Deep Learning for Broad Coverage Semantics: SRL, Coreference, and Beyond

Luke Zettlemoyer†* 

Joint work with Luheng He†, Kenton Lee†, Matthew Peters*, Christopher Clark†, Matthew Gardner*, Mohit Iyyer*, Mandar Joshi†, Mike Lewis‡, Julian Michael†, Mark Neumann* 

† Paul G. Allen School of Computer Science & Engineering, University of Washington, 
‡ Facebook AI Research 
* Allen Institute for Artificial Intelligence
Three Simple Steps that will Revolutionize Your ML Research

Step 1:

Step 2:

Step 3:
Three Simple Steps that will Revolutionize Your ML Research

Step 1: Gather lots of training data!

Step 2:

Step 3:
Three Simple Steps that will Revolutionize Your ML Research

Step 1: Gather lots of training data!

Step 2:

Step 3:
Three Simple Steps that will Revolutionize Your ML Research

Step 1: Gather lots of training data!

Step 2: Apply Deep Learning!!

Step 3:
Three Simple Steps that will Revolutionize Your ML Research

Step 1: Gather lots of training data!

 IMAGENET   European Parliament   twitter

Step 2: Apply Deep Learning!!

Step 3:
Three Simple Steps that will Revolutionize Your ML Research

Step 1: Gather lots of training data!

Step 2: Apply Deep Learning!!

Step 3: Observe Impressive Gains!!
A fire in a Bangladeshi garment factory has left at least 37 people dead and 100 hospitalized. Most of the deceased were killed in the crush as workers tried to flee the blaze in the four-story building.

NASA observed an X-ray flare 400 times brighter than usual on January 5, 2015.

Many applications:

- Question Answering
- Information Extraction
- Machine Translation

Example Tasks:

- Coreference: clustering NPs
- Semantic Role Labeling: who did what, etc.
Does the Recipe Work for Broad Coverage Semantics?

Step 1: Gather lots of training data!

Step 2: Apply Deep Learning!!

Step 3: Observe Impressive Gains!!!
Does the Recipe Work for Broad Coverage Semantics?

Step 1: Gather lots of training data!

Challenge 1: Data is costly and limited (e.g. linguists required to label PennTreebank / OntoNotes)

Step 2: Apply Deep Learning!!

Step 3: Observe Impressive Gains!!!
Does the Recipe Work for Broad Coverage Semantics?

Step 1: Gather lots of training data!

**Challenge 1:** Data is costly and limited (e.g. linguists required to label PennTreebank / OntoNotes)

Step 2: Apply Deep Learning!!

**Challenge 2:** Pipeline of structured prediction problems with cascading errors (e.g. POS->Parsing->SRL->Coref)

Step 3: Observe Impressive Gains!!!
New Learning Approaches

**New state-of-the-art results for two tasks:**

### Coreference:

A fire in a Bangladeshi garment factory has left at least 37 people dead and 100 hospitalized. Most of the deceased were killed in the crush as workers tried to flee the blaze in the four-story building.

### Semantic Role Labeling:

<table>
<thead>
<tr>
<th>ARG0</th>
<th>NASA</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRED</td>
<td>observe</td>
</tr>
<tr>
<td>ARG1</td>
<td>an X-ray flare 400 times brighter than usual</td>
</tr>
<tr>
<td>TMP</td>
<td>On January 5, 2015</td>
</tr>
</tbody>
</table>

**Common themes:**

- End-to-end training of deep neural networks
- No preprocessing (e.g., no POS, no parser, etc.)
- Large gains in accuracy with simpler models and no extra training data
My mug broke into pieces immediately.

Semantic Role Labeling (SRL)

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

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

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

My mug *broke* into pieces immediately.
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.

Semantic Role Labeling (SRL)

- **Predicate**: broke
- **Arguments**:
  - **Subject**: The robot
  - **Object**: my favorite mug
  - **Instrument**: with a wrench
  - **Final State**: into pieces immediately
My mug broke into pieces immediately.

The robot broke my favorite 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</td>
</tr>
<tr>
<td>ARG4</td>
<td>broken away from what?</td>
</tr>
</tbody>
</table>
My mug broke into pieces immediately.

The robot *broke* my favorite 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>
<tr>
<td>ARG4</td>
<td>broken away from what?</td>
</tr>
</tbody>
</table>
SRL is a hard problem …

• Over 10 years, F1 on PropBank: 80.3 (Toutanova et al, 2005) — 80.3 (FitzGerald et al, 2015)

• Many interesting challenges:
  Syntactic alternation
  Prepositional phrase attachment
  Long-range dependencies and common sense
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
- 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
SRL Systems

Pipeline Systems
- sentence, predicate
- syntactic features
- 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

He et al., 2017
SRL as BIO Tagging Problem

Input (sentence and predicate):

The cats love hats.
SRL as BIO Tagging Problem

Input (sentence and predicate):

BIO output: (Begin, Inside, Outside)
Input (sentence and predicate): The cats love hats.

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

(Begin, Inside, Outside)

Final SRL output: ARG0 V ARG1
the [ ]
cats [ ]
love [V]
hats [ ]

[He et al, 2017]
The cats love hats.

(1) Deep BiLSTM tagger

[He et al, 2017]
the cats love hats

(1) Deep BiLSTM tagger

(2) Highway connections

[He et al, 2017]
the cats love hats

(1) Deep BiLSTM tagger

(2) Highway connections

(3) Variational dropout

[He et al, 2017]
(1) Deep BiLSTM tagger
(2) Highway connections
(3) Variational dropout
(4) Viterbi decoding with hard constraints

[He et al, 2017]
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.
### CoNLL 2005 Results

<table>
<thead>
<tr>
<th>Datasets</th>
<th>CoNLL 2012 (OntoNotes) Results</th>
<th>Ablations</th>
</tr>
</thead>
<tbody>
<tr>
<td>WSJ Test</td>
<td>90</td>
<td>*: Ensemble models</td>
</tr>
<tr>
<td>Brown (out-domain) Test</td>
<td>85</td>
<td></td>
</tr>
<tr>
<td>80</td>
<td></td>
<td></td>
</tr>
<tr>
<td>75</td>
<td></td>
<td></td>
</tr>
<tr>
<td>70</td>
<td></td>
<td></td>
</tr>
<tr>
<td>65</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ours*</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>2017</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2017</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zhou</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2015</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FitzGerald*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2015</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Täckström</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2015</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Toutanova*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2008</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Punyakanok*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2008</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
CoNLL 2005 Results

Datasets

- Ours*
  - 2017

- Ours
  - 2017

- Zhou
  - 2015

- FitzGerald*
  - 2015

- Täckström
  - 2015

- Toutanova*
  - 2008

- Punyakanok*
  - 2008

CoNLL 2012 (OntoNotes) Results

- WSJ Test: 84.6
- Brown (out-domain) Test: 83.1
- Zhou: 82.8
- FitzGerald*: 80.3
- Täckström: 79.9
- Toutanova*: 80.3
- Punyakanok*: 79.4

*: Ensemble models

Ablations
CoNLL 2005 Results

Datasets

WSJ Test
Brown (out-domain) Test

Ours: Ensemble models

2017

2008

Ours* 2017
Zhou 2015
FitzGerald* 2015
Täckström 2015
Toutanova* 2008
Punyakanok* 2008

CoNLL 2012 (OntoNotes) Results

Ablations

67.8
68.8
71.3
72.2
69.4
72.1
73.6
82.8
83.1
82.8
80.3
79.9
80.3
68.8
79.4
79.9
80.3
82.8
83.1
84.6

85
90
CoNLL 2005 Results

Datasets

<table>
<thead>
<tr>
<th>Model</th>
<th>Year</th>
<th>Test Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>2017</td>
<td>WSJ Test</td>
</tr>
<tr>
<td>Ours</td>
<td>2017</td>
<td>Brown (out-domain) Test</td>
</tr>
<tr>
<td>Zhou</td>
<td>2015</td>
<td>WSJ Test</td>
</tr>
<tr>
<td>FitzGerald*</td>
<td>2015</td>
<td>Brown (out-domain) Test</td>
</tr>
<tr>
<td>Täckström</td>
<td>2015</td>
<td>WSJ Test</td>
</tr>
<tr>
<td>Toutanova*</td>
<td>2008</td>
<td>Brown (out-domain) Test</td>
</tr>
<tr>
<td>Punyakanok*</td>
<td>2008</td>
<td>Brown (out-domain) Test</td>
</tr>
</tbody>
</table>

CoNLL 2012 (OntoNotes) Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Year</th>
<th>Test Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours*</td>
<td>2017</td>
<td>WSJ Test</td>
</tr>
<tr>
<td>Ours</td>
<td>2017</td>
<td>Brown (out-domain) Test</td>
</tr>
<tr>
<td>Zhou</td>
<td>2015</td>
<td>WSJ Test</td>
</tr>
<tr>
<td>FitzGerald*</td>
<td>2015</td>
<td>Brown (out-domain) Test</td>
</tr>
<tr>
<td>Täckström</td>
<td>2015</td>
<td>WSJ Test</td>
</tr>
<tr>
<td>Toutanova*</td>
<td>2008</td>
<td>Brown (out-domain) Test</td>
</tr>
<tr>
<td>Punyakanok*</td>
<td>2008</td>
<td>Brown (out-domain) Test</td>
</tr>
</tbody>
</table>

Ablations

BiLSTM models

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

F1 on CoNLL-05 Dev.

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

Greedy decoding
Viterbi decoding

F1 on CoNLL-05 Dev.
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.
New Learning Approaches

New state-of-the-art results for two tasks:

Coreference:

A fire in a Bangladeshi garment factory has left at least 37 people dead and 100 hospitalized. Most of the deceased were killed in the crush as workers tried to flee the blaze in the four-story building.

Semantic Role Labeling:

<table>
<thead>
<tr>
<th>ARG0</th>
<th>NASA</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRED</td>
<td>observe</td>
</tr>
<tr>
<td>ARG1</td>
<td>an X-ray flare 400 times brighter than usual</td>
</tr>
<tr>
<td>TMP</td>
<td>On January 5, 2015</td>
</tr>
</tbody>
</table>

Common themes:

• End-to-end training of deep neural networks
• No preprocessing (e.g., no POS, no parser, etc.)
• Large gains in accuracy with simpler models and no extra training data
## Coreference Resolution

<table>
<thead>
<tr>
<th><strong>Input document</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>A fire in a Bangladeshi garment factory has left at least 37 people dead and 100 hospitalized. Most of the deceased were killed in the crush as workers tried to flee the blaze in the four-story building.</td>
</tr>
</tbody>
</table>
Coreference Resolution

<table>
<thead>
<tr>
<th>Input document</th>
</tr>
</thead>
<tbody>
<tr>
<td>A fire in a Bangladeshi garment factory has left at least 37 people dead and 100 hospitalized. Most of the deceased were killed in the crush as workers tried to flee the blaze in the four-story building.</td>
</tr>
</tbody>
</table>

| Cluster #1 | A fire in a Bangladeshi garment factory | the blaze in the four-story building |
Coreference Resolution

A fire in a Bangladeshi garment factory has left at least 37 people dead and 100 hospitalized. Most of the deceased were killed in the crush as workers tried to flee the blaze in the four-story building.

<table>
<thead>
<tr>
<th>Cluster #1</th>
<th>A fire in a Bangladeshi garment factory</th>
<th>the blaze in the four-story building</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster #2</td>
<td>a Bangladeshi garment factory</td>
<td>the four-story building</td>
</tr>
</tbody>
</table>
A fire in a Bangladeshi garment factory has left at least 37 people dead and 100 hospitalized. Most of the deceased were killed in the crush as workers tried to flee the blaze in the four-story building.

<table>
<thead>
<tr>
<th>Cluster #1</th>
<th>A fire in a Bangladeshi garment factory</th>
<th>the blaze in the four-story building</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster #2</td>
<td>a Bangladeshi garment factory</td>
<td>the four-story building</td>
</tr>
<tr>
<td>Cluster #3</td>
<td>at least 37 people</td>
<td>the deceased</td>
</tr>
</tbody>
</table>
Two Subproblems

**Input document**

A fire in a Bangladeshi garment factory has left at least 37 people dead and 100 hospitalized. Most of the deceased were killed in the crush as workers tried to flee the blaze in the four-story building.

**Mention detection**

<table>
<thead>
<tr>
<th>A fire in a Bangladeshi garment factory</th>
</tr>
</thead>
<tbody>
<tr>
<td>at least 37 people</td>
</tr>
<tr>
<td>...</td>
</tr>
<tr>
<td>the four-story building</td>
</tr>
</tbody>
</table>

**Mention clustering**

<table>
<thead>
<tr>
<th>Cluster #1</th>
<th>A fire in a Bangladeshi garment factory</th>
<th>the blaze in the four-story building</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster #2</td>
<td>a Bangladeshi garment factory</td>
<td>the four-story building</td>
</tr>
<tr>
<td>Cluster #3</td>
<td>at least 37 people</td>
<td>the deceased</td>
</tr>
</tbody>
</table>
Previous Approach: Rule-based pipeline

**Input document**
A fire in a Bangladeshi garment factory has left at least 37 people dead and 100 hospitalized.

**Syntactic parser**

**Hand-engineered rules**

**Candidate mentions**

<table>
<thead>
<tr>
<th>A fire in a Bangladeshi garment factory</th>
</tr>
</thead>
<tbody>
<tr>
<td>garment</td>
</tr>
<tr>
<td>factory</td>
</tr>
<tr>
<td>at least 37 people dead and 100 hospitalized</td>
</tr>
</tbody>
</table>

**Mention #1**

<table>
<thead>
<tr>
<th>A fire in a Bangladeshi garment factory</th>
</tr>
</thead>
<tbody>
<tr>
<td>garment</td>
</tr>
<tr>
<td>factory</td>
</tr>
<tr>
<td>at least 37 people dead and 100 hospitalized</td>
</tr>
</tbody>
</table>

**Mention #2**

<table>
<thead>
<tr>
<th>garment</th>
</tr>
</thead>
<tbody>
<tr>
<td>factory</td>
</tr>
<tr>
<td>at least 37 people dead and 100 hospitalized</td>
</tr>
</tbody>
</table>

**Coreferent?**

✓/✗
Previous Approach: Rule-based pipeline

A fire in a Bangladeshi garment factory has left at least 37 people dead and 100 hospitalized.

Mention clustering: main source of improvement for many years!

- Haghighi and Klein (2010)
- Raghunathan et al. (2010)
- ... 
- Clark & Manning (2016)
Previous Approach: Rule-based pipeline

Relies on parser for:
- mention detection
- syntactic features for clustering (e.g. head words)
End-to-end Approach

- Consider all possible spans
- Learn to rank antecedent spans
- Factored model to prune search space
Key Idea: Span Representations

Bidirectional LSTM

Word & character embeddings

General Electric said the Postal Service contacted the company
Key Idea: Span Representations

General Electric said the Postal Service contacted the company.
Key Idea: Span Representations

Boundary representations

Span representation

Bidirectional LSTM

Word & character embeddings

General, Electric, said, the, Postal, Service, contacted, the, company
Key Idea: Span Representations

Attention mechanism to learn headedness

Span representation
Head-finding attention
Bidirectional LSTM
Word & character embeddings

General Electric said the Postal Service contacted the company
Key Idea: Span Representations

Compute all span representations

Span representation
Head-finding attention
Bidirectional LSTM
Word & character embeddings

General Electric
Electric said the
the Postal Service
Service contacted the
the company

General
Electric
said
the
Postal
Service
contacted
the
company
Mention Ranking

Every span independently chooses an antecedent

<table>
<thead>
<tr>
<th>Input document</th>
</tr>
</thead>
<tbody>
<tr>
<td>A fire in a Bangladeshi garment factory has left at least 37 people dead and 100 hospitalized. Most of the deceased were killed in the crush as workers tried to flee the blaze in the four-story building. Witnesses say the only exit door was on the ground floor, and that it was locked when the fire broke out.</td>
</tr>
</tbody>
</table>
Mention Ranking

- Reason over all possible spans

- Assign an antecedent to every span

\[ y_3 \in \{ \epsilon, 1, 2 \} \]
A fire in a Bangladeshi garment factory has left at least 37 people dead and 100 hospitalized. Most of the deceased were killed in the crush as workers tried to flee the blaze in the four-story building. Witnesses say the only exit door was on the ground floor, and that it was locked when the fire broke out.
A fire in a Bangladeshi garment factory has left at least 37 people dead and 100 hospitalized. Most of the deceased were killed in the crush as workers tried to flee the blaze in the four-story building. Witnesses said the only exit door was on the ground floor, and that it was locked when the fire broke out.
A fire in a Bangladeshi garment factory has left at least 37 people dead and 100 hospitalized. Most of the deceased were killed in the crush as workers tried to flee the blaze in the four-story building. Witnesses say the only exit door was on the ground floor, and that it was locked when the fire broke out.

<table>
<thead>
<tr>
<th>Span</th>
<th>Antecedent ($y_i$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>a Bangladeshi garment factory</td>
<td>$\epsilon$</td>
</tr>
<tr>
<td>the four-story building</td>
<td>a Bangladeshi garment factory</td>
</tr>
<tr>
<td>out</td>
<td>$\epsilon$</td>
</tr>
</tbody>
</table>

No link with previously occurring span
A fire in a Bangladeshi garment factory has left at least 37 people dead and 100 hospitalized. Most of the deceased were killed in the crush as workers tried to flee the blaze in the four-story building. Witnesses say the only exit door was on the ground floor, and that it was locked when the fire broke out.

<table>
<thead>
<tr>
<th>Span</th>
<th>Antecedent ($y_i$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>€</td>
</tr>
<tr>
<td>A fire</td>
<td>€</td>
</tr>
<tr>
<td>...</td>
<td>€</td>
</tr>
<tr>
<td>the four-story building</td>
<td>a Bangladeshi garment factory</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>out</td>
<td>€</td>
</tr>
</tbody>
</table>

Predicted coreference link
Span Ranking Model

\[
P(y_1, \ldots, y_M \mid D) = \prod_{i=1}^{M} P(y_i \mid D)
\]

\[
= \prod_{i=1}^{M} \frac{e^{s(i, y_i)}}{\sum_{y' \in \mathcal{Y}(i)} e^{s(i, y')}}
\]

Factor coreference score \( s(i, j) \) to enable span pruning:

\[
s(i, j) = \begin{cases} 
  s_m(i) + s_m(j) + s_a(i, j) & j \neq \epsilon \\
  0 & j = \epsilon 
\end{cases}
\]
Span Ranking Model

\[ P(y_1, \ldots, y_M \mid D) = \prod_{i=1}^{M} P(y_i \mid D) \]

\[ = \frac{\prod_{i=1}^{M} e^{s(i, y_i)}}{\sum_{\forall y' \in \mathcal{Y}(i)} e^{s(i, y')}} \]

Is this span a mention?

Factor coreference score \( s(i, j) \) to enable span pruning:

\[ s(i, j) = \begin{cases} 
    s_m(i) + s_m(j) + s_a(i, j) & j \neq \epsilon \\
    0 & j = \epsilon 
\end{cases} \]
Span Ranking Model

\[ P(y_1, \ldots, y_M \mid D) = \prod_{i=1}^{M} P(y_i \mid D) = \prod_{i=1}^{M} \frac{e^{s(i,y_i)}}{\sum_{i} e^{s(i,y_i)}} \]

Is span \( j \) an antecedent of span \( i \)?

Factor coreference score \( s(i,j) \) to enable span pruning:

\[ s(i,j) = \begin{cases} s_m(i) + s_m(j) + s_a(i,j) & j \neq \epsilon \\ 0 & j = \epsilon \end{cases} \]
Span Ranking Model

\[ P(y_1, \ldots, y_M \mid D) = \prod_{i=1}^{M} P(y_i \mid D) \]

\[ = \prod_{i=1}^{M} \frac{e^{s(i,y_i)}}{\sum_{y' \in \mathcal{Y}(i)} e^{s(i,y')}} \]

Factor coreference score \( s(i, j) \) to enable span pruning:

\[ s(i, j) = \begin{cases} 
    s_m(i) + s_m(j) + s_a(i, j) & j \neq \epsilon \\
    0 & j = \epsilon 
\end{cases} \]

Dummy antecedent has a fixed zero score
Experimental Setup

**Dataset:** English OntoNotes (CoNLL-2012)

**Genres:** Telephone conversations, newswire, newsgroups, broadcast conversation, broadcast news, weblogs

**Documents:** 2802 training, 343 development, 348 test

**Aggressive pruning:** Maximum span width, maximum sentence training, suppress spans with inconsistent bracketing, maximum number of antecedents

**Features:** distance between spans, span width

**Metadata:** speaker information, genre

Longest document has 4009 words!
Coreference Results

Test Avg. F1 (%)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>60.3</td>
<td>61.6</td>
<td>62.5</td>
</tr>
</tbody>
</table>

Linear models
Coreference Results

<table>
<thead>
<tr>
<th>Study</th>
<th>Test Avg. F1 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Durrett &amp; Klein (2013)</td>
<td>60.3</td>
</tr>
<tr>
<td>Björkelund &amp; Kuhn (2014)</td>
<td>61.6</td>
</tr>
<tr>
<td>Martschat &amp; Strube (2015)</td>
<td>62.5</td>
</tr>
<tr>
<td>Wiseman et al. (2016)</td>
<td>64.2</td>
</tr>
<tr>
<td>Clark &amp; Manning (2016)</td>
<td>65.7</td>
</tr>
</tbody>
</table>

Linear models

Neural models
Coreference Results

<table>
<thead>
<tr>
<th>Test Avg. F1 (%)</th>
<th>Authors and Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>60.3</td>
<td>Durrett &amp; Klein (2013)</td>
</tr>
<tr>
<td>61.6</td>
<td>Björkelund &amp; Kuhn (2014)</td>
</tr>
<tr>
<td>62.5</td>
<td>Martschat &amp; Strube (2015)</td>
</tr>
<tr>
<td>64.2</td>
<td>Wiseman et al. (2016)</td>
</tr>
<tr>
<td>65.7</td>
<td>Clark &amp; Manning (2016)</td>
</tr>
</tbody>
</table>

Pipelined models
Coreference Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Test Avg. F1 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Durrett &amp; Klein (2013)</td>
<td>60.3</td>
</tr>
<tr>
<td>Björkelund &amp; Kuhn (2014)</td>
<td>61.6</td>
</tr>
<tr>
<td>Martschat &amp; Strube (2015)</td>
<td>62.5</td>
</tr>
<tr>
<td>Wiseman et al. (2016)</td>
<td>64.2</td>
</tr>
<tr>
<td>Clark &amp; Manning (2016)</td>
<td>65.7</td>
</tr>
<tr>
<td>Our model (single)</td>
<td>67.2</td>
</tr>
<tr>
<td>Our model (ensemble)</td>
<td>68.8</td>
</tr>
</tbody>
</table>

End-to-end models

Pipelined models
Qualitative Analysis

A fire in a Bangladeshi garment factory has left at least 37 people dead and 100 hospitalized. Most of the deceased were killed in the crush as workers tried to flee the blaze in the four-story building.
A fire in a Bangladeshi garment factory has left at least 37 people dead and 100 hospitalized. Most of the deceased were killed in the crush as workers tried to flee the blaze in the four-story building.
A fire in a Bangladeshi garment factory has left at least 37 people dead and 100 hospitalized. Most of the deceased were killed in the crush as workers tried to flee the blaze in the four-story building.

Good head-finding requires word-order information!
The flight attendants have until 6:00 today to ratify labor concessions. The pilots' union and ground crew did so yesterday.
The flight attendants have until 6:00 today to ratify labor concessions. The pilots’ union and ground crew did so yesterday.
Does the Recipe Work for Broad Coverage Semantics?

Step 1: Gather lots of training data!

Challenge 1: Data is costly and limited (e.g. linguists required to label PennTreebank / OntoNotes)

Step 2: Apply Deep Learning!!

Challenge 2: Pipeline of structured prediction problems with cascading errors (e.g. POS->Parsing->SRL->Coref)

Step 3: Observe Impressive Gains!!!
Where Will the Data Come From???
Where Will the Data Come From???

Option 1: Semi-supervised learning

- E.g. word2vec and GloVe are in wide use
  [Mikolov et al., 2013; Pennington et al., 2014]
Where Will the Data Come From???

Option 1: Semi-supervised learning

• E.g. word2vec and GloVe are in wide use
  [Mikolov et al., 2013; Pennington et al., 2014]

• Can we learn better word representations?
Where Will the Data Come From???

Option 1: Semi-supervised learning

• E.g. word2vec and GloVe are in wide use
  [Mikolov et al., 2013; Pennington et al., 2014]
• Can we learn better word representations?

Option 2: Supervised learning
Where Will the Data Come From??

**Option 1:** Semi-supervised learning
- E.g. word2vec and GloVe are in wide use [Mikolov et al., 2013; Pennington et al., 2014]
- Can we learn better word representations?

**Option 2:** Supervised learning
- Can we gather more direct forms of supervision?
Learning Better Word Representations

**Goal:** Model contextualized syntax and semantics

\[ R(w_i, w_1 \ldots w_n) \in \mathbb{R}^n \]

\[ R(\text{plays, "The robot plays piano."}) \neq R(\text{plays, "The robot starred in many plays."}) \]
Word Embeddings from a Language Model

**Step 1:** Train a large BiLM on unlabeled data
Word Embeddings from a Language Model

**Step 1:** Train a large BiLM on unlabeled data

- **Left and Right Per Word Softmaxs**
- **2 Layer Bidirectional LSTM**
- **Character convolutions**

---

General Electric said the Postal Service contacted the company
Word Embeddings from a Language Model

**Step 1:** Train a large BiLM on unlabeled data

**Step 2:** Compute linear function of pre-trained model
Word Embeddings from a Language Model

**Step 1:** Train a large BiLM on unlabeled data

**Step 2:** Compute linear function of pre-trained model

\[ \text{LM Embeddings} = \alpha_1 + \alpha_2 + \alpha_3 \]
Word Embeddings from a Language Model

**Step 1:** Train a large BiLM on unlabeled data

**Step 2:** Compute linear function of pre-trained model

**Step 3:** Learn weights for each end task

\[ \text{LM Embeddings} = \alpha_1 \circ \text{ } + \alpha_2 \circ \text{ } + \alpha_3 \circ \text{ } \]
Best Single System Results

**Test Avg. F1 (%)**

- Feature Based: 79.9, 62.5
- Neural: 81.7, 67.2
- Neural+LM: 84.6, 70.4

**SRL** (+2.9 F1)

**Coreference** (+3.2 F1)
SOTA For Many Others Tasks

- **SNLI**: Previous SOTA 88.1, Baseline 88.7, Baseline+LM 88.7
- **SQuAD**: Previous SOTA 84.3, Baseline 81.1, Baseline+LM 85.3
- **Coref**: Previous SOTA 67.2, Baseline 70.4, Baseline+LM 70.4
- **SRL**: Previous SOTA 81.7, Baseline 81.4, Baseline+LM 84.6
- **NER**: Previous SOTA 91.9, Baseline 90.2, Baseline+LM 92.2
- **Sentiment (SST)**: Previous SOTA 53.7, Baseline 51.4, Baseline+LM 54.7
What Does it Learn?
What Does it Learn?

Semantics:

• Supervised WSD task [Miller et al., 1994]
• Use N-th layer in NN classifier
### What Does it Learn?

#### Semantics:

- Supervised WSD task [Miller et al., 1994]
- Use N-th layer in NN classifier

<table>
<thead>
<tr>
<th>Layer</th>
<th>Avg. F1 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>65.9</td>
</tr>
<tr>
<td>2</td>
<td>69.0</td>
</tr>
<tr>
<td>Iacobacci (2016)</td>
<td>70.1</td>
</tr>
</tbody>
</table>
What Does it Learn?

Semantics:
- Supervised WSD task [Miller et al., 1994]
- Use N-th layer in NN classifier

Syntax:
- Label POS corpus [Marcus et al., 1993]
- Learn classifier on N-th layer
What Does it Learn?

Semantics:
- Supervised WSD task [Miller et al., 1994]
- Use N-th layer in NN classifier

Syntax:
- Label POS corpus [Marcus et al., 1993]
- Learn classifier on N-th layer

Accuracy

Layer 1
Layer 2
Ling et al. (2015)

Avg. F1 (%)

Layer 1
Layer 2
Iacobacci (2016)
Where Will the Data Come From???

Option 1: Semi-supervised learning

• E.g. word2vec and GloVe are in wide use [Mikolov et al., 2013; Pennington et al., 2014]
• Can we learn better word representations?

Option 2: Supervised learning

• Can we gather more direct forms of supervision?
A First Data Step: QA-SRL

- Introduce a new SRL formulation with no frame or role inventory
- Use question-answer pairs to model verbal predicate-argument relations
- Annotated over 3,000 sentences in weeks with non-expert, part-time annotators
- Showed that this data is high-quality and learnable

[He et al, 2015]
The rent rose 10% from $3000 to $3300.

Frameset: rise.01, go up

Arg1-: Logical subject, patient, thing rising
Arg2-EXT: EXT, amount risen
Arg3-DIR: start point
Arg4-LOC: end point
Argm-LOC: medium

- Depends on pre-defined frame inventory, requires syntactic parses
- Annotators need to:
  1) Identify the Frameset
  2) Find arguments in the parse
  3) Assign labels accordingly
- If frame doesn’t exist, create new

The Proposition Bank: An Annotated Corpus of Semantic Roles, Palmer et al., 2005
http://verbs.colorado.edu/propbank/framesets-english/rise-v.html
Our Annotation Scheme

Given sentence and a verb:

They *increased* the rent this year.
Our Annotation Scheme

Given sentence and a verb:

They *increased* the rent this year.

Step 1: Ask a question about the verb:

Who increased something?
Our Annotation Scheme

Given sentence and a verb:
They *increased* the rent this year.

**Step 1: Ask a question about the verb:**
Who increased something?

**Step 2: Answer with words in the sentence:**
They
Our Annotation Scheme

Given sentence and a verb:

They *increased* the rent this year.

**Step 1:** Ask a question about the verb:

Who increased something?

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

They

**Step 3:** Repeat, write as many QA pairs as possible ...

They
Our Annotation Scheme

Given sentence and a verb:

They *increased* the rent this year.

**Step 1:** Ask a question about the verb:

Who increased something?

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

They

**Step 3:** Repeat, write as many QA pairs as possible ...

What is increased?

the rent

When is something increased?

this year
Our Method: Q/A Pairs for Semantic Relations

The rent rose 10% from $3000 to $3300

Wh-Question                      Answer

What rose?                       the rent
How much did something rise?    10%
What did something rise from?   $3000
What did something rise to?     $3300
# Wh-words vs. PropBank Roles

<table>
<thead>
<tr>
<th></th>
<th>Who</th>
<th>What</th>
<th>When</th>
<th>Where</th>
<th>Why</th>
<th>How</th>
<th>HowMuch</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARG0</td>
<td>1575</td>
<td>414</td>
<td>3</td>
<td>5</td>
<td>17</td>
<td>28</td>
<td>2</td>
</tr>
<tr>
<td>ARG1</td>
<td>285</td>
<td>2481</td>
<td>4</td>
<td>25</td>
<td>20</td>
<td>23</td>
<td>95</td>
</tr>
<tr>
<td>ARG2</td>
<td>85</td>
<td>364</td>
<td>2</td>
<td>49</td>
<td>17</td>
<td>51</td>
<td>74</td>
</tr>
<tr>
<td>ARG3</td>
<td>11</td>
<td>62</td>
<td>7</td>
<td>8</td>
<td>4</td>
<td>16</td>
<td>31</td>
</tr>
<tr>
<td>ARG4</td>
<td>2</td>
<td>30</td>
<td>5</td>
<td>11</td>
<td>2</td>
<td>4</td>
<td>30</td>
</tr>
<tr>
<td>ARG5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>AM-ADV</td>
<td>5</td>
<td>44</td>
<td>9</td>
<td>2</td>
<td>25</td>
<td>27</td>
<td>6</td>
</tr>
<tr>
<td>AM-CAU</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>23</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>AM-DIR</td>
<td>0</td>
<td>6</td>
<td>1</td>
<td>13</td>
<td>0</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>AM-EXT</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>AM-LOC</td>
<td>1</td>
<td>35</td>
<td>10</td>
<td>89</td>
<td>0</td>
<td>13</td>
<td>11</td>
</tr>
<tr>
<td>AM-MNR</td>
<td>5</td>
<td>47</td>
<td>2</td>
<td>8</td>
<td>4</td>
<td>108</td>
<td>14</td>
</tr>
<tr>
<td>AM-PNC</td>
<td>2</td>
<td>21</td>
<td>0</td>
<td>1</td>
<td>39</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>AM-PRD</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>AM-TMP</td>
<td>2</td>
<td>51</td>
<td>341</td>
<td>2</td>
<td>11</td>
<td>20</td>
<td>10</td>
</tr>
</tbody>
</table>
Advantages
• Easily explained
• No pre-defined roles, few syntactic assumption
• Can capture implicit arguments
• Generalizable across domains

Limitations
• Only modeling verbs (for now)
• Not annotating verb senses directly
• Can have multiple equivalent questions

Challenges
• What questions to ask?
• How much data do we need?
• Can we generalize to other tasks, such as coref?
Does the Recipe Work for Broad Coverage Semantics?

Step 1: Gather lots of training data!

Challenge 1: Data is costly and limited (e.g. linguists required to label PennTreebank / OntoNotes)

Step 2: Apply Deep Learning!!

Challenge 2: Pipeline of structured prediction problems with cascading errors (e.g. POS->Parsing->SRL->Coref)

Step 3: Observe Impressive Gains!!!
Contributions

Models

• End-to-end deep learning for SRL and coreference
• No preprocessing (e.g. no parser or POS tagger)

Data

• Contextualized word embeddings from a language model
• First steps towards scalable data annotation
The End: Questions?

Future Directions

• Multi-task learning, given architectural similarities
• Multi-lingual should work, in theory…
• Need to scale up data annotation efforts, and focus on out of domain performance

Recent Release

• AllenNLP: Deep Learning Semantic NLP toolkit
• See demos and code at AllenNLP.org