tr>
<tr>
<td>ARGM-TMP</td>
<td>temporal</td>
</tr>
</tbody>
</table>
The Proposition Bank (PropBank)
Paul Kingsbury and Martha Palmer. *From Treebank to PropBank*. 2002
The Proposition Bank (PropBank)

Annotated on top of the Penn Treebank Syntax

PropBank Annotation Guidelines, Bonial et al., 2010
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>
<tr>
<td>ARG3</td>
<td>pieces</td>
</tr>
<tr>
<td>ARG4</td>
<td>broken away from what?</td>
</tr>
</tbody>
</table>

Annotated on top of the Penn Treebank Syntax

PropBank Annotation Guidelines, Bonial et al., 2010
The Proposition Bank (PropBank)

Annotated on top of the Penn Treebank Syntax

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

Frame: \textit{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: \textit{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 bought</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>

PropBank Annotation Guidelines, Bonial et al., 2010

Paul Kingsbury and Martha Palmer. \textit{From Treebank to PropBank}. 2002
The Proposition Bank (PropBank)

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

Annotated on top of the Penn Treebank Syntax

PropBank Annotation Guidelines, Bonial et al., 2010

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 bought</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>
</tbody>
</table>

...
SRL is a hard problem ...
SRL is a hard problem …

• Over 10 years, F1 on the PropBank test set:
  79.4 (Punyakanok 2005) — 80.3 (FitzGerald 2015)
SRL is a hard problem …

• Over 10 years, F1 on the PropBank test set:
  79.4 (Punyakanok 2005) — 80.3 (FitzGerald 2015)

• Many interesting challenges:
  Syntactic alternation
  Prepositional phrase attachment
  Long-range dependencies and common sense
The robot *plays* piano.

ARG0
player

ARG2
instrument

The cafe is *playing* my favorite song.

ARG0
player

ARG1
thing performed

The music *plays* softly.

ARG1
thing performed

ARGM-MNR
The robot *plays* piano.

The cafe is *playing* my favorite song.

The music *plays* softly.
The robot *plays* piano.

The cafe is *playing* my favorite song.

The music *plays* softly.
I eat [pasta] [with delight].

ARG0
-eater

ARG1
-meal

ARGM-MNR
-manner
I eat [pasta] [with delight].

ARG0
  eater

ARG1
  meal

ARGM-MNR
  manner

I eat [pasta with broccoli].

ARG0
  eater

ARG1
  meal
We *flew* to Chicago.
We **flew** to Chicago.

We remember the nice view **flying** to Chicago.
We flew to Chicago.

We remember the nice view flying to Chicago.

We remember John and Mary flying to Chicago.
SRL is even harder for out-domain data …

“Dip chicken breasts into eggs to coat”
SRL is even harder for out-domain data …

“Dip chicken breasts into eggs to coat”
SRL is even harder for out-domain data …

“Dip chicken breasts into eggs to coat”

Active, Ser133-phosphorylated CREB effects transcription of CRE-dependent genes via interaction with the 265-kDa …
Long-term Plan for Improving SRL
Long-term Plan for Improving SRL

Step 1: Collect more data for SRL
- Question-Answer Driven Semantic Role Labeling (QA-SRL)
- Human-in-the-Loop Parsing
Long-term Plan for Improving SRL

Step 1: Collect more data for SRL
  — Question-Answer Driven Semantic Role Labeling (QA-SRL)
  — Human-in-the-Loop Parsing

Step 2: Build accurate SRL model
  — Neural Semantic Role Labeling (for PropBank SRL)
Long-term Plan for Improving SRL

Step 1: Collect more data for SRL
- Question-Answer Driven Semantic Role Labeling (QA-SRL)
- Human-in-the-Loop Parsing

Step 2: Build accurate SRL model
- Neural Semantic Role Labeling (for PropBank SRL)

Step 3: SRL system for many domains
- Future work …
Long-term Plan for Improving SRL

Step 1: Collect more data for SRL
  — Question-Answer Driven Semantic Role Labeling (QA-SRL)
  — Human-in-the-Loop Parsing

Step 2: Build accurate SRL model
  — Neural Semantic Role Labeling (for PropBank SRL)

Step 3: SRL system for many domains
  — Future work …
First Step: Collect more (cheaper) SRL Data
First Step: Collect more (cheaper) SRL Data

**Intuition:** Anyone who understands the meaning of a sentence should be able provide annotation for SRL.
First Step: Collect more (cheaper) SRL Data

**Intuition:** Anyone who understands the meaning of a sentence should be able provide annotation for SRL.

**Challenge:** Complicated annotation process of traditional SRL.
First Step: Collect more (cheaper) SRL Data

**Intuition:** Anyone who understands the meaning of a sentence should be able provide annotation for SRL.

**Challenge:** Complicated annotation process of traditional SRL.

**Solution:** Design a simpler annotation scheme!
Question-Answer Driven SRL (QA-SRL)

Given sentence and a verb:

Last month, we saw the Grand Canyon flying to Chicago.
Question-Answer Driven SRL (QA-SRL)

Given sentence and a verb:

Last month, we saw the Grand Canyon flying to Chicago.

Step 1: Ask a question about the verb:

Who was flying?
Question-Answer Driven SRL (QA-SRL)

Given sentence and a verb:

Last month, we saw the Grand Canyon flying to Chicago.

Step 1: Ask a question about the verb:

Who was flying?

Step 2: Answer with words in the sentence:

we
Question-Answer Driven SRL (QA-SRL)

Given sentence and a verb:

Last month, we saw the Grand Canyon *flying* to Chicago.

Step 1: Ask a question about the verb:

Who was flying?

Step 2: Answer with words in the sentence:

we

Step 3: Repeat, write as many Q/A pairs as possible ...

...
Question-Answer Driven SRL (QA-SRL)

Given sentence and a verb:

Last month, we saw the Grand Canyon flying to Chicago.

**Step 1:** Ask a question about the verb:
Who was flying?

**Step 2:** Answer with words in the sentence:
we

**Step 3:** Repeat, write as many Q/A pairs as possible ...
Where did someone fly to? Chicago
When did someone fly? Last month
Question-Answer Driven SRL (QA-SRL)

Given sentence and a verb:

Last month, we saw the Grand Canyon flying to Chicago.

Step 1: Ask a question about the verb:

Who was flying?

Step 2: Answer with words in the sentence:

we

Step 3: Repeat, write as many Q/A pairs as possible …

Where did someone fly to? Chicago

When did someone fly? Last month

Stop until all Q/A pairs are exhausted.
Comparing QA-SRL to PropBank

**Large Role Inventory**

**Predicate**

**Role**

**Argument**

Traditional SRL (PropBank)

**No Role Inventory!**

**Question**

**(Verbal) Predicate**

**Answer**

Question-Answer Driven SRL (QA-SRL)
Last month, we saw the Grand Canyon *flying* to Chicago.
Last month, we saw the Grand Canyon *flying* to Chicago.
Last month, we saw the Grand Canyon *flying* to Chicago.

Non-expert annotated QA-SRL has about **80% agreement** with PropBank.
Long-term Plan for Improving SRL

Step 1: Collect more data for SRL
- Question-Answer Driven Semantic Role Labeling (QA-SRL)
- Human-in-the-Loop Parsing

Step 2: Build accurate SRL model
- Neural Semantic Role Labeling (for PropBank SRL)

Step 3: SRL system for many domains
- Future work …
Long-term Plan for Improving SRL

Step 1: Collect more data for SRL
  — Question-Answer Driven Semantic Role Labeling (QA-SRL)
  — Human-in-the-Loop Parsing

✓

Step 2: Build accurate SRL model
  — Neural Semantic Role Labeling (for PropBank SRL)

Step 3: SRL system for many domains
  — Future work …
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
SRL as BIO Tagging Problem

Input (sentence and predicate): The cats love hats.
Input (sentence and predicate):
The cats love hats.

BIO output:
B-ARG0 I-ARG0 B-V I-ARG1 O

(Begin, Inside, Outside)
SRL as BIO Tagging Problem

Input (sentence and predicate):
The cats love hats.

BIO output:
(Begin, Inside, Outside)

Final SRL output:
the [ ]
cats [ ]
love [V]
hats [ ]
B-ARG0 0.4
I-ARG0 0.05
B-ARG1 0.5
I-ARG1 0.03

B-ARG0 0.1
I-ARG0 0.5
B-ARG1 0.1
I-ARG1 0.2

B-ARG0 0.001
I-ARG0 0.001
B-ARG1 0.001
I-ARG1 0.2

B-ARG0 0.1
I-ARG0 0.1
B-ARG1 0.7
I-ARG1 0.2

(1) Deep BiLSTM tagger
B-ARG0 0.4
I-ARG0 0.05
B-ARG1 0.5
I-ARG1 0.03

B-ARG0 0.1
I-ARG0 0.5
B-ARG1 0.1
I-ARG1 0.2

B-ARG0 0.001
I-ARG0 0.001
B-ARG1 0.001

B-V 0.95

B-ARG0 0.1
I-ARG0 0.1
B-ARG1 0.7
I-ARG1 0.2

(1) Deep BiLSTM tagger
(2) Highway connections
(1) Deep BiLSTM tagger
(2) Highway connections
(3) Variational dropout

18
The cats love hats.

(1) Deep BiLSTM tagger
(2) Highway connections
(3) Variational dropout
(4) Viterbi decoding with hard constraints
Deep BiLSTM Tagger

Output

B-ARG0

I-ARG0

B-V

B-ARG1

argmax

Softmax

Backward LSTM

Forward LSTM

Word + Tag embeddings

the [ ]
cats [ ]
love [V]
hats [ ]
Deep BiLSTM Tagger

Output

Softmax

Backward LSTM

Forward LSTM

Word + Tag embeddings

Predicate

B-ARG0

I-ARG0

B-V

B-ARG1

argmax

the [ ]

cats [ ]

love [V]

hats [ ]

Concatenate: 100dim + 100dim

Deep BiLSTM Tagger

Highway Connections

Variational Dropout

Viterbi Decoding w/ Hard Constraints

B-ARG0

I-ARG0

B-V

B-ARG1

argmax

the [ ]

cats [ ]

love [V]

hats [ ]

Concatenate: 100dim + 100dim

Predicate
Deep BiLSTM Tagger

Output

B-ARG0

I-ARG0

B-V

B-ARG1

argmax

Softmax

Backward LSTM

Forward LSTM

Word + Tag embeddings

the [ ]
cats [ ]
love [V]
hats [ ]

Predicate

Concatenate: 100dim + 100dim

B-ARG0 0.1
I-ARG0 0.7
B-V O 0.03
B-ARG1 0.1

Deep BiLSTM Tagger

Highway Connections

Variational Dropout

Viterbi Decoding w/ Hard Constraints

Word + Tag embeddings

Deep BiLSTM Tagger

Forward LSTM

Backward LSTM

Softmax

Output

B-ARG0

I-ARG0

B-V

B-ARG1

argmax

Predicate

Concatenate: 100dim + 100dim

B-ARG0 0.1
I-ARG0 0.7
B-V O 0.03
B-ARG1 0.1
Trend: Deeper models for higher accuracy

Grammar as a Foreign Language (Vinyals et al., 2014): 3 layers
End-to-end Semantic Role Labeling (Zhou and Xu, 2015): 8 layers
Google’s Neural Machine Translation (GNMT, Wu et al., 2016): 8 layers

this work: 8 layers

Deep Residual Learning for Image Recognition (He et al, 2016): 152 layers
increase expressive power

BiLSTM layers 1-2

the [ ]  cats [ ]  love [V]  hats [ ]
increase expressive power

BiLSTM layers 1-2

BiLSTM layers 3-4

21
The cats love hats.
The cats love hats.

increase expressive power
the cats love hats

BiLSTM layers 1-2

BiLSTM layers 3-4

BiLSTM layers 5-6

increase expressive power

harder to back-propagate
Grammar as a Foreign Language (Vinyals et al., 2014): 3 layers
End-to-end Semantic Role Labeling (Zhou and Xu, 2015): 8 layers
Google’s Neural Machine Translation (GNMT, Wu et al., 2016): 8 layers

this work: 8 layers

Deep Residual Learning for Image Recognition (He et al, 2016): 152 layers
Grammar as a Foreign Language (Vinyals et al., 2014): 3 layers
End-to-end Semantic Role Labeling (Zhou and Xu, 2015): 8 layers
Google’s Neural Machine Translation (GNMT, Wu et al., 2016): 8 layers
this work: 8 layers

Deep Residual Learning for Image Recognition (He et al, 2016): 152 layers

use different learning rates for different layers (harder to reimplement)
Grammar as a Foreign Language (Vinyals et al., 2014): 3 layers

End-to-end Semantic Role Labeling (Zhou and Xu, 2015): 8 layers

Google’s Neural Machine Translation (GNMT, Wu et al., 2016): 8 layers

this work: 8 layers

Deep Residual Learning for Image Recognition (He et al, 2016): 152 layers

use shortcut connections between layers (“highway” or “residual”)

use different learning rates for different layers (harder to reimplement)
Highway Connections

Non-linearity

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

output to the next layer

\[ h_t \]

recurrent input

\[ h_{t-1} \]

from the prev. timestep

input from the previous layer

\[ x_t \]

References:

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

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

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

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

\[ h_t \]  

\[ h_{t-1} \] from the prev. timestep

\[ x_t \]  

Non-linearity

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

output to the next layer

\[ h_t + x_t \]  

new output:

References:

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

\[
\begin{align*}
\text{recurrent input} & \quad \quad h_{t-1} \\
\text{from the prev. timestep} & \quad \quad \text{output to the next layer} \\
\text{input from the previous layer} & \quad \quad x_t \\
\text{Non-linearity} & \quad \quad F(c_{t-1}, h_{t-1}, x_t) \\
\text{shortcut} & \quad \quad x_t \\
\text{residual net} & \quad \quad h_t + x_t \\
\text{gated highway network:} & \quad \quad r_t \circ h_t + (1 - r_t) \circ x_t \\
& \quad \quad r_t = \sigma(f(h_{t-1}, x_t))
\end{align*}
\]

References:
- Deep Residual Networks, Kaiming He, ICML 2016 Tutorial
- Training Very Deep Networks, Srivastava et al., 2015
Variational Dropout

Model - (3) Variational Dropout
Deep BiLSTM Tagger
Highway Connections
Viterbi Decoding w/ Hard Constraints

the [ ]
cats [ ]
love [V]
Traditionally, dropout masks are only applied to vertical connections.
Traditionally, dropout masks are only applied to vertical connections.

Applying dropout to recurrent connections causes too much noise amplification.

**Variational Dropout**

Deep BiLSTM Tagger

Highway Connections

Viterbi Decoding w/ Hard Constraints

Model - (3) Variational Dropout

Deep BiLSTM Tagger

Highway Connections

Viterbi Decoding w/ Hard Constraints

```
the [ ]
cats [ ]
love [V]
```
Traditionally, dropout masks are only applied to vertical connections.

Applying dropout to recurrent connections causes too much noise amplification.

**Variational dropout:** Reuse the same dropout mask for each timestep. 
Gal and Ghahramani, 2016
Viterbi Decoding w/ Hard Constraints
Viterbi Decoding w/ Hard Constraints

BIO inconsistency

Greedy Output

B-ARG1

I-ARG0

B-V

B-ARG1

Softmax

argmax

BiLSTM layers …

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

\[ s(B-\text{ARG}0 \rightarrow I-\text{ARG}0) = 0 \]
\[ s(B-\text{ARG}1 \rightarrow I-\text{ARG}0) = -\infty \]

...
Heuristic transition scores

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

\[ \cdots \]

Softmax

BiLSTM layers ...

Viterbi decoding with Hard Constraints

Deep BiLSTM Tagger
Highway Connections Variational Dropout
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)
- 5 model ensemble with **product-of-experts** (Hinton 2002)
- Trained for 500 epochs.
## Datasets

<table>
<thead>
<tr>
<th></th>
<th>CoNLL-2005 (PropBank)</th>
<th>CoNLL-2012 (OntoNotes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>40k sentences</td>
<td>140k sentences</td>
</tr>
</tbody>
</table>
| Domains             | newswire              | • telephone conversations  
                     |                       | • newswire              
                     |                       | • newsgroups            
                     |                       | • broadcast news        
                     |                       | • broadcast conversation 
                     |                       | • weblogs                |
| Annotated           | Verbs                 | Added some nominal     
                     | predicates             | predicates              |
CoNLL 2005 Results

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

90
85
80
75
70
65
60

Ours*
Ours
Zhou
FitzGerald*
Täckström
Toutanova*
Punyakanok*
CoNLL 2005 Results

- **WSJ Test**:
  - Ours*: 84.6
  - Ours: 83.1
  - Zhou: 82.8
  - FitzGerald*: 80.3
  - Täckström: 79.9
  - Toutanova*: 80.3
  - Punyakanok*: 79.4

- **Brown (out-domain) Test**: 80.3

*: Ensemble models
CoNLL 2005 Results

<table>
<thead>
<tr>
<th></th>
<th>WSJ Test</th>
<th>Brown (out-domain) Test</th>
<th>Ensemble models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours*</td>
<td>84.6</td>
<td>73.6</td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td>83.1</td>
<td>72.1</td>
<td></td>
</tr>
<tr>
<td>Zhou</td>
<td>82.8</td>
<td>69.4</td>
<td></td>
</tr>
<tr>
<td>FitzGerald*</td>
<td>80.3</td>
<td>72.2</td>
<td></td>
</tr>
<tr>
<td>Täckström</td>
<td>79.9</td>
<td>71.3</td>
<td></td>
</tr>
<tr>
<td>Toutanova*</td>
<td>80.3</td>
<td>68.8</td>
<td></td>
</tr>
<tr>
<td>Punyakanok*</td>
<td>79.4</td>
<td>67.8</td>
<td></td>
</tr>
</tbody>
</table>
### CoNLL 2005 Results

<table>
<thead>
<tr>
<th>Method</th>
<th>WSJ Test</th>
<th>Brown (out-domain) Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours*</td>
<td>84.6</td>
<td>73.6</td>
</tr>
<tr>
<td>Ours</td>
<td>83.1</td>
<td>72.1</td>
</tr>
<tr>
<td>Zhou</td>
<td>82.8</td>
<td>69.4</td>
</tr>
<tr>
<td>FitzGerald*</td>
<td>80.3</td>
<td>72.2</td>
</tr>
<tr>
<td>Täckström</td>
<td>79.9</td>
<td>71.3</td>
</tr>
<tr>
<td>Toutanova*</td>
<td>80.3</td>
<td>68.8</td>
</tr>
<tr>
<td>Punyakanok*</td>
<td>79.4</td>
<td>67.8</td>
</tr>
</tbody>
</table>

- **wsj**: Pipeline models
- **BiLSTM**: BiLSTM models
- **Brown (out-domain)**: Pipeline models

*: Ensemble models
CoNLL 2012 (OntoNotes) Results

CoNLL 2012 Test

*: Ensemble models

F1

BiLSTM models

Pipeline models
Ablations on Number of Layers (2, 4, 6 and 8)

F1 on CoNLL-05 Dev.

- **Greedy decoding**
  - L2: 74.6
  - L4: 79.1
  - L6: 80.1
  - L8: 80.5

- **Viterbi decoding**
  - L2: 80.1
  - L4: 80.5
  - L6: 80.1
  - L8: 80.5
Ablations on Number of Layers (2, 4, 6 and 8)

<table>
<thead>
<tr>
<th>Layer</th>
<th>Greedy Decoding</th>
<th>Viterbi Decoding</th>
</tr>
</thead>
<tbody>
<tr>
<td>L2</td>
<td>74.6</td>
<td>77.2</td>
</tr>
<tr>
<td>L4</td>
<td>79.1</td>
<td>80.5</td>
</tr>
<tr>
<td>L6</td>
<td>80.1</td>
<td>81.4</td>
</tr>
<tr>
<td>L8</td>
<td>80.5</td>
<td>81.6</td>
</tr>
</tbody>
</table>
Ablations on Number of Layers (2, 4, 6 and 8)

Performance increases as model goes deeper. Biggest jump from 2 to 4 layer.
Ablations on Number of Layers (2, 4, 6 and 8)

Shallow models benefit more from constrained decoding.

Performance increases as model goes deeper. Biggest jump from 2 to 4 layer.

- Greedy decoding
- Viterbi decoding

F1 on CoNLL-05 Dev.
Ablations (single model, on CoNLL05 Dev)

- Full model
- No highway
- No orthonormal init.
- No dropout

F1 on Dev. Set vs. Num. Epochs
Ablations (single model, on CoNLL05 Dev)

F1 on Dev. Set

- Full model
- No highway
- No orthonormal init.
- No dropout

Graph shows the F1 score on the development set over the number of epochs for different models with various ablations.
Ablations (single model, on CoNLL05 Dev)

Without initialization, the deep model learns very slowly.
Ablations (single model, on CoNLL05 Dev)

With dropout, model overfits at ~300 epochs.

Without initialization, the deep model learns very slowly.

Full model  No highway  No orthonormal init.  No dropout
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. What are deeper models good at?
What can we learn from the results?

1. What’s in the remaining 17%? When does the model still struggle?
2. What are deeper models good at?
3. 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?
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. pp attachment)
<table>
<thead>
<tr>
<th>Error Breakdown</th>
<th>Labeling Errors</th>
<th>PP Attachment</th>
<th>Long-range Dependencies</th>
<th>Structural Consistency</th>
<th>Can Syntax Still Help?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oracle Transformations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
1) Fix 
Label: 

[We] *fly* to NYC tomorrow.
Error Breakdown

Oracle Transformations

1) Fix
Label: [We] fly to NYC tomorrow.

2) Move
core arg:
[They] wrote [an email] to cancel it.
Error Breakdown
Oracle Transformations

1) Fix Label: [We] fly to NYC tomorrow.

2) Move core arg: [They] wrote [an email] to cancel it.

3) Split/Merge span: I eat [pasta with delight].

I eat [pasta] [with broccoli].
We fly to NYC tomorrow.

They wrote an email to cancel it.

I eat pasta with delight.

With broccoli.

“No broccoli”, I said.
Error Breakdown

Oracle Transformations

1) Fix Label: [We] *fly* to NYC tomorrow.

2) Move core arg: [They] wrote [an email] to *cancel* it.

3) Split/Merge span: I *eat* [pasta with delight].

4) Fix span boundary: [“No broccoli”] I said.

5) Drop/add arg: Hesitantly, they *declined* to elaborate [on that matter].

[Oracle Transformations Diagram with labels and spans for error breakdown.]
Pradhan, Punyakanok: CoNLL-2005 systems
Pradhan, Punyakanok: CoNLL-2005 systems
Error Breakdown

- Labeling Errors
- PP Attachment
- Long-range Dependencies
- Structural Consistency
- Can Syntax Still Help?

Ours  Pradhan  Punyakanok

F1 after error fix

Original  Fix Labels  Move arg.  Merge/Split spans  Fix boundary  Drop/Add arg.

Pradhan, Punyakanok: CoNLL-2005 systems
Pradhan, Punyakanok: CoNLL-2005 systems
Error Breakdown

Oracle Transformations (Error Breakdown) Result

Pradhan, Punyakanok: CoNLL-2005 systems
Error Breakdown

- Labeling Errors
- PP Attachment
- Long-range Dependencies
- Structural Consistency
- Can Syntax Still Help?

Oracle Transformations (Error Breakdown) Result

Pradhan, Punyakanok: CoNLL-2005 systems
Error Breakdown

Ours | Pradhan | Punyakanok

Labeling Errors | PP Attachment | Long-range Dependencies | Structural Consistency | Can Syntax Still Help?

F1 after error fix

100
93.75
90
87.5
81.25
75

Original | Fix Labels | Move arg. | Merge/Split spans | Fix boundary | Drop/Add arg.

Pradhan, Punyakanok: CoNLL-2005 systems

Labeling error 29.3%
Pradhan, Punyakanok: CoNLL-2005 systems
**Labeling Errors**

Confusion matrix for labeling errors (row 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>A0</strong></td>
<td>76</td>
<td>13</td>
<td>6</td>
<td>14</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>A1</strong></td>
<td>16</td>
<td>74</td>
<td>25</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td>9</td>
<td>11</td>
<td>19</td>
<td>2</td>
</tr>
<tr>
<td><strong>A2</strong></td>
<td>2</td>
<td>5</td>
<td>31</td>
<td>52</td>
<td>10</td>
<td>45</td>
<td>26</td>
<td>46</td>
<td>19</td>
<td>0</td>
</tr>
<tr>
<td><strong>A3</strong></td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>57</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>19</td>
</tr>
<tr>
<td><strong>ADV</strong></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>33</td>
<td>0</td>
<td>11</td>
<td>33</td>
<td>19</td>
<td>5</td>
</tr>
<tr>
<td><strong>DIR</strong></td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>5</td>
<td>0</td>
<td>27</td>
<td>9</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>LOC</strong></td>
<td>1</td>
<td>2</td>
<td>7</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>34</td>
<td>11</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td><strong>MNR</strong></td>
<td>1</td>
<td>0</td>
<td>7</td>
<td>29</td>
<td>21</td>
<td>0</td>
<td>0</td>
<td>43</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td><strong>PNC</strong></td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>0</td>
<td>9</td>
<td>3</td>
<td>2</td>
<td>44</td>
<td>0</td>
</tr>
<tr>
<td><strong>TMP</strong></td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>26</td>
<td>9</td>
<td>20</td>
<td>7</td>
<td>0</td>
<td>71</td>
</tr>
</tbody>
</table>
### Confusion matrix for labeling errors (row 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>A0</td>
<td>76</td>
<td>13</td>
<td>6</td>
<td>14</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A1</td>
<td>16</td>
<td>74</td>
<td>25</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td>9</td>
<td>11</td>
<td>19</td>
<td>2</td>
</tr>
<tr>
<td>A2</td>
<td>2</td>
<td>5</td>
<td>31</td>
<td>52</td>
<td>10</td>
<td>45</td>
<td>26</td>
<td>46</td>
<td>19</td>
<td>0</td>
</tr>
<tr>
<td>A3</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>57</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>19</td>
<td>2</td>
</tr>
<tr>
<td>ADV</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>33</td>
<td>0</td>
<td>11</td>
<td>33</td>
<td>19</td>
<td>5</td>
</tr>
<tr>
<td>DIR</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>5</td>
<td>0</td>
<td>27</td>
<td>9</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LOC</td>
<td>1</td>
<td>2</td>
<td>7</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>34</td>
<td>11</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>MNR</td>
<td>1</td>
<td>0</td>
<td>7</td>
<td>29</td>
<td>21</td>
<td>0</td>
<td>0</td>
<td>43</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>PNC</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>0</td>
<td>9</td>
<td>3</td>
<td>2</td>
<td>44</td>
<td>0</td>
</tr>
<tr>
<td>TMP</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>26</td>
<td>9</td>
<td>20</td>
<td>7</td>
<td>0</td>
<td>71</td>
</tr>
</tbody>
</table>

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

Confusion matrix for labeling errors (row 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
Confusion matrix for labeling errors (row 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
Confusion matrix for labeling errors (row 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 (row 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 human annotators!
**Labeling Errors**

Confusion matrix for labeling errors (row normalized)

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

> “After many attempts to find a reliable test to distinguish between arguments and adjuncts, we abandoned structurally marking this difference.”

—The Penn Treebank: An Overview (Taylor et al., 2003)

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

Sumimoto *financed* the acquisition from Sears

Correct PP attachment
(attach low)
Wrong PP attachment (attach high)

Sumimoto *financed* the acquisition from Sears

Correct PP attachment (attach low)

Wrong SRL spans

merge

Correct SRL spans
Wrong PP attachment (attach high)

Sumimoto *financed* the acquisition from Sears

Correct PP attachment (attach low)

Merge/split span operations: 25.3% of the mode mistakes.

Categorize the Y spans in:

- 
  
[XY]→[X][Y] and
  
[X][Y]→[XY] operations

using gold syntactic labels
Sumimoto financed the acquisition from Sears

Wrong PP attachment (attach high)

Correct PP attachment (attach low)

Merge/split span operations: 25.3% of the mode mistakes.

Categorize the Y spans in:

\[[XY] \rightarrow [X][Y]\] and
\[[X][Y] \rightarrow [XY]\] operations using gold syntactic labels

Wrong SRL spans

merge

Correct SRL spans
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. pp attachment)
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. pp attachment)

**Takeaway**
- Traditionally hard tasks, such as argument-adjunct distinction and PP attachment decisions are still challenging!
- Use external information to improve PP attachment.
Question (2): What are deeper models good at?

Analysis

— Long-range dependencies: model performance on arguments that are far away from the predicates.
— Structural consistency: amount of inconsistent BIO tag pairs in greedy prediction.
Long-range Dependencies

Num. gold arguments

Distance to the predicate (by num. words)

Avg. F1

L8
L6
L4
L2
Punyakanok
Pradhan

Can Syntax Still Help?
Long-range Dependencies

Distance to the predicate (by num. words):

- 0: 4,469
- 1-2: 2,151
- 3-7: 1,174
- 7-max: 657

Num. gold arguments:
- 0: 4,469
- 1-2: 2,151
- 3-7: 1,174
- 7-max: 657

Avg. F1:
- 0: 88
- 1-2: 79.5
- 3-7: 71
- 7-max: 62.5

Error Breakdown:
- Labeling Errors
- PP Attachment

Structural Consistency:
- Can Syntax Still Help?
Long-range Dependencies

Error Breakdown | Labeling Errors | PP Attachment
---|---|---

| Num. gold arguments |
|---|---|---|---|---|
| 0 | 1-2 | 3-7 | 7-max |
| 4,469 | 2,151 | 1,174 | 657 |

Distance to the predicate (by num. words)

Avg. F1

0 | 1-2 | 3-7 | 7-max
---|---|---|---
| 54 | 88 | 79.5 | 71 |

Num. gold arguments

Error Breakdown

Structural Consistency

Can Syntax Still Help?

L8 | L6 | L4 | L2 | Punyakanok | Pradhan
---|---|---|---|---|---

Distance to the predicate (by num. words)
Deep models’ performance deteriorates slower on long-distance predictions.
e.g. (B-ArgX, I-ArgY) or (O, I-ArgY)
Deeper models (with 4+ layers) generate more consistent BIO sequences.

e.g. (B-ArgX, I-ArgY) or (O, I-ArgY)
Question (3): 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.
— See if SRL performance increases.
Can Syntax Still Help?

BiLSTM-based models

Syntax-aware models:

- Gold

Error Breakdown
Labeling Errors
PP Attachment
Long-range Dependencies
Structural Consistency

Chart showing F1 scores for percentage of arguments in gold syntax tree.
Can Syntax Still Help?

BiLSTM-based models

Syntax-aware models:

Improve SRL accuracy by adding syntactic information?

% Arguments in gold syntax tree

- L2
- L4
- L6
- L8
- L8+PoE
- Gold

Error Breakdown
Labeling Errors
PP Attachment
Long-range Dependencies
Structural Consistency
[The cats] ∈ Syntax Tree
[hats and the dogs] ∉ Syntax Tree

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

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

Penalize sequence score
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

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

Penalty strength

Num. arguments disagree w/ syntax
• Constraints are not locally decomposable.
• A* search (Lewis and Steedman 2014) for a sequence with highest score.

Sequence score:
\[
\sum_{i=1}^{t} \log p(\text{tag}_t \mid \text{sentence}) - C \times \sum_{\text{span}} 1(\text{span} \notin \text{Syntax Tree})
\]
Can Syntax Still Help?
Constrained Decoding with Syntax

<table>
<thead>
<tr>
<th>Error Breakdown</th>
<th>Labeling Errors</th>
<th>PP Attachment</th>
<th>Long-range Dependencies</th>
<th>Structural Consistency</th>
</tr>
</thead>
</table>

**Charniak:** A maximum-entropy-inspired parser, Charniak, 2000

**Choe:** Parsing as language modeling, Choe and Charniak, 2016
(State of the art)
**Charniak:** A maximum-entropy-inspired parser, Charniak, 2000

**Choe:** Parsing as language modeling, Choe and Charniak, 2016

(State of the art)
Can Syntax Still Help?

- **BiLSTM-based models**
  - Syntax-aware models

Constrained decoding \( w\backslash \text{Syntax} \)

Gold

% Arguments agree with gold syntax
Can Syntax Still Help?

**Takeaway**

- Modest gain observed with predicted syntax.
- Joint training could bring more improvement.
Contributions (Neural SRL)

- New state-of-the-art deep network for end-to-end SRL.
- Code and models will be 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.
Long-term Plan for Improving SRL

Step 1: Collect more data for SRL
- Question-Answer Driven Semantic Role Labeling (QA-SRL)
- Human-in-the-Loop Parsing

Step 2: Build accurate SRL model
- Neural Semantic Role Labeling (for PropBank SRL)

Step 3: SRL system for many domains
- Future work …
Long-term Plan for Improving SRL

Step 1: Collect more data for SRL
  — Question-Answer Driven Semantic Role Labeling (QA-SRL)
  — Human-in-the-Loop Parsing

Step 2: Build accurate SRL model
  — Neural Semantic Role Labeling (for PropBank SRL)

Step 3: SRL system for many domains
  — Future work …
Thanks!

Code will be available at: https://github.com/luheng/deep_srl

Problem

My mug broke into pieces immediately.

Frame: break.01

role | description
--- | ---
ARG0 | breaker
ARG1 | thing broken
ARG2 | instrument
ARG3 | pieces
ARG4 | broken away from what?

Model

sentence, predicate

Deep BiLSTM

BIO sequence

Viterbi

prediction

Analysis

![Analysis graph]