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: 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 
observe

ARG0

ARG1 an X-ray flare 400 times brighter than usual

TMP
On January 5, 2015

Example Tasks:

Coreference: clustering NPs

Semantic Role Labeling: who did what, etc.

Many applications:

Question Answering
Information Extraction
Machine Translation
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.

The robot broke my favorite mug with a wrench.

My mug broke into pieces immediately.

<table>
<thead>
<tr>
<th>Frame: break.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>role</td>
</tr>
<tr>
<td>ARG0</td>
</tr>
<tr>
<td>ARG1</td>
</tr>
<tr>
<td>ARG2</td>
</tr>
<tr>
<td>ARG3</td>
</tr>
<tr>
<td>ARG4</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

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
- prediction
- Hard constraints

He et al., 2017
The cats love hats.

Input (sentence and predicate):

BIO output:

(Begin, Inside, Outside)

Final SRL output:
(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.
Ablations

Datasets

CoNLL 2005

Results

CoNLL 2012 (OntoNotes)

Results

Full model

No highway

No orthonormal init.

No dropout

Without dropout, model overfits at ~300 epochs.

Without orthonormal initialization, the deep model learns very slowly

(single model, on CoNLL05 Dev)
Error Breakdown
Oracle Transformations

Fix Label: [We] *fly* to NYC tomorrow.
Labeling error 29%

Split/Merge span:
I *eat* [pasta with delight].
Attachment error 25%
Confusion matrix for labeling errors (column normalized)

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

- Argument-adjunct distinctions are difficult even for expert annotators!
Takeaway

— Traditionally hard tasks, such as argument-adjunct distinction and PP attachment decisions are still challenging!
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 an X-ray flare 400 times brighter</td>
<td>On January</td>
</tr>
<tr>
<td>than usual</td>
<td>5, 2015</td>
</tr>
<tr>
<td>TMP</td>
<td></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
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.
# Coreference Resolution

## 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. |

## Cluster #1

| Cluster #1 | A fire in a Bangladeshi garment factory | 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.

<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>
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.

**Input document**

<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**

**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>

**Hand-engineered rules**

<table>
<thead>
<tr>
<th>Mention #1</th>
<th>Mention #2</th>
<th>Coreferent?</th>
</tr>
</thead>
<tbody>
<tr>
<td>A fire in a Bangladeshi garment factory</td>
<td>garment</td>
<td>✓</td>
</tr>
<tr>
<td>garment</td>
<td>factory</td>
<td>✓</td>
</tr>
<tr>
<td>factory</td>
<td>at least 37 people dead and 100 hospitalized</td>
<td>✓</td>
</tr>
</tbody>
</table>
Previous Approach:
Rule-based pipeline

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

Replies 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

Span representation

Bidirectional LSTM

Word & character embeddings

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

Bidirectional LSTM

Word & character embeddings

Boundary representations

the Postal Service

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
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
Electric said the
the Postal Service
Service contacted the
the company
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.
Mention Ranking

- Reason over all possible spans
- Assign an antecedent to every span

\[ y_3 \in \{\epsilon, 1, 2\} \]

<table>
<thead>
<tr>
<th>Span</th>
<th>Antecedent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>A fire</td>
</tr>
<tr>
<td>3</td>
<td>A fire in</td>
</tr>
<tr>
<td></td>
<td>...</td>
</tr>
<tr>
<td>M</td>
<td>out</td>
</tr>
</tbody>
</table>
Mention Ranking

- Reason over all possible spans
- Assign an antecedent to every span

\[ y_3 \in \{\epsilon, 1, 2\} \]

<table>
<thead>
<tr>
<th>Span</th>
<th>Antecedent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>A fire</td>
</tr>
<tr>
<td>3</td>
<td>A fire in</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>M</td>
<td>out</td>
</tr>
</tbody>
</table>

\( \epsilon \) : no coreference link
Mention Ranking

• Reason over all possible spans
• Assign an antecedent to every span

\[ y_3 \in \{\epsilon, 1, 2\} \]

<table>
<thead>
<tr>
<th>Span</th>
<th>Antecedent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>A fire</td>
</tr>
<tr>
<td>3</td>
<td>A fire in</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>M</td>
<td>out</td>
</tr>
</tbody>
</table>

Coreference link from span 1 to span 3
Mention Ranking

- Reason over all possible spans
- Assign an antecedent to every span

\[ y_3 \in \{\varepsilon, 1, 2\} \]

<table>
<thead>
<tr>
<th>Span</th>
<th>Antecedent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>A fire</td>
</tr>
<tr>
<td>3</td>
<td>A fire in</td>
</tr>
<tr>
<td></td>
<td>...</td>
</tr>
<tr>
<td>M</td>
<td>out</td>
</tr>
</tbody>
</table>

Coreference link from span 2 to span 3
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.

<table>
<thead>
<tr>
<th>Span</th>
<th>Antecedent ((y_i))</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>(\epsilon)</td>
</tr>
<tr>
<td>A fire</td>
<td>(\epsilon)</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>a Bangladeshi garment factory</td>
<td>(\epsilon)</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>(\epsilon)</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. 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>...</td>
<td>...</td>
</tr>
<tr>
<td>a Bangladeshi garment factory</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>
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>
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

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

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

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

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
Coreference Results

Durrett & Klein (2013) 60.3
Björkelund & Kuhn (2014) 61.6
Martschat & Strube (2015) 62.5

Linear 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>
</tbody>
</table>

- Neural models: Wiseman et al. (2016), Clark & Manning (2016)
Coreference Results

Pipelined models

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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<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>
<td>64.2</td>
<td>65.7</td>
<td>67.2</td>
<td>68.8</td>
</tr>
</tbody>
</table>

Pipelined models

End-to-end models
Mention Recall

- Raghunathan et al. (2010)
- Our model (actual threshold)
- Our model (various thresholds)
Mention Recall

- Raghunathan et al. (2010)
- Our model (actual threshold)
- Our model (various thresholds)

92.7% @ 0.4 spans per word
Head-finding Agreement

% of constituent spans with predicted heads that agree with syntactic heads

% agreement

Span width

1 2 3 4 5 6 7 8 9 10
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.
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.

Conflating relatedness with paraphrasing
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???

**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

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

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

General Electric said the Postal Service contacted the company

$$\text{LM Embeddings} = \alpha_1 \ + \ \alpha_2 \ + \ \alpha_3$$

Step 2: Compute linear function of pre-trained 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

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

**Step 3:** Learn weights for each end task
Best Single System Results

<table>
<thead>
<tr>
<th></th>
<th>Feature Based</th>
<th>Neural</th>
<th>Neural+LM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Avg. F1 (%)</td>
<td>79.9</td>
<td>81.7</td>
<td>84.6</td>
</tr>
<tr>
<td><strong>SRL (+2.9 F1)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Feature Based</th>
<th>Neural</th>
<th>Neural+LM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Avg. F1 (%)</td>
<td>62.5</td>
<td>67.2</td>
<td>70.4</td>
</tr>
<tr>
<td><strong>Coreference (+3.2 F1)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
SOTA For Many Others Tasks

- SNLI: Previous SOTA 88.1, Baseline 88.8, Baseline+LM 88.7
- SQuAD: Previous SOTA 84.3, Baseline 81.1, Baseline+LM 88
- 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?

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
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]
Previous Method: Annotation with Frames

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.

**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

Wh-Question

What rose?

How much did something rise?

What did something rise from?

What did something rise to?

Answer

the rent

10%

$3000

$3300
Cost and Speed

- Cost per Verb:
  - Newswire: $0.58
  - Wikipedia: $0.45

- Cost per Sentence:
  - Newswire: $1.57
  - Wikipedia: $1.01

- Time per Sentence:
  - Newswire: 9min
  - Wikipedia: 6min

- Part-time freelancers from upwork.com (hourly rate: $10)
- ~2h screening process for native English proficiency
## 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