Semantic Role Labeling with Labeled Span Graph Networks

Luheng He
Google A.I.
Oct. 2018

Work done at the University of Washington
On January 13, 2018, a false ballistic missile alert was issued via the Emergency Alert System and Commercial Mobile Alert System over television, radio, and cellphones in the U.S. state of Hawaii. The alert stated that there was an incoming ballistic missile threat to Hawaii, advised residents to seek shelter, and concluded "This is not a drill". The message was sent at 8:07 a.m. local time.
On January 13, 2018, a false ballistic missile alert was issued via the Emergency Alert System and Commercial Mobile Alert System over television, radio, and cellphones in the U.S. state of Hawaii. The alert stated that there was an incoming ballistic missile threat to Hawaii, advised residents to seek shelter, and concluded “This is not a drill”.

The message was sent at 8:07 a.m. local time.

From Wikipedia: 2018 Hawaii false missile alert. Only part of the structures are visualized.
On January 13, 2018, a false ballistic missile alert was issued via the Emergency Alert System and Commercial Mobile Alert System over television, radio, and cellphones in the U.S. state of Hawaii. The alert stated that there was an incoming ballistic missile threat to Hawaii, advised residents to seek shelter, and concluded "This is not a drill".

The message was sent at 8:07 a.m. local time.

From Wikipedia: 2018 Hawaii false missile alert. Only part of the structures are visualized.
Another example: Relation Extraction on Scientific Documents

To reduce ambiguity, the [MORphological PArser] (MORPA) method is provided with a [PCFG] method... It combines context-free grammar method with...

[MORPA] method is a fully implemented parser method developed for a [text-to-speech system] task.

Annotation with entities, relations, and coreference

SciERC (Entity, Relation, Coreference): Luan et al., 2018
Another example: Relation Extraction on Scientific Documents

To reduce [ambiguity]OtherS, the [MORphological PArser]MORPA method is provided with a [PCFG] method...

[It]Generic combines [context-free grammar] method with...


Used-for developed for a [text-to-speech system] task.

Annotation with entities, relations, and coreference

SciERC (Entity, Relation, Coreference): Luan et al., 2018
Another example: Relation Extraction

Goal: A unified model for all these tasks.

Challenge: Very different structures, task-specific pipelines/features/architectures …

This talk:
1) Build end-to-end models for SRL.
2) Generalizes such model to all tasks.

Annotation with entities, relations, and coreference.

SciERC (Entity, Relation, Coreference): Luan et al., 2018
Contributions
Contributions

• **End-to-end** prediction of SRL structure, without relying on NLP pipeline.
Contributions

- **End-to-end** prediction of SRL structure, without relying on NLP pipeline.
- **Almost 40% error reduction** over best pre-neural model despite being much simpler.
Contributions

- **End-to-end** prediction of SRL structure, without relying on NLP pipeline.

- **Almost 40% error reduction** over best pre-neural model despite being much simpler.

- First end-to-end result for **jointly predicting predicates and argument spans**.
Contributions

• **End-to-end** prediction of SRL structure, without relying on NLP pipeline.

• **Almost 40% error reduction** over best pre-neural model despite being much simpler.

• First end-to-end result for **jointly predicting predicates and argument spans**.

• Joint modeling for a variety of span-based tasks, opens up opportunities for **full-text understanding**.
Semantic Role Labeling (SRL)

**Predicate** | **Arguments**
---|---
visit | Who is the **visitor**: [Many tourists]
 | What is **visited**: [Disney]
 | What **purpose**: [to meet their favorite cartoon characters]
Semantic Role Labeling (SRL)

Many tourists visit Disney to meet their favorite cartoon characters.

Frame: visit.01

**Predicate**: visit

**Arguments**

**ARG0**: [Many tourists]
**ARG1**: [Disney]
**AM-PRP**: [to meet … characters]

The Proposition Bank: An Annotated Corpus of Semantic Roles, Palmer et al., 2005
Semantic Role Labeling (SRL)

Many tourists visit Disney to meet their favorite cartoon characters

Frame: visit.01

<table>
<thead>
<tr>
<th>role</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARG0</td>
<td>visitor</td>
</tr>
<tr>
<td>ARG1</td>
<td>visited</td>
</tr>
</tbody>
</table>

**Core arguments:** Verb-specific roles (A0-A5)

**Adjuncts:** Arg-modifier (AM-) roles shared across verbs

The Proposition Bank: An Annotated Corpus of Semantic Roles, Palmer et al., 2005
Many tourists visit Disney to meet their favorite cartoon characters.

**Predicate**
- visit
- meet

**Arguments**
- **ARG0**: [Many tourists]
- **ARG1**: [Disney]
- **AM-PRP**: [to meet their favorite cartoon characters]
SRL Task: Given (gold) predicate, predict arguments

**Predicate**  |  **Arguments**
--- | ---
visit  |  **ARG0**: [Many tourists]  
         |  **ARG1**: [Disney]  
         |  **AM-PRP**: [to meet their favorite cartoon characters]
meet  |  **ARG0**: [Many tourists]  
         |  **ARG1**: [their favorite cartoon characters]

**Most Span-based SRL Tasks**: given gold predicates, predict the argument spans and labels.
Outline

**Predicting SRL with Deep BiLSTMs**
— DeepSRL (He et al., 2017)

**An End-to-End, Span-based SRL Model**
— Labeled Span Graph Network (He et al., 2018)

**Towards Unified and Full-text Semantic Analysis**
— Multi-task learning with LSGN; ScienceIE (Luan et al., 2018)
SRL Systems: Pipelined vs. BIO-based

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: Pipelined vs. BIO-based

Pipeline Systems

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

Syntactic Parser

Punyakanok et al., 2008
Täckström et al., 2015
FitzGerald et al., 2015
SRL Systems: Pipelined vs. BIO-based

Pipe line Systems

sentence, predicate

syntactic features

argument id.

candidate argument spans

labeling

labeled arguments

ILP/DP

prediction

Syntactic Parser

Hand-engineereded Rules
1. Arguments cannot overlap with the predicate.
2. If a predicate is outside a clause, its argument is a clause.
3. Arguments cannot exclusively overlap with the

Punyakanok et al., 2008
Täckström et al., 2015
FitzGerald et al., 2015
SRL Systems: Pipelined vs. BIO-based

Pipeline Systems

- **Sentence, predicate**
- **Syntactic features**
  - **Argument id.**
  - **Candidate argument spans**
  - **Labeling**
  - **Labeled arguments**

- **ILP/DP**
  - **Prediction**

Syntactic Parser

Hand-engineered Rules

1. Arguments cannot overlap with the predicate.
2. If a predicate is outside a clause, its argument spans.

Hand-engineered Features

- Starting word of $s$
- Ending word of $s$
- Head word of $s$
- Bag of words in $s$
- A bias feature
- Dependency path between $s$’s head and $t$
- The set of dependency labels of $t$’s children
- Dependency path conjoined with the tag of $s$’s head
- Dep. path conjoined with the cluster of $s$’s head
- Position of $s$ w.r.t. $t$ (before, after, overlap or same)

Punyakanok et al., 2008
Täckström et al., 2015
FitzGerald et al., 2015
SRL Systems: Pipelined vs. BIO-based

Pipeline Systems

Syntactic Parser

Hand-engineered Rules
1. Arguments cannot overlap with the predicate.
2. If a predicate is outside a clause, its argument is a clause.

Hand-engineered Features

Post-Processing

Punyakanok et al., 2008
Täckström et al., 2015
FitzGerald et al., 2015
SRL Systems: Pipelined vs. BIO-based

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

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

**BIO-based Systems**
- sentence, predicate
- word-level features
- Deep BiLSTM + CRF layer
- BIO sequence
- Viterbi
- prediction

*Collobert et al., 2011*  
*Zhou and Xu, 2015*  
*Wang et. al, 2015*
SRL Systems: Pipelined vs. BIO-based

**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

**BIO-based Systems**
- sentence, predicate
- word-level features
- Deep BiLSTM + CRF layer
- BIO sequence
- Viterbi
- prediction

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

**DeepSRL**
- sentence, predicate
- Deep BiLSTM
- BIO sequence
- Hard constraints
- prediction

He et al., 2017

No global normalization
the [V=0] cats [V=0] love [V=1] hats [V=0]
The cats love hats.

(2) Deep BiLSTM tagger

(1) Inputs words and target predicate
(1) Inputs words and target predicate

(2) Deep BiLSTM tagger

(3) Highway connections, variational dropouts, etc.
(1) Inputs words and target predicate

(2) Deep BiLSTM tagger

(3) Highway connections, variational dropouts, etc.

(4) Viterbi decoding with hard constraints at test time

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 0.1  I-ARG0 0.1  B-ARG1 0.7  I-ARG1 0.2
B-ARG0 0.001 I-ARG0 0.001 B-ARG1 0.001 I-ARG1 0.001

B-V 0.95

the [V=0]  cats [V=0]  love [V=1]  hats [V=0]
Strengths:
No syntactic preprocessing; Easy to implement (can use off-the-shelf sequential tagger)

Limitations:
Needs to re-process the same sentence multiple times, if sentence has multiple predicates
CoNLL 2005 (Original PropBank) Results

F1

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

90
85
80
75
70
65
60
Toutanova05*
Täckström15
FitzGerald15*
Zhou15
DeepSRL
DeepSRL*

Pipeline models
Deep BIO Models
# CoNLL 2005 (Original PropBank) Results

<table>
<thead>
<tr>
<th>Model</th>
<th>WSJ Test</th>
<th>Brown (out-domain) Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Toutanova05*</td>
<td>80.3</td>
<td></td>
</tr>
<tr>
<td>Täckström15</td>
<td>79.9</td>
<td></td>
</tr>
<tr>
<td>FitzGerald15*</td>
<td>80.3</td>
<td></td>
</tr>
<tr>
<td>Zhou15</td>
<td>82.8</td>
<td>83.1</td>
</tr>
<tr>
<td>DeepSRL</td>
<td>80.3</td>
<td></td>
</tr>
<tr>
<td>DeepSRL*</td>
<td>80.3</td>
<td></td>
</tr>
<tr>
<td>DeepSRL*</td>
<td>84.6</td>
<td></td>
</tr>
</tbody>
</table>

*: Ensemble models

**Pipeline models**

**Deep BIO Models**
CoNLL 2005 (Original PropBank) Results

F1 scores for different models on WSJ Test and Brown (out-domain) Test. * denotes ensemble models.

- Toutanova05*: 80.3 (WSJ), 68.8 (Brown)
- Täckström15: 79.9 (WSJ), 71.3 (Brown)
- FitzGerald15*: 80.3 (WSJ), 72.2 (Brown)
- Zhou15: 82.8 (WSJ), 69.4 (Brown)
- DeepSRL: 83.1 (WSJ), 72.1 (Brown)
- DeepSRL*: 84.6 (WSJ), 73.6 (Brown)

Models are divided into pipeline models and deep BIO models.
CoNLL 2005 (Original PropBank) 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>Toutanova05*</td>
<td>80.3</td>
<td>68.8</td>
<td>80.3</td>
</tr>
<tr>
<td>Täckström15</td>
<td>79.9</td>
<td>71.3</td>
<td>79.9</td>
</tr>
<tr>
<td>FitzGerald15*</td>
<td>80.3</td>
<td>71.3</td>
<td>80.3</td>
</tr>
<tr>
<td>Zhou15</td>
<td>82.8</td>
<td>82.8</td>
<td>82.8</td>
</tr>
<tr>
<td>DeepSRL</td>
<td>83.1</td>
<td>83.1</td>
<td>83.1</td>
</tr>
<tr>
<td>DeepSRL*</td>
<td>84.6</td>
<td>73.6</td>
<td>84.6</td>
</tr>
</tbody>
</table>

- 20% Error reduction over pipelined systems!
- Still way to go for out-domain data!
CoNLL 2012 (OntoNotes) Results

F1
- Pradhan12: 77.5
- Täckström15: 79.4
- FitzGerald15*: 80.2
- Zhou15: 81.5
- DeepSRL: 81.7
- DeepSRL*: 83.4

Larger dataset with 6 domains, contains nominal predicates

*: Ensemble models
CoNLL 2012 (OntoNotes) Results

Larger dataset with 6 domains, contains **nominal predicates**

<table>
<thead>
<tr>
<th>Model</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pradhan12</td>
<td>77.5</td>
</tr>
<tr>
<td>Täckström15</td>
<td>79.4</td>
</tr>
<tr>
<td>FitzGerald15*</td>
<td>80.2</td>
</tr>
<tr>
<td>Zhou15</td>
<td>81.5</td>
</tr>
<tr>
<td>DeepSRL</td>
<td>81.7</td>
</tr>
<tr>
<td>DeepSRL*</td>
<td>83.4</td>
</tr>
</tbody>
</table>

**Pipeline models**

**Deep BIO models**

*Ensemble models*
CoNLL 2012 (OntoNotes) Results

F1

CoNLL 2012 Test
*: Ensemble models

Larger dataset with 6 domains, contains nominal predicates

83.4 81.7 81.5 80.2 79.4 77.5

However, the model still relies on gold predicates …

Pipeline models

Deep BIO models

Pradhan12 Täckström15 FitzGerald15* Zhou15 DeepSRL DeepSRL*
Real Scenario: No Gold Predicates!

**End-to-end SRL:**
Given sentence, predict all predicates as well as their arguments.

- Sentence, predicate
- Deep BiLSTM
- BIO sequence
- Hard constraints
- Prediction
Real Scenario: No Gold Predicates!

**End-to-end SRL:**
Given sentence, predict all predicates as well as their arguments.

**Pipelined approach:**
Identify predicates first, then run the BIO tagger for each predicate.

(No way to recover from recall loss at predicate ID stage …)
Real Scenario: No Gold Predicates!

**End-to-end SRL:**
Given sentence, predict all predicates as well as their arguments.

**Pipelined approach:**
Identify predicates first, then run the BIO tagger for each predicate.

(No way to recover from recall loss at predicate ID stage …)
Real Scenario: No Gold Predicates!

**End-to-end SRL:**
Given sentence, predict all predicates as well as their arguments.

**Pipelined approach:**
Identify predicates first, then run the BIO tagger for each predicate.

(No way to recover from recall loss at predicate ID stage ...)

---

sentence

BiLSTM

predicates

for each predicate
End-to-End SRL Result

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Gold Predicate</th>
<th>End-to-End</th>
</tr>
</thead>
<tbody>
<tr>
<td>WSJ</td>
<td>84.6</td>
<td>83.4</td>
</tr>
<tr>
<td>Brown</td>
<td>73.6</td>
<td></td>
</tr>
<tr>
<td>OntoNotes</td>
<td></td>
<td>83.4</td>
</tr>
</tbody>
</table>
End-to-End SRL Result

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Gold Predicate</th>
<th>End-to-End</th>
</tr>
</thead>
<tbody>
<tr>
<td>WSJ</td>
<td>84.6</td>
<td>82.7</td>
</tr>
<tr>
<td>Brown</td>
<td>73.6</td>
<td>70.1</td>
</tr>
<tr>
<td>OntoNotes</td>
<td>83.4</td>
<td>78.4</td>
</tr>
</tbody>
</table>
End-to-End SRL Result

Larger performance drop on Brown (out-domain) and OntoNotes (nominal predicates)
Predicting SRL with Deep BiLSTMs
— DeepSRL

An End-to-End, Span-based SRL Model
— Labeled Span Graph Network (LSGN)

Towards Unified and Full-text Semantic Analysis
— Multi-task learning with LSGN; ScienceIE (Luan et al., 2018)
Intuition: SRL as Span-Span Relations

Many tourists visit Disney to meet their favorite cartoon characters.
Intuition: SRL as Span-Span Relations

Challenges:
1. Span can nest within each other.
2. Too many possible edges (n^2 argument spans & n predicates).

Many tourists visit Disney to meet their favorite cartoon characters.
Labeled Span Graph Network (LSGN)

**LSG**: A graph with nodes as spans and labeled edges.

**LSGN**: An end-to-end network for predicting an LSG.
Labeled Span Graph Network (LSGN)

**LSG**: A graph with nodes as spans and labeled edges.  
**LSGN**: An end-to-end network for predicting an LSG.

Many NLP structures can be considered as an LSG:

- **SRL** (*First end-to-end model!*)
- Coreference resolution (Lee et al. 2017)
- Named entity recognition and relation extraction (Luan et al., 2018)
DeepSRL Architecture (Revisit)

Many tourists visit Disney to meet their favorite cartoon characters
DeepSRL Architecture (Revisit)

```
Many [0]    tourists [0]    visit [1]    Disney [0]    to [0]    meet [0]    their [0]    favorite [0]    cartoon [0]    characters [0]
```

**Input sentence**
Many tourists visit Disney to meet their favorite cartoon characters.

**Target Predicate**
```
[0] [0] [1] [0] [0] [0] [0] [0] [0] [0]
```

**Word & Pred. Embeddings**

**Highway BiLSTMs**

**Tagging Softmax**

**Output Labels**
``` B-A0 I-A0 B-V B-V ... ```
Many tourists visit Disney to meet their favorite cartoon characters.
LSGN Architecture: Overview

(1) Construct span representations for all \( n^2 \) spans!

Span Representation
Highway BiLSTMs
Word & Char Embeddings
Input sentence

No predicate input!

Many tourists visit Disney to meet their favorite cartoon characters
Many tourists visit Disney to meet their favorite cartoon characters.

1) Construct span representations for all $n^2$ spans!

2) Local classifier over labels (including NULL) for all possible (predicate, argument) pairs.
Many tourists visit Disney to meet their favorite cartoon characters.
(1) Span Representations

(2) Local Label Classifiers

(3) Span Pruning

Many tourists visit Disney to meet their favorite cartoon characters

(Same as Lee et al., 2017)
Many tourists visit Disney to meet their favorite cartoon characters.

(Same as Lee et al., 2017)
(1) Span Representations

\[ [\text{BiLSTM}(w_1 : w_n)_{\text{START}}, \text{BiLSTM}(w_1 : w_n)_{\text{END}}] \]

\[ \sum_{i=\text{START}}^{\text{END}} \text{SOFTMAX}(a_{\text{START}} : a_{\text{END}})_i w_i \]

(2) Local Label Classifiers

(3) Span Pruning

Span Representation

Highway BiLSTMs

Word & Char Embeddings

Input sentence

Many tourists visit Disney to meet their favorite cartoon characters

(Same as Lee et al., 2017)
Many tourists visit Disney to meet their favorite cartoon characters.
Disney to meet many tourists their favorite cartoon characters

(1) Span Representations

- Word & Char Embeddings
- Highway BiLSTMs
- Span Representation
- Node & Edge Scores
- Labeling Softmax

(2) Local Label Classifiers

- (1) Span Representations
- (2) Local Label Classifiers
- (3) Span Pruning

(3) Span Pruning

- ARG0
- ARG1
- ARG2
- AM-TMP
- \( \epsilon \) (No Edge)
- many tourists
- meet
Many tourists visit Disney to meet their favorite cartoon characters.
Disney to meet favorite cartoon characters

Many tourists visit Disney to meet their favorite cartoon characters

\[
P(y_{\text{pred}, \text{arg}} = l \mid X) \propto \exp(\phi(\text{pred}, \text{arg}, l))
\]
\[ \phi(\text{pred}, \text{arg}, l) = \Phi_a(\text{arg}) + \Phi_p(\text{pred}) + \Phi_{rel}(\text{arg}, \text{pred}) \]
$$\phi(\text{pred}, \text{arg}, l) = \Phi_a(\text{arg}) + \Phi_p(\text{pred}) + \Phi_{\text{rel}}^{(l)}(\text{arg}, \text{pred})$$
Many tourists meet

\( \phi(\text{pred}, \text{arg}, l) = \Phi_a(\text{arg}) + \Phi_p(\text{pred}) + \Phi_{rel}^{(l)}(\text{arg}, \text{pred}) \)

\( \phi(\text{“Many tourists”}, \text{“meet”}, \epsilon) = 0 \)

\( \Phi_{rel}^{(\text{ARG0})}(\text{“Many tourists”}, \text{“meet”}) \)

\( \Phi_{rel}^{(\text{ARG1})}(\text{“Many tourists”}, \text{“meet”}) \)

\( \Phi_a(\text{“Many tourists”}) \)

\( \Phi_p(\text{“meet”}) \)

Many tourists

meet
Many tourists visit Disney to meet their favorite cartoon characters.

O(n²) arguments, O(n) predicates, \(\rightarrow O(n^3)\) edges!
Many tourists visit Disney to meet their favorite cartoon characters.

Span Representations

(1) Span Representations

Highway BiLSTMs

Word & Char Embeddings

(2) Local Label Classifiers

(3) Span Pruning

\[ \Phi_a(\text{"many tourists"}) = 2.5 \]

\[ \Phi_a(\text{"tourists visit Disney"}) = -0.8 \]

Only keep top \( O(n) \) spans using their unary scores.
End-to-End SRL Results

BIO-based, pipelined predicate ID

<table>
<thead>
<tr>
<th>Test Set</th>
<th>DeepSRL</th>
<th>LSGN</th>
<th>DeepSRL (Ensemble)</th>
<th>LSGN + ELMo</th>
</tr>
</thead>
<tbody>
<tr>
<td>CoNL05 WSJ Test</td>
<td>81.2</td>
<td>82.5</td>
<td>82.5</td>
<td>79.8</td>
</tr>
<tr>
<td>CoNL05 Brown Test</td>
<td>68.5</td>
<td>70.8</td>
<td>70.8</td>
<td></td>
</tr>
<tr>
<td>CoNLL2012 (OntoNotes)</td>
<td>76.8</td>
<td>79.8</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
End-to-End SRL Results

With **ELMo**, over 3 points improvement over SotA ensemble!

*ELMo: Deep Contextualized Word Representations, Peters et al., 2018*
End-to-End SRL

- New SotA (Strubell et al., 2018) with syntax-informed transformer model.

With ELMo, over 3 points improvement over SotA ensemble!

*ELMo: Deep Contextualized Word Representations, Peters et al., 2018
Gold Predicates: CoNLL 2005 SRL Results

- Täckström15: 79.9
- FitzGerald15*: 80.3
- Zhou15: 82.8
- DeepSRL17*: 84.6
- Tan18*: 86.1
- LSGN+ELMo: 87.4

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

Pipeline models

Deep BIO models
Gold Predicates: CoNLL 2005 SRL Results

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

<table>
<thead>
<tr>
<th>Model</th>
<th>WSJ Test</th>
<th>Brown (out-domain) Test</th>
<th>Deep BIO models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Täckström15</td>
<td>79.9</td>
<td>71.3</td>
<td>Pipeline models</td>
</tr>
<tr>
<td>FitzGerald15*</td>
<td>80.3</td>
<td>72.2</td>
<td>Transformer-style BIO-tagger</td>
</tr>
<tr>
<td>Zhou15</td>
<td>82.8</td>
<td>69.4</td>
<td></td>
</tr>
<tr>
<td>DeepSRL17*</td>
<td>84.6</td>
<td>73.6</td>
<td></td>
</tr>
<tr>
<td>Tan18*</td>
<td>86.1</td>
<td>74.8</td>
<td></td>
</tr>
<tr>
<td>LSGN+ELMo</td>
<td>87.4</td>
<td>80.4</td>
<td></td>
</tr>
</tbody>
</table>

Gold Predicates: CoNLL 2005 SRL Results

- **WSJ Test**
- **Brown (out-domain) Test**
- *Ensemble models*
Gold Predicates: CoNLL-2005 SRL Results

- **In-domain**: 40% error reduction over pre-neural models.
- **Out-domain**: Reaching 80% F1.
Gold Predicates: CoNLL 2005

- New results (Ouchi. et al, 2018) with span-selection model +ELMo.
## Related Work on Span-based Models

<table>
<thead>
<tr>
<th>Model</th>
<th>Nesting Spans</th>
<th>Span Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>BIO-Taggers</strong> (Collobert et al., 2010, Chiu and Nichols, 2016, DeepSRL)</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td><strong>Semi-Markov Models</strong> (Kong et al., 2016)</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>LSGN</strong></td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>
## Span-based vs. BIO

<table>
<thead>
<tr>
<th></th>
<th>Inputs</th>
<th>DeepSRL (BIO)</th>
<th>LSGN (Span-based)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Predicate Identification</strong></td>
<td>(Sentence, Predicate)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Global Consistency</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Long-range Dependency</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Sentences and Pipelined vs. Joint**

- **Inputs**:
  - DeepSRL (BIO): (Sentence, Predicate)
  - LSGN (Span-based): Sentence

- **Predicate Identification**:
  - DeepSRL (BIO): Pipelined
  - LSGN (Span-based): Joint

- **Global Consistency**:
  - DeepSRL (BIO): Pipelined
  - LSGN (Span-based): Joint

- **Long-range Dependency**:
  - DeepSRL (BIO): Pipelined
  - LSGN (Span-based): Joint
Span-based vs. BIO

<table>
<thead>
<tr>
<th>Inputs</th>
<th>DeepSRL (BIO)</th>
<th>LSGN (Span-based)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentence, Predicate</td>
<td>Sentence</td>
<td>Joint</td>
</tr>
</tbody>
</table>

- Due to the strong independence assumption LSGN makes

Global Consistency

Long-range Dependency
### Span-based vs. BIO

<table>
<thead>
<tr>
<th>Inputs</th>
<th>DeepSRL (BIO)</th>
<th>LSGN (Span-based)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Sentence, Predicate)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentence</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pipelined</td>
<td></td>
<td>Joint</td>
</tr>
<tr>
<td>Global Consistency</td>
<td>Due to the strong independence assumption LSGN makes</td>
<td></td>
</tr>
<tr>
<td>Long-range Dependency</td>
<td>By allowing direct interaction between predicates and arguments</td>
<td></td>
</tr>
</tbody>
</table>
Outline

Predicting SRL with Deep BiLSTMs
— DeepSRL

☑️ Accurate
☑️ No NLP pipeline
☑️ Joint predicate ID
☐ Full-text Semantics

An End-to-End, Span-based SRL Model
— Labeled Span Graph Network (LSGN)

Towards Unified and Full-text Semantic Analysis
— Multi-task learning with LSGN; ScienceIE (Luan et al., 2018)
On January 13, 2018, a false ballistic missile alert was issued via the Emergency Alert System and Commercial Mobile Alert System over television, radio, and cellphones in the U.S. state of Hawaii. The alert stated that there was an incoming ballistic missile threat to Hawaii, advised residents to seek shelter, and concluded "This is not a drill". The message was sent at 8:07 a.m. local time.

From Wikipedia: 2018 Hawaii false missile alert. Only part of the structures are visualized.
On January 13, 2018, a false ballistic missile alert was issued via the Emergency Alert System and Commercial Mobile Alert System over television, radio, and cellphones in the U.S. state of Hawaii. The alert stated that there was an incoming ballistic missile threat to Hawaii, advised residents to seek shelter, and concluded "This is not a drill". The message was sent at 8:07 a.m. local time.

From Wikipedia: 2018 Hawaii false missile alert. Only part of the structures are visualized.
Multi-task LSGN Architecture

Many tourists visit Disney to meet their favorite cartoon characters
Multi-task LSGN Architecture

Many tourists visit Disney to meet their favorite cartoon characters.
Multi-task LSGN Architecture

- **Input Document**: Many tourists visit Disney to meet their favorite cartoon characters

- **Word & Char Embeddings**

- **Highway BiLSTMs**

- **Span Representation**

- **Node & Edge Scores**

- **Labeling Softmax**

- **Lightweight, task-specific span classifiers**

- **Shared span representations**
Task-specific Span Classifiers

Softmax
Combined score
Edge score
Node score
Span Rep.

Coreference
(Lee et al., 2017)

Many tourists
Disney
their

SRL/Relation Extraction

... ARG0
ARG1

Many tourists
meet

NER

null
null
null
null

null
Task-specific Span Classifiers

Coreference
(Lee et al., 2017)

SRL/Relation Extraction

NER

Shared span representations! (Efficiency gain)
Task-specific Span Classifiers

Multi-task learning objective

Softmax
Combined score
Edge score
Node score
Span Rep.

Coreference
(SRL/Relation Extraction)
( Lee et al., 2017

Shared span representations! (Efficiency gain)
Two multi-tasking setups

1. “One model for everything”: Train one model to predict all the $n$ tasks. Performance is tuned on the average of the $n$ metrics …
Two multi-tasking setups

1. “One model for everything”: Train one model to predict all the $n$ tasks. Performance is tuned on the average of the $n$ metrics …

2. “Let the tasks help each other”: Train $n$ models, each predicts a target task, and treat the rest $n-1$ tasks as auxiliary. (Swayamdipta et al., 2018 …)
OntoNotes: Is it possible to have one model to do them all?

Previous systems

- DeepSRL17*
- Lee17*
- Li17
- LSGN (3 models)
- LSGN (1 model)

*: Ensemble models

<table>
<thead>
<tr>
<th>Task</th>
<th>DeepSRL17*</th>
<th>Lee17*</th>
<th>Li17</th>
<th>LSGN (3 models)</th>
<th>LSGN (1 model)</th>
</tr>
</thead>
<tbody>
<tr>
<td>E2E SRL</td>
<td>78.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coref</td>
<td></td>
<td>68.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NER</td>
<td></td>
<td></td>
<td></td>
<td>87.2</td>
<td></td>
</tr>
</tbody>
</table>

*Ensemble models
OntoNotes: Is it possible to have one model to do them all?

LSGN (3 models, no task sharing): Almost 1 point improvement over all previous SotA! (Coref. improvement due to larger model capacity)
OntoNotes: Is it possible to have one model to do them all?

One model for all tasks, treating them as equally important. (Only modest accuracy loss …)
ScienceERC Data (Luan et al., 2018): Can the tasks help each other?

To reduce ambiguity, the [MORphological PArser] is provided with a [PCFG] method...

It [Generic] combines [context-free grammar] method with...

[MORPA] method is a fully implemented [parser] method developed for a [text-to-speech system] task.

Softmax

Combined score

Edge score

Node score

Span Rep.

null

null

null

... MORPA PCFG

... Hyponym-of Used-for

... TASK METHOD

Coreference

Relation Extraction

Entity Extraction
ScienceERC: Can the tasks help each other?

Previous SotA. Uses dependency features.
ScienceERC: Can the tasks help each other?

ELMo is very helpful on this small datasets (500 documents)!
ScienceERC: Can the tasks help each other?

<table>
<thead>
<tr>
<th>Model</th>
<th>Relation</th>
<th>Coref.</th>
<th>Entity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Miwa16</td>
<td>31.4</td>
<td>36.6</td>
<td>39.3</td>
</tr>
<tr>
<td>Miwa16 +ELMo</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lee17 +ELMo</td>
<td>46.2</td>
<td>48.2</td>
<td>61.2</td>
</tr>
<tr>
<td>SciIE +ELMo +MTL</td>
<td>61.2</td>
<td>63.7</td>
<td>64.2</td>
</tr>
</tbody>
</table>

3 models for 3 tasks: Each task is tuned individually, bringing large gains from MTL.
ScienceERC: Ablations
ScienceERC: Ablations

Single-task
Multi-task

70
60
50
40
30

Entity  Relation  Coreference
ScienceERC: Ablations

<table>
<thead>
<tr>
<th>Task</th>
<th>Single-task</th>
<th>Multi-task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entity</td>
<td>65.7</td>
<td></td>
</tr>
<tr>
<td>Relation</td>
<td>37.9</td>
<td></td>
</tr>
<tr>
<td>Coreference</td>
<td>55.3</td>
<td></td>
</tr>
</tbody>
</table>
ScienceERC: Ablations

<table>
<thead>
<tr>
<th></th>
<th>Single-task</th>
<th>Multi-task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entity</td>
<td>65.7</td>
<td>68.1</td>
</tr>
<tr>
<td>Relation</td>
<td>37.9</td>
<td>39.5</td>
</tr>
<tr>
<td>Coreference</td>
<td>55.3</td>
<td>58</td>
</tr>
</tbody>
</table>
ScienceERC: Qualitative Analysis

With predicted coreference links, the system extracted less generic terms and more specific ones!
Conclusion

• A general framework for a variety of tasks.
Conclusion

- A general framework for a variety of tasks.

- Our recipe:
  1. Contextualized span representations
  2. Local label classifiers
  3. Greedy span pruning
Conclusion

• A general framework for a variety of tasks.

• Our recipe:
  1. Contextualized span representations
  2. Local label classifiers
  3. Greedy span pruning

• Multi-task learning works (sometimes)
Future Work

1. LSGNs for more NLP tasks.
2. Improve global consistency of the LSGN outputs with joint inference (e.g. Singh et al., 2013).
3. Pre-train transferrable span embeddings.
Links to Code and Data

1. DeepSRL:
   https://github.com/luheng/deep_srl

2. LSGN:
   https://github.com/luheng/lsgn

3. ScilE/ScienceERC (by Yi Luan):
   http://nlp.cs.washington.edu/scilE/
Many Thanks to my Collaborators!

in collaboration with Kenton Lee, Mike Lewis, Omer Levy, Yi Luan, and Luke Zettlemoyer