Learning Challenges in Natural Language Processing

Swabha Swayamdipta
April 04, 2019
NLP today
NLP today

Contextualized Representations

[Clark et. al., 2018]

[Howard & Ruder, 2018]

[Radford et. al., 2018]

[Devlin et. al., 2018]
NLP today

Large Language Model

Contextualized Representations
NLP today

Large Language Model

Contextualized Representations

Downstream Tasks

[Devlin et. al., 2018]

[Clark et. al., 2018]

[Howard & Ruder, 2018]

[Radford et. al., 2018]

[Clark et. al., 2018]

[Howard & Ruder, 2018]

[Radford et. al., 2018]

[Devlin et. al., 2018]

[Devlin et. al., 2018]
NLP today

Large Language Model

Contextualized Representations

Downstream Tasks
NLP today

Large Language Model

Downstream Tasks

Unsupervised

Supervised

Contextualized Representations

[Devlin et. al., 2018]

[Clark et. al., 2018]

[Howard & Ruder, 2018]

[Radford et. al., 2018]

[Clark et. al., 2018]
Wessex is chivalrous and charming, but semi-betrothed to Lady Ursula Glynde, whom he has not seen since her infancy. Wessex is repelled by the idea of having his wife thrust upon him and purposely avoids Lady Ursula. Unknown to Wessex, the Queen jealously guards him against Ursula, who is extremely beautiful. As soon as she realizes the Queen is keeping her away from Wessex, Ursula is angered. She believes she loves Wessex, for his nobility and goodness, and she is invested heavily in the betrothal. Although Ursula does not want to lose her independence by marrying, she seeks to frustrate the Queen’s plans and make Wessex notice her.

Who seeks to frustrate the Queen’s plans?
Wessex is chivalrous and charming, but semi-betrothed to Lady Ursula Glynde, whom he has not seen since her infancy. Wessex is repelled by the idea of having his wife thrust upon him and purposely avoids Lady Ursula. Unknown to Wessex, the Queen jealously guards him against Ursula, who is extremely beautiful. As soon as she realizes the Queen is keeping her away from Wessex, Ursula is angered. She believes she loves Wessex, for his nobility and goodness, and she is invested heavily in the betrothal. Although Ursula does not want to lose her independence by marrying, she seeks to frustrate the Queen’s plans and make Wessex notice her.

Who seeks to frustrate the Queen’s plans?
Wessex is chivalrous and charming, but semi-betrothed to Lady Ursula Glynde, whom he has not seen since her infancy. Wessex is repelled by the idea of having his wife thrust upon him and purposely avoids Lady Ursula. Unknown to Wessex, the Queen jealously guards him against Ursula, who is extremely beautiful. As soon as she realizes the Queen is keeping her away from Wessex, Ursula is angered. She believes she loves Wessex, for his nobility and goodness, and she is invested heavily in the betrothal. Although Ursula does not want to lose her independence by marrying, she seeks to frustrate the Queen’s plans and make Wessex notice her.

Who seeks to frustrate the Queen’s plans?

Wessex
Wessex is chivalrous and charming, but semi-betrothed to Lady Ursula Glynde, whom he has not seen since her infancy. Wessex is repelled by the idea of having his wife thrust upon him and purposely avoids Lady Ursula. Unknown to Wessex, the Queen jealously guards him against Ursula, who is extremely beautiful. As soon as she realizes the Queen is keeping her away from Wessex, Ursula is angered. She believes she loves Wessex, for his nobility and goodness, and she is invested heavily in the betrothal. Although Ursula does not want to lose her independence by marrying, she seeks to frustrate the Queen’s plans and make Wessex notice her.

Who seeks to frustrate the Queen’s plans?

Wessex
Wessex is chivalrous and charming, but semi-betrothed to Lady Ursula Glynde, whom he has not seen since her infancy. Wessex is repelled by the idea of having his wife thrust upon him and purposely avoids Lady Ursula. Unknown to Wessex, the Queen jealously guards him against Ursula, who is extremely beautiful. As soon as she realizes the Queen is keeping her away from Wessex, Ursula is angered. She believes she loves Wessex, for his nobility and goodness, and she is invested heavily in the betrothal. Although Ursula does not want to lose her independence by marrying, she seeks to frustrate the Queen’s plans and make Wessex notice her.

Who seeks to frustrate the Queen’s plans?

Wessex
Learning Challenges
Learning Challenges

Part I

Can we incorporate some priors about language to improve our models?

- Syntactic Scaffolds for Semantic Structures
  (EMNLP 2018)
Learning Challenges

Part I
Can we incorporate some priors about language to improve our models?

- Syntactic Scaffolds for Semantic Structures (EMNLP 2018)

Part II
What in our data is causing models to achieve high performance?

- Annotation Artifacts in Natural Language Inference Data (NAACL 2018)
Learning Challenge #1

Can we incorporate some priors about language?

Wessex is chivalrous and charming, but semi-betrothed to Lady Ursula Glynde, whom he has not seen since her infancy. Wessex is repelled by the idea of having his wife thrust upon him and purposely avoids Lady Ursula. Unknown to Wessex, the Queen jealously guards him against Ursula, who is extremely beautiful. As soon as she realizes the Queen is keeping her away from Wessex, Ursula is angered. She believes she loves Wessex, for his nobility and goodness, and she is invested heavily in the betrothal. Although Ursula does not want to lose her independence by marrying, she seeks to frustrate the Queen's plans and make Wessex notice her.

Who seeks to frustrate the Queen's plans?
Learning Challenge #1

Can we incorporate some priors about language?

One kind of prior - Linguistic Structure

Wessex is chivalrous and charming, but semi-betrothed to Lady Ursula Glynde, whom he has not seen since her infancy. Wessex is repelled by the idea of having his wife thrust upon him and purposely avoids Lady Ursula. Unknown to Wessex, the Queen jealously guards him against Ursula, who is extremely beautiful. As soon as she realizes the Queen is keeping her away from Wessex, Ursula is angered. She believes she loves Wessex, for his nobility and goodness, and she is invested heavily in the betrothal. Although Ursula does not want to lose her independence by marrying, she seeks to frustrate the Queen's plans and make Wessex notice her.

Who seeks to frustrate the Queen's plans?
Learning Challenge #1

- Can we incorporate some priors about language?
- One kind of prior - Linguistic Structure
- Can linguistic structure act as an informative prior?

Wessex is chivalrous and charming, but semi-betrothed to Lady Ursula Glynde, whom he has not seen since her infancy. Wessex is repelled by the idea of having his wife thrust upon him and purposely avoids Lady Ursula. Unknown to Wessex, the Queen jealously guards him against Ursula, who is extremely beautiful. As soon as she realizes the Queen is keeping her away from Wessex, Ursula is angered. She believes she loves Wessex, for his nobility and goodness, and she is invested heavily in the betrothal. Although Ursula does not want to lose her independence by marrying, she seeks to frustrate the Queen's plans and make Wessex notice her.

Who seeks to frustrate the Queen's plans?
Linguistic Structure: Semantics
Linguistic Structure: Semantics

➤ Who did what to whom?
Linguistic Structure: Semantics

Who did what to whom?

After encouraging them, he told them goodbye and left for Macedonia.
Linguistic Structure: Semantics

Who did what to whom?

After **encouraging** them, he **told** them goodbye and **left** for Macedonia

---

PropBank [Palmer et al., 05]
Linguistic Structure: Semantics

Who did what to whom?

After encouraging them, he told them goodbye and left for Macedonia.
Linguistic Structure: Semantics

► Who did what to whom?

After **encouraging** them, he **told** them **goodbye** and **left** for Macedonia

Semantics

Who did what to whom?
After encouraging them, he told them goodbye and left for Macedonia.

Who did what to whom?

Linguistic Structure: Semantics

- ARG0: encourage.02
- ARG1: leave.04
- ARG2: tell.01

PropBank [Palmer et al., 05]
After encouraging them, he told them goodbye and left for Macedonia.

Who did what to whom?

Linguistic Structure: Semantics

PropBank [Palmer et al., 05]
After encouraging them, he told them goodbye and left for Macedonia.
Linguistic Structure: Semantics

Who did what to whom?

- After encouraging them, he told them goodbye and left for Macedonia.
Linguistic Structure: Semantics

Who did what to whom?

This talk: Span-based semantics.

After encouraging them, he told them goodbye and left for Macedonia.
Linguistic Structure: Semantics

▷ Who did what to whom?

▷ This talk: Span-based semantics.

▷ Can span-based semantics serve as a linguistic prior?

PropBank [Palmer et. al., 05]
Linguistic Structure: Semantics

Who did what to whom?

This talk: Span-based semantics.

Can span-based semantics serve as a linguistic prior?

- After encouraging them, he told them goodbye and left for Macedonia.
A Prior for Semantics
A Prior for Semantics

Syntax - a foundation for sentence meaning / semantics
A Prior for Semantics

- Syntax - a foundation for sentence meaning / semantics

After encouraging them, he told them goodbye and left for Macedonia.
A Prior for Semantics

Syntax - a foundation for sentence meaning / semantics

After encouraging them, he told them goodbye and left for Macedonia
A Prior for Semantics

**Syntax** - a foundation for sentence meaning / semantics

**Phrase-based syntax** (node → span)

After encouraging them, he told them goodbye and left for Macedonia.
A Prior for Semantics

Syntax - a foundation for sentence meaning / semantics

Phrase-based syntax (node $\rightarrow$ span)
A Prior for Semantics

Syntax - a foundation for sentence meaning / semantics

Phrase-based syntax (node → span)

Key Intuition: Learn from a complementary structure
Syntactic Scaffolds for Semantic Structures

EMNLP 2018

Structured prediction with an auxiliary structure
Structured prediction with an auxiliary structure

Auxiliary structure: syntax
Structured prediction with an auxiliary structure

Auxiliary structure: syntax
Structured prediction with an auxiliary structure

- Auxiliary structure: **syntax**

- Traditionally a pipeline, both at train and test time [Gildea & Jurafsky, 2002]

Diagram:
- **Input**
- **Auxiliary Structure (Syntax)**
- **Primary Structure (Span-based Semantics)**
Structured prediction with an auxiliary structure

- Auxiliary structure: **syntax**

- Traditionally a pipeline, both at train and test time [Gildea & Jurafsky, 2002]

- More structured data

Primary Structure (Span-based Semantics)

Auxiliary Structure (Syntax)

Input
Structured prediction with an auxiliary structure

- Auxiliary structure: syntax

- Traditionally a pipeline, both at train and test time [Gildea & Jurafsky, 2002]
  - More structured data
  - Cascading errors
Structured prediction with an auxiliary structure

- Auxiliary structure: **syntax**

- Traditionally a pipeline, both at train and test time [Gildea & Jurafsky, 2002]
  - More structured data
  - Cascading errors

- Forsaken in most end-to-end models, but at a cost [He et. al, 2017]
Training Paradigms

Syntax-free training

Syntax for training

Difficulty
Training Paradigms

Syntax-free training

End-to-end modeling
[He et al., 17]

Syntax for training

Syntactic Pipelines
[Gildea & Jurafsky, 02]

Difficulty
Training Paradigms

Syntax-free training
- End-to-end modeling [He et al., 17]

Syntax for training
- Latent variables for syntax [Zettlemoyer & Collins, 05]
- Syntactic Pipelines [Gildea & Jurafsky, 02]
Training Paradigms

Syntax-free training
- End-to-end modeling [He et al., 17]

Syntax for training
- Joint Modeling [Swayamdipta et al., 16]
- Latent variables for syntax [Zettlemoyer & Collins, 05]
- Syntactic Pipelines [Gildea & Jurafsky, 02]

Difficulty
Training Paradigms

Syntax-free training

End-to-end modeling [He et al., 17]

Syntax for training

Multitask Parameter Sharing [Fitzgerald et al., 15]

Joint Modeling [Swayamdipta et al., 16]

Syntactic Pipelines [Gildea & Jurafsky, 02]

Latent variables for syntax [Zettlemoyer & Collins, 05]

Difficulty

FREDA [Daumé III, 2009]
Training Paradigms

Syntax-free training
- End-to-end modeling [He et. al., 17]
- Latent variables for syntax [Zettlemoyer & Collins, 05]

Syntax for training
- Multitask Parameter Sharing [Fitzgerald et. al., 15]
- Joint Modeling [Swayamdipta et. al., 16]
- Syntactic Pipelines [Gildea & Jurafsky, 02]

Syntactic Scaffolds

Difficulty

FREDA [Daumé III, 2009]
Syntactic Scaffolds
Syntactic Scaffolds

Multitask setting

Input
Syntactic Scaffolds

- Multitask setting

- Primary Task → Span-based Semantics

- PropBank Semantic Role Labeling
- Frame-Semantic Role Labeling
- Coreference Resolution

Span-based Semantics

Input
Syntactic Scaffolds

- Multitask setting
- Primary Task → Span-based Semantics
- Scaffold “Task” → Syntax

- PropBank Semantic Role Labeling
- Frame-Semantic Role Labeling
- Coreference Resolution

Input

Span-based Semantics

Syntactic Scaffold
Syntactic Scaffolds

- Multitask setting
- Primary Task $\rightarrow$ Span-based Semantics
- Scaffold “Task” $\rightarrow$ Syntax
  - Full Trees Shallow syntax
Syntactic Scaffolds

- Multitask setting
- Primary Task $\rightarrow$ Span-based Semantics
- Scaffold “Task” $\rightarrow$ Syntax
- Full Trees Shallow syntax
- Soft syntax-aware representations avoid cascaded errors
Syntactic Scaffolds

- Multitask setting
  - Primary Task → Span-based Semantics
    - Scaffold “Task” → Syntax
      - Full Trees Shallow syntax
      - Soft syntax-aware representations avoid cascaded errors

- Not required during test

- PropBank Semantic Role Labeling
- Frame-Semantic Role Labeling
- Coreference Resolution
- Span-based Semantics

Input
Desired parts of syntactic tree:

- After encouraging them, he told them goodbye and left for Macedonia.
Shallow Syntactic Prediction

Desired parts of syntactic tree:

After encouraging them, he told them goodbye and left for Macedonia.
### Shallow Syntactic Prediction

**Desired** parts of syntactic tree:

```
After encouraging them, he told goodbye and left for Macedonia
```

**Span-level classification:** For every span, predict phrase category

\[
\mathcal{L}_2(x, z) = - \sum_{1 \leq i \leq j \leq n} \log p(z_{i:j} | x_{i:j})
\]
Training with syntactic scaffolds

\[ x = \text{Input} \]
\[ y = \text{Output Structure} \]
\[ z = \text{Scaffold Structure} \]
Training with syntactic scaffolds

\[ \sum_{(x,z) \in \mathcal{D}_2} \mathcal{L}_2(x, z; \theta, \psi) \]

\( x = \text{Input} \)
\( y = \text{Output Structure} \)
\( z = \text{Scaffold Structure} \)

Scaffold Task Objective

Scaffold Dataset
Training with syntactic scaffolds

\[ \sum_{(x,y) \in \mathcal{D}_1} \mathcal{L}_1(x, y; \theta, \phi) \]  

Primary Task Objective  
Primary Dataset

\[ \sum_{(x,z) \in \mathcal{D}_2} \mathcal{L}_2(x, z; \theta, \psi) \]  

Scaffold Task Objective  
Scaffold Dataset

\( x = \text{Input} \)  
\( y = \text{Output Structure} \)  
\( z = \text{Scaffold Structure} \)
Training with syntactic scaffolds

\[ \sum_{(x,y) \in \mathcal{D}_1} \mathcal{L}_1(x, y; \theta, \phi) + \delta \sum_{(x,z) \in \mathcal{D}_2} \mathcal{L}_2(x, z; \theta, \psi) \]

- \( x = \text{Input} \)
- \( y = \text{Output Structure} \)
- \( z = \text{Scaffold Structure} \)

Primary Task Objective

Scaffold Task Objective

Primary Dataset

Scaffold Dataset

Mixing Ratio
Training with syntactic scaffolds

\[ \sum_{(x,y) \in D_1} L_1(x, y; \theta, \phi) + \delta \sum_{(x,z) \in D_2} L_2(x, z; \theta, \psi) \]

Primary Dataset
Primary Task Objective

Scaffold Dataset
Scaffold Task Objective

Shared input parameters

x = Input
y = Output Structure
z = Scaffold Structure
The primary objective
The primary objective

Same structures must be scored in both the primary and the scaffold task.
The primary objective

Same structures must be scored in both the primary and the scaffold task.

- Span-based classification, with aggressive pruning [Lee et. al., 2017]
The primary objective

Same structures must be scored in both the primary and the scaffold task.

- Span-based classification, with aggressive pruning [Lee et. al., 2017]

- Semi-Markov Conditional Random Fields [Sarawagi et. al. 2004]
After encouraging them, he told them goodbye and left for Macedonia.
Semi-Markov CRFs

Globally normalized model for segmentations ($s$) of a sentence ($x$).
Semi-Markov CRFs

Globally normalized model for segmentations \((s)\) of a sentence \((x)\).

\[ p(s \mid x) \]
Semi-Markov CRFs

Globally normalized model for segmentations \( (s) \) of a sentence \( (x) \).

Generalization of CRFs:

\[
p(s | x)
\]
Semi-Markov CRFs

<table>
<thead>
<tr>
<th>After</th>
<th>encouraging</th>
<th>them</th>
<th>he</th>
<th>told</th>
<th>them</th>
<th>goodbye</th>
<th>and</th>
<th>left</th>
<th>for</th>
<th>Macedonia</th>
</tr>
</thead>
</table>

- Globally normalized model for segmentations ($s$) of a sentence ($x$).

$$p(s \mid x)$$

- Generalization of CRFs:
  - label and length of an input segment
Semi-Markov CRFs

Globally normalized model for segmentations ($s$) of a sentence ($x$).

Generalization of CRFs:

- label and length of an input segment

$$p(s | x)$$

$$s = \langle i, j, y_{i:j} \rangle$$
Semi-Markov CRFs

After encouraging them he told them goodbye and left for Macedonia

- Globally normalized model for segmentations \((s)\) of a sentence \((x)\).

- Generalization of CRFs:
  - label and length of an input segment

\[
p(s \mid x) \quad \Phi(x, s) = \sum_{k=1}^{m} \phi(s_k, x_{i_k:j_k})
\]

\[
s = \langle i, j, y_{i:j} \rangle
\]
Semi-Markov CRFs

Globally normalized model for segmentations ($s$) of a sentence ($x$).

Generalization of CRFs:

- label and length of an input segment
- Training and inference given by $O(ndl)$ dynamic programs, with a 0th-order Markovian assumption.

$$p(s \mid x)$$

$$s = \langle i, j, y_{i:j} \rangle$$

$$\Phi(x, s) = \sum_{k=1}^{m} \phi(s_k, x_{i_k:j_k})$$

After encouraging them he told them goodbye and left for Macedonia leave.04 ARG2
Model architecture

After encouraging them, he said goodbye and left for Macedonia.
After encouraging them, he said goodbye and left for Macedonia.
Model architecture

After encouraging them, he said goodbye and left for Macedonia.
After encouraging them, he said goodbye and left for Macedonia.
After encouraging them, he said goodbye and left for Macedonia.
After encouraging them, he said goodbye and left for Macedonia.
After encouraging them, he said goodbye and left for Macedonia.
After encouraging them, he said goodbye and left for Macedonia.

Learn scaffold score when syntactic annotations available.
Results
Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Test F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yang &amp; Mitchell, 2017</td>
<td>65.5</td>
</tr>
<tr>
<td>Semi-CRF Baseline</td>
<td>67.0</td>
</tr>
<tr>
<td>NP-PP Scaffold</td>
<td>69.1</td>
</tr>
<tr>
<td>Frame-SRL</td>
<td>67.0</td>
</tr>
</tbody>
</table>
Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Frame-SRL</th>
<th>CoNLL 2012 Span SRL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yang &amp; Mitchell, 2017</td>
<td>69.1</td>
<td>83.8</td>
</tr>
<tr>
<td>Semi-CRF Baseline</td>
<td>67.0</td>
<td>82.1</td>
</tr>
<tr>
<td>NP-PP Scaffold</td>
<td>65.5</td>
<td>82.7</td>
</tr>
<tr>
<td>He et. al., 2017</td>
<td>81.7</td>
<td></td>
</tr>
<tr>
<td>He et. al., 2018</td>
<td>82.1</td>
<td>83.0</td>
</tr>
<tr>
<td>Tan et. al., 2018</td>
<td>82.7</td>
<td></td>
</tr>
<tr>
<td>Semi-CRF Baseline</td>
<td>82.7</td>
<td></td>
</tr>
<tr>
<td>NP-PP Scaffold</td>
<td>83.0</td>
<td></td>
</tr>
</tbody>
</table>
Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Frame-SRL</th>
<th>CoNLL 2012 Span SRL</th>
<th>Coreference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yang &amp; Mitchell, 2017</td>
<td>69.1</td>
<td>83.8</td>
<td></td>
</tr>
<tr>
<td>Semi-CRF Baseline</td>
<td>65.5</td>
<td>81.7</td>
<td>67.2</td>
</tr>
<tr>
<td>NP-PP Scaffold</td>
<td>67.0</td>
<td>82.1</td>
<td>67.8</td>
</tr>
</tbody>
</table>

He et. al., 2017
He et. al., 2018
Tan et. al., 2018
Lee et. al., 2017
Tan et. al., 2018
Semi-CRF Baseline
NP-PP Scaffold
NP Scaffold
Effect of Contextualized Representations

<table>
<thead>
<tr>
<th></th>
<th>Without ELMo</th>
<th>With ELMo</th>
</tr>
</thead>
<tbody>
<tr>
<td>FN 1.5 Test F1</td>
<td>67.0</td>
<td>73.5</td>
</tr>
<tr>
<td>Semi-CRF</td>
<td>69.1</td>
<td>72.8</td>
</tr>
<tr>
<td>NP-PP Scaf.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: These results are not included in the paper.
Recap: Learning Challenge #1

Can linguistic structure act as an informative prior for improving our models?
Recap: Learning Challenge #1

Can linguistic structure act as an informative prior for improving our models?
Recap: Learning Challenge #1

Can linguistic structure act as an informative prior for improving our models?
Recap: Learning Challenge #1

Can linguistic structure act as an informative prior for improving our models?
Recap: Learning Challenge #1

Can linguistic structure act as an informative prior for improving our models?
Looking ahead: Predicted Structure

Sentence ➔ Semantics
Looking ahead: Predicted Structure
Looking ahead: Predicted Structure

- Syntax
- Semantics
- Sentence
- Downstream Applications e.g. Reading Comprehension
Looking ahead: Predicted Structure

- Syntax
- Semantics
- Downstream Applications e.g. Reading Comprehension

Sentence Learning
Looking ahead:
Structured Transformation

Input → Syntax → Semantics
Looking ahead: Structured Transformation

Iyyer et al. [NAACL 2018]
Looking ahead:
Structured Transformation

- Input
- Syntax
- Semantics
- Transformed Input

Iyyer et. al. [NAACL 2018]
Looking ahead: Structured Transformation

Iyyer et. al. [NAACL 2018]
Part II
Recap:
Confusion of the Muppets

Wessex is chivalrous and charming, but semi-betrothed to Lady Ursula Glynde, whom he has not seen since her infancy. Wessex is repelled by the idea of having his wife thrust upon him and purposely avoids Lady Ursula. Unknown to Wessex, the Queen jealously guards him against Ursula, who is extremely beautiful. As soon as she realizes the Queen is keeping her away from Wessex, Ursula is angered. She believes she loves Wessex, for his nobility and goodness, and she is invested heavily in the betrothal. Although Ursula does not want to lose her independence by marrying, she seeks to frustrate the Queen's plans and make Wessex notice her.

Who seeks to frustrate the Queen's plans? Wessex
# Learning Challenges

<table>
<thead>
<tr>
<th>Part I</th>
<th>Part II</th>
</tr>
</thead>
<tbody>
<tr>
<td>Can linguistic structure act as an informative prior for improving our models?</td>
<td>What in our data is causing models to achieve high performance?</td>
</tr>
<tr>
<td>✓ Syntactic Scaffolds for Semantic Structures (EMNLP 2018)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>☐ Annotation Artifacts in Natural Language Inference Data (NAACL 2018)</td>
</tr>
</tbody>
</table>
Annotation Artifacts in Natural Language Inference Data

NAACL 2018

Suchin Gururangan*
S.*
Omer Levy
Roy Schwartz
Sam Bowman
Noah A. Smith

*equal contribution
Natural Language Inference (NLI)

Given a premise, is a hypothesis true, false or neither?

**Premise**
Two dogs are running through a field.

**Hypothesis**
The pets are sitting on a couch.

- **True** $\rightarrow$ **Entailment**
- **False** $\rightarrow$ **Contradiction**
- **Cannot Say** $\rightarrow$ **Neutral**
NLI Datasets

**Stanford NLI** [Bowman et. al, 2015]  570 K

**Multi-genre NLI** [Williams et. al., 2017]  433 K
NLI Datasets

Two dogs are running through a field.

Premise

Stanford NLI [Bowman et. al, 2015] 570 K
Multi-genre NLI [Williams et. al., 2017] 433 K
NLI Datasets

Two dogs are running through a field.

Premise

Stanford NLI [Bowman et. al, 2015] 570 K
Multi-genre NLI [Williams et. al., 2017] 433 K
NLI Datasets

Two dogs are running through a field.

Premise

Stanford NLI [Bowman et. al, 2015] 570 K
Multi-genre NLI [Williams et. al., 2017] 433 K
NLI Datasets

Stanford NLI [Bowman et. al, 2015]  570 K
Multi-genre NLI [Williams et. al., 2017]  433 K

Premise

Two dogs are running through a field.

Entailment

There are animals outdoors.

There are animals outdoors.
NLI Datasets

**Stanford NLI** [Bowman et. al, 2015]  570 K
**Multi-genre NLI** [Williams et. al., 2017]  433 K

---

There are animals outdoors.

Entailment

Neutral

There are animals outdoors.

Some puppies are running to catch a stick.

Two dogs are running through a field.

Premise
NLI Datasets

Stanford NLI [Bowman et. al, 2015] 570 K
Multi-genre NLI [Williams et. al., 2017] 433 K

Premise

Entailment

Neutral

Contradiction

There are animals outdoors.

Some puppies are running to catch a stick.

The pets are sitting on a couch.

Two dogs are running through a field.
Lots of progress

<table>
<thead>
<tr>
<th>Publication</th>
<th>Model</th>
<th>Parameters</th>
<th>Train (% acc)</th>
<th>Test (% acc)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Feature-based models</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bowman et al. '15</td>
<td>Unlexicalized features</td>
<td></td>
<td>49.4</td>
<td>50.4</td>
</tr>
<tr>
<td>Bowman et al. '15</td>
<td>+ Unigram and bigram features</td>
<td></td>
<td>99.7</td>
<td>78.2</td>
</tr>
<tr>
<td>Peters et al. '18</td>
<td>ESIM + ELMo</td>
<td>8.0m</td>
<td>91.6</td>
<td>88.7</td>
</tr>
<tr>
<td>Boyuan Pan et al. '18</td>
<td>300D DMAN</td>
<td>9.2m</td>
<td>95.4</td>
<td>88.8</td>
</tr>
<tr>
<td>Zhiguo Wang et al. '17</td>
<td>BIMPM <strong>Ensemble</strong></td>
<td>6.4m</td>
<td>93.2</td>
<td>88.8</td>
</tr>
<tr>
<td>Yichen Gong et al. '17</td>
<td>448D Densely Interactive Inference Network (DIIN, code) <strong>Ensemble</strong></td>
<td>17m</td>
<td>92.3</td>
<td>88.9</td>
</tr>
<tr>
<td>Seonhoon Kim et al. '18</td>
<td>Densely-Connected Recurrent and Co-Attentive Network</td>
<td>6.7m</td>
<td>93.1</td>
<td>88.9</td>
</tr>
<tr>
<td>Zhuosheng Zhang et al. '18</td>
<td>SLRC</td>
<td>6.1m</td>
<td>89.1</td>
<td>89.1</td>
</tr>
<tr>
<td>Qian Chen et al. '17</td>
<td>KIM <strong>Ensemble</strong></td>
<td>43m</td>
<td>93.6</td>
<td>89.1</td>
</tr>
<tr>
<td>Ghaeini et al. '18</td>
<td>450D DR-BILSTM <strong>Ensemble</strong></td>
<td>45m</td>
<td>94.8</td>
<td>89.3</td>
</tr>
<tr>
<td>Peters et al. '18</td>
<td>ESIM + ELMo <strong>Ensemble</strong></td>
<td>40m</td>
<td>92.1</td>
<td>89.3</td>
</tr>
<tr>
<td>Yi Tay et al. '18</td>
<td>300D CAFE <strong>Ensemble</strong></td>
<td>17.5m</td>
<td>92.5</td>
<td>89.3</td>
</tr>
<tr>
<td>Chuanqi Tan et al. '18</td>
<td>150D Multiway Attention Network <strong>Ensemble</strong></td>
<td>58m</td>
<td>95.5</td>
<td>89.4</td>
</tr>
<tr>
<td>Boyuan Pan et al. '18</td>
<td>300D DMAN <strong>Ensemble</strong></td>
<td>79m</td>
<td>96.1</td>
<td>89.6</td>
</tr>
<tr>
<td>Radford et al. '18</td>
<td>Fine-Tuned LM-Pretrained Transformer</td>
<td>85m</td>
<td>96.6</td>
<td><strong>89.9</strong></td>
</tr>
<tr>
<td>Seonhoon Kim et al. '18</td>
<td>Densely-Connected Recurrent and Co-Attentive Network <strong>Ensemble</strong></td>
<td>53.3m</td>
<td>95.0</td>
<td><strong>90.1</strong></td>
</tr>
</tbody>
</table>
NLI as Text Classification

Premise:
Two dogs are running through a field.

Hypothesis:
The pets are sitting on a couch.
A simple experiment
A simple experiment

Premise

Hypothesis

fastText [Joulin et. al. 2017]
Performance of hypothesis-only

Over 50% of NLI examples can be correctly classified **without** ever observing the premise

[Polik et. al., 2018, Glockner et. al., 2018]
Can we filter out examples with artifacts?
Can we filter out examples with artifacts?

Premise

Hypothesis

E

N

C

Easy

Hard
Revisiting NLI models

**DAM** - Decomposable Attention Model [Parikh et. al. 2016]

**ESIM** - Enhanced Sequential Inference Model [Chen et. al., 2017]

**DIIN** - Densely Interactive Inference Network [Gong et. al. 2018]
Revisiting NLI models

**MultiNLI Mismatched**

- **DAM**: 72.1
- **ESIM**: 73.1
- **DIIN**: 76.6

**MultiNLI Matched**

- **DAM**: 72.0
- **ESIM**: 74.1
- **DIIN**: 77.0

**DAM** - Decomposable Attention Model [Parikh et al. 2016]

**ESIM** - Enhanced Sequential Inference Model [Chen et al., 2017]

**DIIN** - Densely Interactive Inference Network [Gong et al. 2018]
Revisiting NLI models

MultiNLI Mismatched

MultiNLI Matched

DAM - Decomposable Attention Model [Parikh et. al. 2016]
ESIM - Enhanced Sequential Inference Model [Chen et. al., 2017]
DIIN - Densely Interactive Inference Network [Gong et. al. 2018]
Revisiting NLI models

**MultiNLI**

Mismatched

Mismatched

<table>
<thead>
<tr>
<th>Model</th>
<th>Full</th>
<th>Hard</th>
<th>Easy</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAM</td>
<td>72.1</td>
<td>66.2</td>
<td>85.7</td>
</tr>
<tr>
<td>ESIM</td>
<td>85.7</td>
<td>85.2</td>
<td>58.9</td>
</tr>
<tr>
<td>DIIN</td>
<td>76.6</td>
<td>64.4</td>
<td>86.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Full</th>
<th>Hard</th>
<th>Easy</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAM</td>
<td>87.6</td>
<td>75.8</td>
<td>66.2</td>
</tr>
<tr>
<td>ESIM</td>
<td>85.3</td>
<td>89.3</td>
<td>64.1</td>
</tr>
<tr>
<td>DIIN</td>
<td>77.0</td>
<td>74.1</td>
<td>64.1</td>
</tr>
</tbody>
</table>

**DAM** - Decomposable Attention Model [Parikh et al. 2016]

**ESIM** - Enhanced Sequential Inference Model [Chen et al., 2017]

**DIIN** - Densely Interactive Inference Network [Gong et al. 2018]
Artifacts by NLI Class
Some men and boys are playing frisbee in a grassy area.

Premise

Generalization

People play frisbee outdoors.

Entailment

Hypothesis
Artifacts by NLI Class

Premise

Some men and boys are playing frisbee in a grassy area.

A middle-aged man works under the engine of a train on rail tracks.

Generalization

People play frisbee outdoors.

Entailment Hypothesis

A man is doing work on a black Amtrak train.

Modifiers

Neutral Hypothesis
Artifacts by NLI Class

Generalization

Premise: Some men and boys are playing frisbee in a grassy area.

Entailment Hypothesis: People play frisbee outdoors.

Modifiers

Premise: A middle-aged man works under the engine of a train on rail tracks.

Neutral Hypothesis: A man is doing work on a black Amtrak train.

Cats!

Premise: Three dogs racing on racetrack.

Contradiction Hypothesis: Three cats race on a track.
Two dogs are running through a field.

Premise

There are animals outdoors.
Entailment

Some puppies are running to catch a stick.
Neutral

The pets are sitting on a couch.
Contradiction
Two dogs are running through a field.

Premise

There are animals outdoors.

Entailment

Some puppies are running to catch a stick.

Neutral

The pets are sitting on a couch.

Contradiction
Can we filter out examples with artifacts?
Can we filter out examples with artifacts?

- Hard examples exhibit their own artifacts!
Can we filter out examples with artifacts?

- Hard examples exhibit their own artifacts!
- Artifacts are still valid examples...
Looking ahead: Learning from Datasets with Artifacts
Looking ahead: Learning from Datasets with Artifacts

- Intuition: Models which exploit artifacts == models which can detect artifacts

- Stylistic global features
Looking ahead: Learning from Datasets with Artifacts

- Intuition: Models which exploit artifacts == models which can detect artifacts

- Stylistic global features

- Subsampling large datasets → weight each example based on how representative it could be [Coleman et. al., 2018]
Looking ahead: Learning from Datasets with Artifacts

- Intuition: Models which exploit artifacts == models which can detect artifacts

- Stylistic global features

- Subsampling large datasets $\rightarrow$ weight each example based on how representative it could be [Coleman et. al., 2018]

Easy

Hard
Looking Ahead: Improved Data Collection
Looking Ahead: Improved Data Collection

- Partial input baselines. E.g. SWAG [Zellers et. al., 2018], DROP [Dua et. al., 2019], Diverse NLI [Poliak et. al., 2018]
Looking Ahead: Improved Data Collection

- Partial input baselines. E.g. SWAG [Zellers et. al., 2018], DROP [Dua et. al., 2019], Diverse NLI [Poliak et. al., 2018]

- Alternatives to human elicitation for building datasets?
Looking Ahead: Improved Data Collection

- Partial input baselines. E.g. SWAG [Zellers et. al., 2018], DROP [Dua et. al., 2019], Diverse NLI [Poliak et. al., 2018]

- Alternatives to human elicitation for building datasets?
In conclusion:
It’s an exciting time for NLP!
In conclusion:
It’s an exciting time for NLP!

Finally, a Machine That Can Finish Your Sentence

Completing someone else’s thought is not an easy trick for A.I. But new systems are starting to crack the code of natural language.
In conclusion - Learning Challenges

Part I

Can linguistic structure act as an informative prior to improve our models?

Predicted structure can help representation learning.

Part II

What in our data is causing models to achieve high performance?

Need models robust to artifacts.
Thanks!

http://www.cs.cmu.edu/~sswayamd

swabhs  swabhz