Natural Language Inference: Challenges and Opportunities

Vered Shwartz

Natural Logic Meets Machine Learning, June 2021
Microsoft DeBERTa surpasses human performance on... of NLU tasks, including question answering, natural language inference, coreference resolution, word sense disambiguation, and others.
Jan 6, 2021

Facebook Inch Closer To General-Purpose Intelligence...
The Facebook AI researchers started with the question: "Could a Transformer trained for natural language inference on textual input also...
Mar 3, 2021
Microsoft DeBERTa surpasses human performance on ... of NLU tasks, including question answering, natural language inference, co-reference resolution, word sense disambiguation, and others.

Jan 6, 2021

Facebook Inch Closer To General-Purpose Intelligence ...
The Facebook AI researchers started with the question: "Could a Transformer trained for natural language inference on textual input also ...

Mar 3, 2021

Human Performance:

- Facebook: 84.6
- Google AI: 89.3
- Microsoft: 90.3

SuperGLUE

Jul '19 Aug '19 Jan '20 Jan '21
Microsoft DeBERTa surpasses human performance on NLU tasks, including question answering, natural language inference, coreference resolution, word sense disambiguation, and others.

Jan 6, 2021

Facebook Inch Closer To General-Purpose Intelligence

The Facebook AI researchers started with the question: “Could a Transformer trained for natural language inference on textual input also...

Mar 3, 2021

Human Performance:

<table>
<thead>
<tr>
<th>Date</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jul '19</td>
<td>84.6</td>
</tr>
<tr>
<td>Aug '19</td>
<td>89.3</td>
</tr>
<tr>
<td>Jan '20</td>
<td>90.2</td>
</tr>
<tr>
<td>Jan '21</td>
<td>90.3</td>
</tr>
</tbody>
</table>

Is NLI (nearly) solved?
I baked a chocolate cake but accidentally used regular instead of self-rising flour.
I baked a chocolate cake but accidentally used regular instead of self-rising flour.

I've made a mistake.
I baked a chocolate cake but accidentally used regular instead of self-rising flour.

The cake was difficult to cut.

I baked a bad chocolate cake.
I baked a chocolate cake but accidentally used regular instead of self-rising flour.

I've made a mistake.

The cake turned out flat.
The cake was difficult to cut.
I baked a bad chocolate cake.

The cake was soft and tasty.
I baked a chocolate cake but accidentally used regular instead of self-rising flour.

I've made a mistake.

The cake turned out flat.
The cake was difficult to cut.
I baked a bad chocolate cake.

The cake was soft and tasty.

I've baked a chocolate rock.
Outline

Limitations of neural NLI Models

Incorporating symbolic knowledge into neural NLI models

NLI is too easy? What’s next?
Limitations of neural NLI Models

Incorporating symbolic knowledge into neural NLI models

NLI is too easy? What’s next?
#1 Learning Paradigm

**Dataset**

![Dataset Image]

**Labels**

parrot, duck, duck, parrot...
#1 Learning Paradigm

**Dataset**

**Labels**

parrot, duck, duck, parrot...
#1 Learning Paradigm

**Dataset**

**Labels**
parrot, duck, duck, parrot...

**Training set**

**Labels**
parrot, duck, duck, parrot...

**Classifier**
#1 Learning Paradigm

Dataset

Labels
parrot, duck, duck, parrot...

Training set

Labels
parrot, duck, duck, parrot...

Classifier

Test set

Labels
parrot, duck, duck, parrot...
#1 Learning Paradigm

**Dataset**

**Labels**
parrot, duck, duck, parrot...

**Training set**

**Labels**
parrot, duck, duck, parrot...

**Classifier**

duck, parrot, duck, parrot...

**Test set**

**Labels**
parrot, duck, duck, parrot...
#1 Learning Paradigm

**Dataset**

- **Labels**
  - parrot, duck, duck, parrot...

**Training set**

- **Labels**
  - parrot, duck, duck, parrot...

**Test set**

- **Labels**
  - parrot, duck, duck, parrot...

**Classifier**

- duck, parrot, duck, parrot...

100% accuracy
#1 Learning Paradigm

Dataset

Training set

Labels
parrot, duck, duck, parrot...

Labels
parrot, duck, duck, parrot...

Classifier
#1 Learning Paradigm

Dataset

Labels
parrot, duck, duck, parrot...

Training set

Labels
parrot, duck, duck, parrot...

Yellow head → parrot

Classifier
#1 Learning Paradigm

**Dataset**

**Labels**
parrot, duck, duck, parrot...

**Training set**

**Labels**
parrot, duck, duck, parrot...

Yellow head → parrot

Classifier → parrot
#1 Learning Paradigm

**Dataset**

**Labels**
- parrot, duck, duck, parrot...

**Training set**

**Labels**
- parrot, duck, duck, parrot...

Green body $\rightarrow$ parrot, Gray body $\rightarrow$ duck

**Classifier**
#1 Learning Paradigm

**Dataset**

Labels: parrot, duck, duck, parrot...

**Training set**

Labels: parrot, duck, duck, parrot...

Green body → parrot,
Gray body → duck

Classifier → duck
#1 Learning Paradigm

**Dataset**

Labels: parrot, duck, duck, parrot...

**Training set**

Labels: parrot, duck, duck, parrot...

**Classifier**
#1 Learning Paradigm

**Dataset**

- Labels: parrot, duck, duck, parrot...

**Training set**

- Labels: parrot, duck, duck, parrot...

**Classifier**
#1 Learning Paradigm

1. Spurious correlations

Dataset

Labels: parrot, duck, duck, parrot...

Training set

Labels: parrot, duck, duck, parrot...

Classifier
#1 Learning Paradigm

1. Spurious correlations
2. Train & test from the same distribution
Annotation artifacts, hypothesis-only baseline  
(Gururangan, Swayamdipta, et al., 2018; Poliak et al., 2018; Tsuchiya, 2018)
#1 Learning Paradigm

Annotation artifacts, hypothesis-only baseline
(Gururangan, Swayamdipta, et al., 2018; Poliak et al., 2018; Tsuchiya, 2018)

**p:** I only had a soup but it was very filling.

**h:** I didn't eat a salad.

🤖: contradiction (98.2%)
#1 Learning Paradigm

Annotation artifacts, hypothesis-only baseline
(Gururangan, Swayamdipta, et al., 2018; Poliak et al., 2018; Tsuchiya, 2018)

\[ \text{p: I only had a soup but it was very filling.} \]
\[ \text{h: I didn't eat a salad.} \]
\[ \text{🤖: contradiction (98.2%)} \]

NLI models rely on syntactic heuristics
(McCoy et al., 2019)
#2 Representations

Relatedness
Relatedness → Finer-grained Semantic Relations
#2 Representations

**P:** Charlie will visit his mother in London on **Wednesday** evening.

**H:** Charlie will visit his mother in London on **Thursday** evening.
#2 Representations

2018

(ELMo-based Decomposable Attention)

P: Charlie will visit his mother in London on Wednesday evening.
H: Charlie will visit his mother in London on Thursday evening.

https://demo.allennlp.org/textual-entailment
#2 Representations

2018
(ELMo-based Decomposable Attention)

**P:** Charlie will visit his mother in London on **Wednesday** evening.

**H:** Charlie will visit his mother in London on **Thursday** evening.

Prediction: Entailment (94.1%) ✗
P: Charlie will visit his mother in London on **Wednesday** evening.

H: Charlie will visit his mother in London on **Thursday** evening.

Prediction: Entailment (94.1%) ✗
#2 Representations

2018

(ELMo-based Decomposable Attention)

**P:** Charlie will visit his mother in London on **Wednesday** evening.

**H:** Charlie will visit his mother in London on **Thursday** evening.

Prediction: Entailment (94.1%) ✗

- Errors: similar but mutually-exclusive words
- Accuracy increases with frequency in training set → Limited generalization ability!
- Similar findings in “NLI stress tests” (Naik et al., 2018)
#2 Representations

2021

(RoBERTa)

https://demo.allennlp.org/textual-entailment
#2 Representations

2021
(RoBERTa)

**P:** Charlie will visit his mother in London on *Wednesday* evening.

**H:** Charlie will visit his mother in London on *Thursday* evening.

Prediction: Contradiction (73.2%) ✓
#2 Representations

2021
(RoBERTa)

**P:** Charlie will visit his mother in London on **Wednesday** evening.

**H:** Charlie will visit his mother in London on **Thursday** evening.

Prediction: Contradiction (73.2%) ✓

**P:** Charlie said on **Wednesday** that he is busy on **Thursday** so he will visit his mother next week.

**H:** Charlie said on **Thursday** that he is busy on **Wednesday** so he will visit his mother next week.

Prediction: Entailment (92%) ✗
The definition of the textual entailment recognition task, like that of any other text understanding task, refers to human understanding of language. Such definition necessarily assumes common background knowledge, on which the (human) entailment judgment relies. […] this knowledge should cover both extra-linguistic world knowledge […] as well as knowledge of the language itself.
#3 World Knowledge

Knowledge in Language Models

Capture facts not explicitly mentioned in the corpus
(Petroni et al. 2019; Feldman et al. 2019, this talk)
#3 World Knowledge

Knowledge in Language Models

- Capture facts not explicitly mentioned in the corpus (Petroni et al. 2019; Feldman et al. 2019, this talk)
- Associate concepts with their properties (Weir et al. 2020)

<table>
<thead>
<tr>
<th>Context</th>
<th>Human Response</th>
<th>PF</th>
<th>RoBERTa-L Response</th>
<th>PLM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Everyone knows that a bear has ____</td>
<td>fur 27</td>
<td>teeth .36</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>claws 15</td>
<td>claws .18</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>teeth 11</td>
<td>eyes .05</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>cubs 7</td>
<td>ears .03</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>paws 7</td>
<td>horns .02</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
#3 World Knowledge

Knowledge in Language Models

✅ Capture facts not explicitly mentioned in the corpus (Petroni et al. 2019; Feldman et al. 2019, this talk)

✅ Associate concepts with their properties (Weir et al. 2020)

<table>
<thead>
<tr>
<th>Context</th>
<th>Human Response</th>
<th>PF</th>
<th>ROBERTA-L Response</th>
<th>PLM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Everyone knows that a bear has ____</td>
<td>fur 27</td>
<td>teeth .36</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>claws 15</td>
<td>claws .18</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>teeth 11</td>
<td>eyes .05</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>cubs 7</td>
<td>ears .03</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>paws 7</td>
<td>horns .02</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
#3 World Knowledge

Knowledge in Language Models

- Not sensitive to negation (Kassner et al. 2020; Ettinger, 2020)

- Capture facts not explicitly mentioned in the corpus (Petroni et al. 2019; Feldman et al. 2019, this talk)

- Associate concepts with their properties (Weir et al. 2020)

<table>
<thead>
<tr>
<th>Context</th>
<th>Human Response</th>
<th>PF</th>
<th>ROBERTA-L Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Everyone knows that a bear has ____</td>
<td>fur 27</td>
<td>teeth .36</td>
<td></td>
</tr>
<tr>
<td></td>
<td>claws 15</td>
<td>claws .18</td>
<td></td>
</tr>
<tr>
<td></td>
<td>teeth 11</td>
<td>eyes .05</td>
<td></td>
</tr>
<tr>
<td></td>
<td>cubs 7</td>
<td>ears .03</td>
<td></td>
</tr>
<tr>
<td></td>
<td>paws 7</td>
<td>horns .02</td>
<td></td>
</tr>
</tbody>
</table>
#3 World Knowledge

Knowledge in Language Models

- Capture facts not explicitly mentioned in the corpus (Petroni et al. 2019; Feldman et al. 2019, this talk)
- Associate concepts with their properties (Weir et al. 2020)
- Not sensitive to negation (Kassner et al. 2020; Ettinger, 2020)
- Predict similar but mutually-exclusive facts (Jiang et al., 2020)

DirectX is developed by

<table>
<thead>
<tr>
<th>Rank</th>
<th>Company</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Intel</td>
<td>-1.06</td>
</tr>
<tr>
<td>2</td>
<td>Microsoft</td>
<td>-2.21</td>
</tr>
<tr>
<td>3</td>
<td>IBM</td>
<td>-2.76</td>
</tr>
<tr>
<td>4</td>
<td>Google</td>
<td>-3.40</td>
</tr>
<tr>
<td>5</td>
<td>Nokia</td>
<td>-3.58</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Context</th>
<th>Human Response</th>
<th>PF</th>
<th>ROBERTA-L Response</th>
<th>\textit{PLM}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Everyone knows that a bear has</td>
<td>fur 27</td>
<td>teeth .36</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>claws 15</td>
<td>claws .18</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>teeth 11</td>
<td>eyes .05</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>cubs 7</td>
<td>ears .03</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>paws 7</td>
<td>horns .02</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
#3 World Knowledge

## Knowledge in Language Models

- **Capture facts not explicitly mentioned in the corpus** (Petroni et al. 2019; Feldman et al. 2019, this talk)
- **Associate concepts with their properties** (Weir et al. 2020)

### Limitations

- **Not sensitive to negation** (Kassner et al. 2020; Ettinger, 2020)
- **Predict similar but mutually-exclusive facts** (Jiang et al., 2020)
- **Lack perceptual and physical knowledge** (Forbes et al. 2019, Weir et al., 2020, Bisk et al. 2020)

<table>
<thead>
<tr>
<th>Context</th>
<th>Human Response</th>
<th>PF</th>
<th>ROBERTA-L Response PLM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Everyone knows that a bear has ____</td>
<td>fur 27</td>
<td>teeth .36</td>
<td></td>
</tr>
<tr>
<td></td>
<td>claws 15</td>
<td>claws .18</td>
<td></td>
</tr>
<tr>
<td></td>
<td>teeth 11</td>
<td>eyes .05</td>
<td></td>
</tr>
<tr>
<td></td>
<td>cubs 7</td>
<td>ears .03</td>
<td></td>
</tr>
<tr>
<td></td>
<td>paws 7</td>
<td>horns .02</td>
<td></td>
</tr>
</tbody>
</table>

DirectX is developed by

1. Intel -1.06
2. Microsoft -2.21
3. IBM -2.76
4. Google -3.40
5. Nokia -3.58
#3 World Knowledge

**Knowledge in Language Models**

- **Capture facts not explicitly mentioned in the corpus**
  (Petroni et al. 2019; Feldman et al. 2019, this talk)

- **Associate concepts with their properties**
  (Weir et al. 2020)

- **Not sensitive to negation** (Kassner et al. 2020; Ettinger, 2020)

- **Predict similar but mutually-exclusive facts**
  (Jiang et al., 2020)

- **Lack perceptual and physical knowledge**
  (Forbes et al. 2019, Weir et al., 2020, Bisk et al. 2020)

- **Don’t differentiate constant vs. contingent facts**

<table>
<thead>
<tr>
<th>Context</th>
<th>Human Response</th>
<th>PF</th>
<th>ROBERTA-L Response</th>
<th>PLM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Everyone knows that a bear has ____</td>
<td>fur 27</td>
<td>teeth .36</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>claws 15</td>
<td>claws .18</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>teeth 11</td>
<td>eyes .05</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>cubs 7</td>
<td>ears .03</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>paws 7</td>
<td>horns .02</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

DirectX is developed by

<table>
<thead>
<tr>
<th>Company</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intel</td>
<td>-1.06</td>
</tr>
<tr>
<td>Microsoft</td>
<td>-2.21</td>
</tr>
<tr>
<td>IBM</td>
<td>-2.76</td>
</tr>
<tr>
<td>Google</td>
<td>-3.40</td>
</tr>
<tr>
<td>Nokia</td>
<td>-3.58</td>
</tr>
</tbody>
</table>

Zebras are black and white.

My shirt is blue / red.
#3 World Knowledge

Knowledge in Language Models

- Capture facts not explicitly mentioned in the corpus ([Petroni et al. 2019; Feldman et al. 2019, this talk](#)).
- Associate concepts with their properties ([Weir et al. 2020](#)).

<table>
<thead>
<tr>
<th>Context</th>
<th>Human Response</th>
<th>PF</th>
<th>ROBERTA-L Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Everyone knows that a bear has</td>
<td>fur 27</td>
<td></td>
<td>teeth .36</td>
</tr>
<tr>
<td></td>
<td>claws 15</td>
<td></td>
<td>claws .18</td>
</tr>
<tr>
<td></td>
<td>teeth 11</td>
<td></td>
<td>eyes .05</td>
</tr>
<tr>
<td></td>
<td>cubs 7</td>
<td></td>
<td>ears .03</td>
</tr>
<tr>
<td></td>
<td>paws 7</td>
<td></td>
<td>horns .02</td>
</tr>
</tbody>
</table>

- Not sensitive to negation ([Kassner et al. 2020; Ettinger, 2020](#)).
- Predict similar but mutually-exclusive facts ([Jiang et al., 2020](#)).
- Lack perceptual and physical knowledge ([Forbes et al. 2019, Weir et al., 2020, Bisk et al. 2020](#)).
- Don’t differentiate constant vs. contingent facts.
- Reporting bias.

 DirectX is developed by:

1. Intel -1.06
2. Microsoft -2.21
3. IBM -2.76
4. Google -3.40
5. Nokia -3.58

Zebras are black and white.

My shirt is blue / red.

#3 World Knowledge

Knowledge in Language Models

- Capture facts not explicitly mentioned in the corpus (Petroni et al. 2019; Feldman et al. 2019, this talk)
- Associate concepts with their properties (Weir et al. 2020)
- Not sensitive to negation (Kassner et al. 2020; Ettinger, 2020)
- Predict similar but mutually-exclusive facts (Jiang et al., 2020)
- Lack perceptual and physical knowledge (Forbes et al. 2019, Weir et al., 2020, Bisk et al. 2020)
- Don’t differentiate constant vs. contingent facts
- Reporting bias

<table>
<thead>
<tr>
<th>Context</th>
<th>Human Response</th>
<th>PF</th>
<th>ROBERTA-L Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Everyone knows that a bear has</td>
<td>fur</td>
<td>27</td>
<td>teeth</td>
</tr>
<tr>
<td></td>
<td>claws</td>
<td>15</td>
<td>claws</td>
</tr>
<tr>
<td></td>
<td>teeth</td>
<td>11</td>
<td>eyes</td>
</tr>
<tr>
<td></td>
<td>cubs</td>
<td>7</td>
<td>ears</td>
</tr>
<tr>
<td></td>
<td>paws</td>
<td>7</td>
<td>horns</td>
</tr>
</tbody>
</table>

DirectX is developed by

- Intel -1.06
- Microsoft -2.21
- IBM -2.76
- Google -3.40
- Nokia -3.58

Zebras are black and white.

My shirt is blue / red.

The man turned on the faucet. As a result,
#3 World Knowledge

Knowledge in Language Models

- Capture facts not explicitly mentioned in the corpus (Petroni et al. 2019; Feldman et al. 2019, this talk)
- Associate concepts with their properties (Weir et al. 2020)

<table>
<thead>
<tr>
<th>Context</th>
<th>Human Response</th>
<th>PF</th>
<th>RoBERTa-L Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Everyone knows that a bear has ____</td>
<td>fur 27</td>
<td>teeth .36</td>
<td></td>
</tr>
<tr>
<td></td>
<td>claws 15</td>
<td>claws .18</td>
<td></td>
</tr>
<tr>
<td></td>
<td>teeth 11</td>
<td>eyes .05</td>
<td></td>
</tr>
<tr>
<td></td>
<td>cubs 7</td>
<td>ears .03</td>
<td></td>
</tr>
<tr>
<td></td>
<td>paws 7</td>
<td>horns .02</td>
<td></td>
</tr>
</tbody>
</table>

- Not sensitive to negation (Kassner et al. 2020; Ettinger, 2020)
- Predict similar but mutually-exclusive facts (Jiang et al., 2020)
- Lack perceptual and physical knowledge (Forbes et al. 2019, Weir et al., 2020, Bisk et al. 2020)
- Don’t differentiate constant vs. contingent facts
- Reporting bias

DirectX is developed by Intel -1.06, Microsoft -2.21, IBM -2.76, Google -3.40, Nokia -3.58

The man turned on the faucet. As a result, the man’s blood was sprayed everywhere.

Zebras are black and white. My shirt is blue / red.

Incorporating symbolic knowledge into neural NLI models

Limitations of neural NLI Models

NLI is too easy? What’s next?
The office is located in Baytown [SEP]
The office is located in Texas

**label: entailment**
Learning Semantic Phenomena

The office is located in Baytown [SEP] The office is located in Texas

label: entailment

Model expected to learn:

Fact: LocatedIn(Baytown, Texas)
Learning Semantic Phenomena

The office is located in Baytown [SEP] Texas

label: entailment

Model expected to learn:

Fact: LocatedIn(Baytown, Texas)

Rule: Entailment between a city and its state in upward monotone sentences
Probing and Inoculation
Does the model know the location of cities?

The office is located in Baytown [SEP]
The office is located in Texas

Probing and Inoculation
Does the model know the location of cities?

The office is located in Baytown [SEP]
The office is located in Texas

NLI Model

Contradiction
Does the model know the location of cities?

The office is located in Baytown [SEP]
The office is located in Texas

Probing and Inoculation

Probing and Inoculation

Does the model know the location of cities?

The office is located in Baytown [SEP]
The office is located in Texas

Contradiction

The office is located in London [SEP]
She lives in Dallas [SEP]
He moved to Chicago [SEP]
He lives in Illinois

label: entailment


Blindspot in the original NLI dataset
Inherent model limitation
Does the model know the location of cities?

The office is located in Baytown [SEP]
The office is located in Texas

The office is located in London [SEP]
She lives in Dallas [SEP]
He moved to Chicago [SEP]
He lives in Illinois

I got my locatedIn VACCINE!

Contradiction

Blindspot in the original NLI dataset
Inherent model limitation

Probing and Inoculation

Does the model know the location of cities?

The office is located in Baytown [SEP]
The office is located in Texas

The office is located in London [SEP]

She lives in Dallas [SEP]
He moved to Chicago [SEP]
He lives in Illinois

label: entailment

Contradiction

Possible Outcomes

Does the model know the location of cities?

The office is located in Baytown [SEP]
The office is located in Texas

The office is located in London [SEP]

She lives in Dallas [SEP]
He lives in Dallas [SEP]

He moved to Chicago [SEP]
He lives in Illinois

label: entailment

Possible Outcomes

Success

Blindspot in the original NLI dataset

Does the model know the location of cities?

The office is located in Baytown [SEP]
The office is located in Texas

The office is located in London [SEP]
She lives in Dallas [SEP]
He moved to Chicago [SEP]
He lives in Illinois

label: entailment

Contradiction

Possible Outcomes

Success
Blindspot in the original NLI dataset

Failure
Inherent model limitation

Probing and **Inoculation**

Does the model know the location of cities?

- The office is located in **Baytown** [SEP]
- The office is located in **Texas**

**Entailment**

- The office is located in **London** [SEP]
- She lives in **Dallas** [SEP]
- He moved to **Chicago** [SEP]
- He lives in **Illinois**

*label: entailment*

**Possible Outcomes**

- **Success**
- **Failure**

---

Probing and **Inoculation**

Does the model know the location of cities?

The office is located in **Baytown** [SEP]
The office is located in **Texas**

The office is located in **London** [SEP]

She lives in **Dallas** [SEP]

He moved to **Chicago** [SEP]

He lives in **Illinois** [SEP]

Has the NLI model learned a general notion of the target relation?

Possible Outcomes

- **Success**
  - Blindspot in the original NLI dataset
  - I GOT MY LocatedIn

- **Failure**
  - Inherent model limitation

---

Analyzing Generalization via Controlled Variance

**Numerical Reasoning**

**P:** I see 260 coins in the bucket.

**H:** I see more than 232 coins in the bucket.

Label: Entailment
Analyzing Generalization via Controlled Variance

1. Train/dev/test split across the dimension in focus

**Numerical Reasoning**

P: I see 260 coins in the bucket.
H: I see more than 232 coins in the bucket.

Label: Entailment

Axis: number range

Range: 100-200
Range: 400-500

Analyzing Generalization via Controlled Variance

1. Train/dev/test split across the dimension in focus
2. Fine-tune on one set and test on another

**Numerical Reasoning**

**P:** I see 260 coins in the bucket.

**H:** I see more than 232 coins in the bucket.

Label: Entailment

---

Analyzing Generalization via Controlled Variance

1. Train/dev/test split across the dimension in focus
2. Fine-tune on one set and test on another

Numerical Reasoning

**P:** I see 260 coins in the bucket.

**H:** I see more than 232 coins in the bucket.

Label: Entailment
Analyzing Generalization via Controlled Variance

1. Train/dev/test split across the dimension in focus
2. Fine-tune on one set and test on another

**Numerical Reasoning**

**P:** I see 260 coins in the bucket.
**H:** I see more than 232 coins in the bucket.

Label: Entailment
Analyzing Generalization via Controlled Variance

1. Train/dev/test split across the dimension in focus
2. Fine-tune on one set and test on another

Numerical Reasoning

P: I see 260 coins in the bucket.
H: I see more than 232 coins in the bucket.

Label: Entailment

- ✗ Axis: number range
- ✓ Axis: syntactic complexity

Good generalization capacity

Poor generalization capacity
Analyzing Generalization via Controlled Variance

1. Train/dev/test split across the dimension in focus
2. Fine-tune on one set and test on another

Numerical Reasoning
P: I see \textit{260 coins} in the bucket.
H: I see \textit{more than 232 coins} in the bucket.
Label: Entailment

Dative Alternation
P: I baked my mom a cake.
H: I baked a cake for my mom.
Label: Entailment

✓ Axis: lexical variability
✗ Axis: syntactic complexity

Analyzing Generalization via Controlled Variance

1. Train/dev/test split across the dimension in focus
2. Fine-tune on one set and test on another

**Numerical Reasoning**

P: I see 260 coins in the bucket.
H: I see more than 232 coins in the bucket.

Label: Entailment

**Dative Alternation**

P: I baked my mom a cake.
H: I baked a cake for my mom.

Label: Entailment

✅ Axis: lexical variability
❌ Axis: syntactic complexity

May decrease performance on the main task
(Richardson et al., 2020)
Incorporating Knowledge into NLI Models

The office is located in Baytown [SEP]
The office is located in Texas

label: entailment
Incorporating Knowledge into NLI Models

The office is located in Baytown [SEP]
The office is located in Texas

*label: entailment*

Model learns:

*Fact:* LocatedIn(Baytown, Texas)

*Rule:* Entailment between a city and its state in upward monotone sentences
Incorporating Knowledge into NLI Models

The office is located in **Baytown** [SEP] The office is located in **Texas**

**label: entailment**

**Model learns:**

**Fact:** LocatedIn(Baytown, Texas)

**Rule:** Entailment between a city and its state in upward monotone sentences

May be learned given enough data.
Incorporating Knowledge into NLI Models

The office is located in **Baytown** [SEP]
The office is located in **Texas**

**label: entailment**

**Model learns:**

**Fact:** LocatedIn(Baytown, Texas)

**Rule:** Entailment between a city and its state in upward monotone sentences

Impossible (and inefficient) to teach an NLI model every fact it might need.

May be learned given enough data.
Incorporating Knowledge into NLI Models

The office is located in **Baytown** [SEP] The office is located in **Texas**

**label: entailment**

---

**Fact:** LocatedIn(Baytown, Texas)

Model learns:

- **Rule:** Entailment between a city and its state in upward monotone sentences
Incorporating Knowledge into NLI Models

The office is located in **Baytown** [SEP]
The office is located in **Texas**

**label: entailment**

Fact: LocatedIn(Baytown, Texas)

Model learns:

Knowledge-Informed NLI Model

**Rule:** Entailment between a city and its state in upward monotone sentences

WIKIPEDIA
The Free Encyclopedia

How to incorporate relational knowledge?
Incorporating Factual Knowledge into LMs

Entity-Centric Knowledge

KnowBERT

Baytown is a city in the U.S. state of Texas

Incorporating Factual Knowledge into LMs

Entity-Centric Knowledge

KnowBERT

Baytown is a city in the U.S. state of Texas

State of Texas
Republic of Texas
Texas, Alabama

Baytown, Texas
Baytown culture
Operation Baytown

😔 performance improvement on relation extraction, entity typing, WSD

Incorporating Factual Knowledge into LMs

Entity-Centric Knowledge

KnowBERT

- can’t learn new facts after training
- performance improvement on relation extraction, entity typing, WSD
- The office is located in Nanhui [SEP]
- The office is located in China

Incorporating Factual Knowledge into LMs

Entity-Centric Knowledge

KnowBERT

Entity-Centric Knowledge

.entity-centric-knowledge

Performance improvement on relation extraction, entity typing, WSD

Entity-centric knowledge

Performance improvement on relation extraction, entity typing, WSD

Performance improvement on relation extraction, entity typing, WSD

Can’t learn new facts after training

The office is located in Nanhui [SEP]
The office is located in China

Performance improvement on relation extraction, entity typing, WSD

Computationally expensive to re-train

Incorporating Factual Knowledge into LMs

Relation-Centric Knowledge

The office is located in Baytown. [SEP] The office is located in Texas. 

**label: entailment**

**Fact:** LocatedIn(Baytown, Texas)
Incorporating Factual Knowledge into LMs

Relation-Centric Knowledge

The office is located in Baytown. [SEP] The office is located in Texas.

**label: entailment**

Fact: LocatedIn(Baytown, Texas)

**Premise:**

Model learns:

**Rule:** Entailment between a city and its state in upward monotone sentences

Relation Embeddings

**Hypothesis:**

InferBERT
Incorporating Factual Knowledge into LMs

Relation-Centric Knowledge

The office is located in Baytown. [SEP]

The office is located in Texas.

**Label:** entailment

**Fact:** LocatedIn(Baytown, Texas)

< 700 training examples
Incorporating Factual Knowledge into LMs

Relation-Centric Knowledge

The office is located in Baytown. [SEP]
The office is located in Texas.

_label: entailment_

InferBERT

Fact: LocatedIn(Baytown, Texas)

😊 performance improvement on NLI challenge sets

< 700 training examples
Incorporating Factual Knowledge into LMs

Relation-Centric Knowledge

The office is located in **Baytown**. [SEP]
The office is located in **Texas**.

**label: entailment**

Fact: LocatedIn(Baytown, Texas)

- 😃 performance improvement on NLI challenge sets
- 😞 can retrieve and use facts about unseen entities

The office is located in **Nanhui** [SEP]
The office is located in **China**

< 700 training examples

*Teach the Rules, Provide the Facts: Targeted Relational-knowledge Enhancement for Textual Inference.* Ohad Rozen, Shmuel Amar, **Vered Shwartz** and Ido Dagan. *SEM 2021*
Incorporating Factual Knowledge into LMs

Relation-Centric Knowledge

The office is located in Baytown. [SEP]
The office is located in Texas.

**label: entailment**

Fact: LocatedIn(Baytown, Texas)

😊 performance improvement on NLI challenge sets

😊 can retrieve and use facts about unseen entities

The office is located in Nanhui [SEP]
The office is located in China

😊 computationally efficient

< 700 training examples

Teach the Rules, Provide the Facts: Targeted Relational-knowledge Enhancement for Textual Inference. Ohad Rozen, Shmuel Amar, Vered Shwartz and Ido Dagan. *SEM 2021
Incorporating Factual Knowledge into LMs

Relation-Centric Knowledge

The office is located in **Baytown**. [SEP] The office is located in **Texas**.

(label: entailment)

Fact: LocatedIn(Baytown, Texas)

InferBERT

- performance improvement on NLI challenge sets
- can retrieve and use facts about unseen entities
- computationally efficient
- task-specific knowledge incorporation

< 700 training examples

Teach the Rules, Provide the Facts: Targeted Relational-knowledge Enhancement for Textual Inference. Ohad Rozen, Shmuel Amar, **Vered Shwartz** and Ido Dagan. *SEM 2021*
Limitations of neural NLI Models

Incorporating symbolic knowledge into neural NLI models

NLI is too easy? What’s next?
Real-World NLI #1

Partial Entailment

\( S_1: \) Amazon to acquire Whole Foods Market for $13.7 Billion.

\( S_2: \) Amazon is buying Whole Foods Market for almost $14 Billion in cash.
Real-World NLI #1

Partial Entailment

$S_1$: Amazon to acquire Whole Foods Market for $13.7 Billion.

$S_2$: Amazon is buying Whole Foods Market for almost $14 Billion in cash.
Partial Entailment

$S_1$: Amazon to acquire Whole Foods Market for $13.7$ Billion.

$S_2$: Amazon is buying Whole Foods Market for almost $14$ Billion in cash.

Subjectivity and inherent disagreements (Pavlick and Kwiatkowski, 2019)
Defeasible Natural Language Inference

P: Tweety is a bird.
H: Tweety flies.
P: Tweety is a bird.
H: Tweety flies.

NLI model: Entailment
P: Tweety is a bird.
H: Tweety flies.

NLI model: Entailment

Skeptical NLI model: given the information I currently have, I suppose so, but I can think of cases in which this is false.
Defeasible Inference (Reiter, 1980)

Given premise P, a hypothesis H is **defeasible** if there exists an update U (consistent with P) such that a human would find H less likely to be true after learning U.
Defeasible Inference (Reiter, 1980)

Given premise P, a hypothesis H is **defeasible** if there exists an update U (consistent with P) such that a human would find H less likely to be true after learning U.
Defeasible Inference (Reiter, 1980)

Given premise $P$, a hypothesis $H$ is **defeasible** if there exists an update $U$ (consistent with $P$) such that a human would find $H$ less likely to be true after learning $U$. 

P: Tweety is a bird.
Defeasible Inference (Reiter, 1980)

Given premise P, a hypothesis H is **defeasible** if there exists an update U (consistent with P) such that a human would find H less likely to be true after learning U.

P: Tweety is a bird.

H: Tweety flies.
Defeasible Inference (Reiter, 1980)

Given premise P, a hypothesis H is **defeasible** if there exists an update U (consistent with P) such that a human would find H less likely to be true after learning U.

P: Tweety is a bird.

H: Tweety flies.

U: Tweety is a penguin.
Defeasible Inference *(Reiter, 1980)*

Given premise P, a hypothesis H is **defeasible** if there exists an update U (consistent with P) such that a human would find H less likely to be true after learning U.

**Useful for Real-time Summarization:** Facts change as the story unfolds.

---

**Example:**

- **P:** Tweety is a bird.
- **H:** Tweety flies.
- **U:** Tweety is a penguin.
An update $U$ is called a **weaken**er if, given a premise $P$ and hypothesis $H$, a human would most likely find $H$ *less likely to be true* after learning $U$; if they would find $H$ *more likely to be true*, then we call $U$ a **strength**ener.
Defeasible Inference in Natural Language

An update $U$ is called a **weakener** if, given a premise $P$ and hypothesis $H$, a human would most likely find $H$ *less likely to be true* after learning $U$; if they would find $H$ *more likely to be true*, then we call $U$ a **strengthener**.

- **P**: Tweety is a bird. 
- **H**: Tweety flies.
- **Weakener**: Tweety is a penguin.
- **Strengthener**: Tweety is on a tree.

*Thinking Like a Skeptic: Defeasible Inference in Natural Language.*

Defeasible Inference in Natural Language
Defeasible Inference in Natural Language

**Discriminative Task**
Determine whether an update weakens or strengthens the hypothesis.

A group of people sitting around a rectangular table having either pieces of paper or laptops in front of them. They have a work meeting.

- They are in a library.  
  - Weakener

+ They are in a conference room.  
  + Strengthener
Defeasible Inference in Natural Language

**Discriminative Task**
Determine whether an update weakens or strengthens the hypothesis.

- A group of people sitting around a rectangular table having either pieces of paper or laptops in front of them.
- They have a work meeting.

**Generative Task**
Generate a weakening or strengthening update for a given premise-hypothesis pair.

- A group of people sitting around a rectangular table having either pieces of paper or laptops in front of them.
- They have a work meeting.

They are in a conference room. + Strengthener

They are in a library. - Weakened

They are in a conference room. +

They are in a library. -
Defeasible Inference in Natural Language

**Discriminative Task**
Determine whether an update weakens or strengthens the hypothesis.

| A group of people sitting around a rectangular table having either pieces of paper or laptops in front of them. |
| They have a work meeting. |
| They are in a conference room. |
| They are in a library. |
| + Strengthener |
| - Weaken |

**Generative Task**
Generate a weakening or strengthening update for a given premise-hypothesis pair.

<p>| A group of people sitting around a rectangular table having either pieces of paper or laptops in front of them. |
| They have a work meeting. |
| They are in a conference room. |
| They are in a library. |</p>
<table>
<thead>
<tr>
<th>+</th>
</tr>
</thead>
</table>

Language models leave plenty of room for improvement on the generative task!
Rationale Generation for Defeasible Inference

Rationale Generation for Defeasible Inference

A group of people sitting around a rectangular table having either pieces of paper or laptops in front of them. They have a work meeting.

• They are in a library.

• They are in a conference room.

A conference room is where people have meetings at work.

You must be quiet in the library, while work meetings involve talking.

A group of people sitting around a rectangular table having either pieces of paper or laptops in front of them. They have a work meeting.

They are in a conference room. A conference room is where people have meetings at work.

They are in a library. You must be quiet in the library, while work meetings involve talking.

Distant supervision:

They have a work meeting.

They are in a conference room.

They are in a library.

A conference room is where people have meetings at work.

You must be quiet in the library, while work meetings involve talking.

Distant supervision:

The definition of a library is...
Rationale Generation for Defeasible Inference
Rationale Generation for Defeasible Inference

Post hoc Rationalization  Generates a rationale for a given decision (label).

A group of people sitting around a rectangular table having either pieces of paper or laptops in front of them. They have a work meeting.

- They are in a conference room.  
  A conference room is where people have meetings at work.

+ They are in a conference room.  
  A conference room is where people have meetings at work.

- They are in a library.  
  You must be quiet in the library, while work meetings involve talking.
Rationale Generation for Defeasible Inference

**Post hoc Rationalization**  Generates a rationale for a given decision (label).

- A group of people sitting around a rectangular table having either pieces of paper or laptops in front of them. They have a work meeting.

- They are in a conference room.
  - A conference room is where people have meetings at work.

- They are in a library.
  - You must be quiet in the library, while work meetings involve talking.

Trivially rephrasing the label! (“[+] implies that [H]”)
Rationale Generation for Defeasible Inference

Post hoc Rationalization  Generates a rationale for a given decision (label).

A group of people sitting around a rectangular table having either pieces of paper or laptops in front of them. They have a work meeting.  

They are in a conference room.  

A conference room is where people have meetings at work.

They are in a library.  

You must be quiet in the library, while work meetings involve talking.

Trivially rephrasing the label! (“[+] implies that [H]”)

Joint Prediction & Rationalization  Predict the label (strengthener / weaker) and rationalize it.

A group of people sitting around a rectangular table having either pieces of paper or laptops in front of them. They have a work meeting.

They are in a conference room.  

A conference room is where people have meetings at work.

They are in a library.  

You must be quiet in the library, while work meetings involve talking.
Rationale Generation for Defeasible Inference

Post hoc Rationalization  Generates a rationale for a given decision (label).

A group of people sitting around a rectangular table having either pieces of paper or laptops in front of them.  They have a work meeting.  

+ They are in a conference room.  A conference room is where people have meetings at work.  

- They are in a library.  You must be quiet in the library, while work meetings involve talking.  

Trivially rephrasing the label! (“[+] implies that [H]”)

Joint Prediction & Rationalization  Predict the label (strengthener / weaker) and rationalize it.

A group of people sitting around a rectangular table having either pieces of paper or laptops in front of them.  They have a work meeting.  

They are in a conference room.  A conference room is where people have meetings at work.  

They are in a library.  You must be quiet in the library, while work meetings involve talking.  

More realistic but very challenging task!
Recap

Neural models achieve impressive gains on NLI
- But make stupid unhuman like errors
- “Human performance” is debatable
Recap

🤖 Neural models achieve impressive gains on NLI
  - But make stupid unhuman like errors
  - “Human performance” is debatable

🤖 Symbolic knowledge is useful
Recap

🤖 Neural models achieve impressive gains on NLI
  • But make stupid unhuman like errors
  • “Human performance” is debatable

🤖 Symbolic knowledge is useful
  • Accurate
Recap

🤖 Neural models achieve impressive gains on NLI
  • But make stupid unhuman like errors
  • “Human performance” is debatable

🤖 Symbolic knowledge is useful
  • Accurate
  • More efficient than “learning all the facts”
Recap

🤖 Neural models achieve impressive gains on NLI
  • But make stupid unhuman like errors
  • “Human performance” is debatable

🤖 Symbolic knowledge is useful
  • Accurate
  • More efficient than “learning all the facts”
  • Combination is not trivial
Recap

🤖 Neural models achieve impressive gains on NLI
  • But make stupid unhuman like errors
  • “Human performance” is debatable

🤖 Symbolic knowledge is useful
  • Accurate
  • More efficient than “learning all the facts”
  • Combination is not trivial

🤖 Time to work on more real-world NLI tasks
  • Partial and defeasible inferences
Recap

🤖 Neural models achieve impressive gains on NLI
  • But make stupid unhuman like errors
  • “Human performance” is debatable

🤖 Symbolic knowledge is useful
  • Accurate
  • More efficient than “learning all the facts”
  • Combination is not trivial

🤖 Time to work on more real-world NLI tasks
  • Partial and defeasible inferences

Thank You!

@VeredShwartz  vereds@allenai.org
References (1)

References (2)


