Commonsense **Knowledge and Reasoning** in Natural Language

Vered Shwartz
Allen Institute for AI (AI2)
Allen School of Computer Science & Engineering, University of Washington
Stevie Wonder announces he’ll be having kidney surgery during London concert
What happens during the London concert?

1) The kidney surgery
2) The announcement

Stevie Wonder announces he’ll be having kidney surgery during London concert
Stevie Wonder announces he’ll be having kidney surgery during London concert.
Stevie Wonder announces he’ll be having kidney surgery during London concert.

Language

What happens during the London concert?
1) The kidney surgery
2) The announcement

Knowledge

- Kidney surgery is performed under general anesthesia
- People are unconscious under general anesthesia
- Performing actions requires being conscious
- ...

Reasoning

It’s impossible to perform a concert while undergoing surgery.
The Deep Learning Revolution
The Deep Learning Revolution

Translation
Google's AI translation system is approaching human-level accuracy
The Deep Learning Revolution

Translation
Google's AI translation system is approaching human-level accuracy

Reading Comprehension
ALIBABA AI BEATS HUMANS IN READING-COMPREHENSION TEST
CHRISTINE CHOU | JULY 9, 2019
The Deep Learning Revolution

Translation

Google's AI translation system is approaching human-level accuracy.

Reading Comprehension

ALIBABA AI BEATS HUMANS COMPREHENSION TEST

CHRISTINE CHOU | JULY 9, 2019

Chatbots

Artificial Intelligence / Voice assistants

Your next doctor’s appointment might be with an AI

A new wave of chatbots are replacing physicians and providing frontline medical advice—but are they as good as the real thing?

by Will Douglas Heaven | October 16, 2018
The Deep Learning Revolution

Translation
Google's AI translation system is approaching human-level accuracy

Reading Comprehension
ALIBABA AI BEATS HUMANS COMPREHENSION TEST
CHRISTINE CHOU | JULY 9, 2019

Chatbots
Artificial intelligence / Voice assistants

Your next doctor’s appointment might be with an AI
A new wave of chatbots are replacing physicians and providing frontline medical advice—but are they as good as the real thing?
by Will Douglas Heaven October 16, 2018

Human Performance:
Baseline

SuperGLUE

Jul '19  Aug '19  Jan '20  Jan '21
The Deep Learning Revolution

Translation
Google's AI translation system is approaching human-level accuracy

Reading Comprehension
ALIBABA AI BEATS HUMANS COMPREHENSION TEST
CHRISTINE CHOU | JULY 9, 2019

Chatbots
Artificial intelligence / Voice assistants
Your next doctor’s appointment might be with an AI
A new wave of chatbots are replacing physicians and providing frontline medical advice—but are they as good as the real thing?
by Will Douglas Heaven October 16, 2018

Does this mean language understanding is nearly solved?

Human Performance: 90.2 90.3

SuperGLUE

Jul ‘19 Aug ‘19 Jan ‘20 Jan ‘21
The Deep Learning Revolution

Translation

Google's AI translation system is approaching human-level accuracy

Reading Comprehension

ALIBABA AI BEATS HUMANS
COMPREHENSION TEST
CHRISTINE CHOU | JULY 9, 2019

Chatbots

Artificial intelligence / Voice assistants

Your next doctor’s appointment might be with an AI

A new wave of chatbots are replacing physicians and providing frontline medical advice—but are they as good as the real thing?

by Will Douglas Heaven October 16, 2018

Does this mean language understanding is nearly solved?
What are the remaining challenges?

Human Performance:

Baseline

Facebook

Google AI

Microsoft

SuperGLUE
Is Natural Language Understanding Nearly Solved?
Is Natural Language Understanding Nearly Solved?

- Pre-training
- Syntax
- Word meanings
- Factual Knowledge
- ...
Is Natural Language Understanding Nearly Solved?

Pre-training

- Syntax
- Word meanings
- Factual Knowledge
- ...

Fine-tuning:

Language Model
Is Natural Language Understanding Nearly Solved?

Pre-training

Language Model

Fine-tuning:

The chocolate cake is amazing

✅ Syntax
✅ Word meanings
✅ Factual Knowledge
✅ ...

Google AI
Is Natural Language Understanding Nearly Solved?

Pre-training: Google AI

WIKIPEDIA The Free Encyclopedia

Fine-tuning: ✅ Syntax
✅ Word meanings
✅ Factual Knowledge
✅ ...

5.4% 94.6%

Language Model

The chocolate cake is amazing
Is Natural Language Understanding Nearly Solved?

Pre-training:
- Syntax
- Word meanings
- Factual Knowledge
- ...

Fine-tuning:
- Understanding the task
- Learning to solve the task

The chocolate cake is amazing
Is Natural Language Understanding Nearly Solved?

Pre-training

Google AI

Wikipedia
The Free Encyclopedia

Fine-tuning:

Language Model

The chocolate cake is amazing

Syntax
Word meanings
Factual Knowledge
...

Understanding the task
Learning to solve the task

What are the remaining challenges?

Generalization to unknown situations

5.4% 94.6%
Overfitting to Data-specific Spurious Correlations
Overfitting to Data-specific Spurious Correlations

เผย: A horse standing in the grass.

(Szegedy et al., 2015)
Overfitting to Data-specific Spurious Correlations

🤖: A horse standing in the grass.
(Szegedy et al., 2015)
Overfitting to Data-specific Spurious Correlations

💬: A horse standing in the grass.
(Szegedy et al., 2015)

🤖: 2
(Agrawal et al., 2016)

How many zebras?
Overfitting to Data-specific Spurious Correlations

🤖: A horse standing in the grass.
(Szegedy et al., 2015)

🤖: 2
(Agrawal et al., 2016)

How many zebras? 2
(Agrawal et al., 2016)

How many giraffes? 2

How many dogs? 2
Overfitting to Data-specific Spurious Correlations

**: A horse standing in the grass.**
(Szegedy et al., 2015)

**: 2**
(Agrawal et al., 2016)

**: I only had a soup but it was very filling.**
**: I didn't eat a salad.**
**: contradiction (91.7%)**
(Gururangan, Swayamdipta, et al., 2018; Poliak et al., 2018)

How many dogs? 2
How many giraffes? 2
Overfitting to Data-specific Spurious Correlations

- **p:** I only had a soup but it was very filling.
  **h:** I *didn't* eat a salad.
  - contradiction (91.7%)
  
  (Gururangan, Swayamdipta, et al., 2018; Poliak et al., 2018)

- **p:** The boy ran in the park.
  **h:** The boy *didn't* run in the park.
  - contradiction

- **p:** A horse standing in the grass.
  - (Szegedy et al., 2015)

- **p:** 2
  - (Agrawal et al., 2016)

- **p:** How many zebra?
  - 2
  - How many giraffes? 2
  - How many dogs? 2

- **p:** The boy ran in the park.
  **h:** The boy *didn't* run in the park.
  - contradiction
  
  (Szegedy et al., 2015)

- **p:** I didn't eat a salad.
  
  (Agrawal et al., 2016)
Overfitting to Data-specific Spurious Correlations

How many zebras? 2

(Agrawal et al., 2016)

How many giraffes? 2

How many zebras? 2

How many gira
ffes? 2

How many dogs? 2

Contradiction (91.7%)

(Gururangan, Swayamdipta, et al., 2018; Poliak et al., 2018)

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

h: I didn't eat a salad.

Contradiction

...Solving datasets but not underlying tasks!
Addressing Unknown Situations with Commonsense

**Translation**

Google's AI translation system is approaching human-level accuracy

**Chatbots**

*Artificial intelligence / Voice assistants*

Your next doctor’s appointment might be with an AI

A new wave of chatbots are replacing physicians and providing frontline medical advice—but are they as good as the real thing?

by Will Douglas Heaven

October 16, 2018

**Reading Comprehension**

ALIBABA AI BEATS HUMANS IN READING-COMPREHENSION TEST

CHRISTINE CHOU | JULY 9, 2019
Addressing Unknown Situations with Commonsense

Translation

<table>
<thead>
<tr>
<th>ENGLISH - DETECTED</th>
<th>HEBREW</th>
<th>HEbrew</th>
<th>ENGLISH</th>
<th>SPANISH</th>
</tr>
</thead>
<tbody>
<tr>
<td>grass-fed yogurt</td>
<td>יוגורט עם דשא</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

= yogurt with grass

Chatbots

Artificial intelligence / Voice assistants

Your next doctor’s appointment might be with an AI

A new wave of chatbots are replacing physicians and providing frontline medical advice—but are they as good as the real thing?

by Will Douglas Heaven

October 16, 2018

Reading Comprehension

ALIBABA AI BEATS HUMANS IN READING-COMPREHENSION TEST

CHRISTINE CHOU | JULY 9, 2019
Addressing Unknown Situations with Commonsense

Translation

<table>
<thead>
<tr>
<th>ENGLISH - DETECTED</th>
<th>HEBREW</th>
<th>HEBREW</th>
<th>ENGLISH</th>
<th>SPANISH</th>
</tr>
</thead>
<tbody>
<tr>
<td>grass-fed yogurt</td>
<td>יוגורט עם דשא</td>
<td>יוגורט עם דשא</td>
<td>yogurt with grass</td>
<td></td>
</tr>
</tbody>
</table>

Chatbots

Artificial intelligence / Voice assistants

Your next doctor’s appointment might be with an AI

A new wave of chatbots are replacing physicians and providing frontline medical advice—but are they as good as the real thing?

by Will Douglas Heaven  October 16, 2018

Reading Comprehension

Stevie Wonder announces
he'll be having kidney surgery
during London concert
Addressing Unknown Situations with Commonsense

Translation

<table>
<thead>
<tr>
<th>ENGLISH - DETECTED</th>
<th>HEBREW</th>
<th>HEBREW</th>
<th>ENGLISH</th>
<th>SPANISH</th>
</tr>
</thead>
<tbody>
<tr>
<td>grass-fed yogurt</td>
<td>יוגרט עם דשא</td>
<td>יוגרט עם דשא</td>
<td>yogurt with grass</td>
<td>yogurt with grass</td>
</tr>
</tbody>
</table>

Chatbots

Medical chatbot using OpenAI’s GPT-3 told a fake patient to kill themselves

Reading Comprehension

Stevie Wonder announces he'll be having kidney surgery during London concert
What is Commonsense?

The basic level of practical knowledge and reasoning concerning everyday situations and events that are commonly shared among most people.
What is Commonsense?

The basic level of practical knowledge and reasoning concerning everyday situations and events that are commonly shared among most people.

It’s a bad idea to touch a hot stove.
What is Commonsense?

The basic level of **practical knowledge** and **reasoning** concerning **everyday situations** and **events** that are **commonly** shared among **most** people.

- It’s a bad idea to touch a hot stove.
- It’s impolite to comment on people’s weight.

*Introductory Tutorial on Commonsense Reasoning* by Maarten Sap, **Vered Shwartz**, Antoine Bosselut, Dan Roth, and Yejin Choi. ACL 2020.
What is Commonsense?

The basic level of **practical knowledge** and **reasoning** concerning **everyday situations** and **events** that are **commonly** shared among **most** people.

- It’s a bad idea to touch a hot stove.
- It’s impolite to comment on people’s weight.
- Eating dinner comes before going to bed.
What is Commonsense?

The basic level of **practical knowledge** and **reasoning** concerning **everyday situations** and **events** that are **commonly** shared among **most** people.

- It’s a bad idea to touch a hot stove.
- It’s impolite to comment on people’s weight.
- Eating dinner comes before going to bed.
- …

*Introductory Tutorial on Commonsense Reasoning.* Maarten Sap, **Vered Shwartz**, Antoine Bosselut, Dan Roth, and Yejin Choi. ACL 2020.
Language

<table>
<thead>
<tr>
<th>Lexical Semantics</th>
</tr>
</thead>
<tbody>
<tr>
<td>• ACL 2016 [outstanding paper]</td>
</tr>
<tr>
<td>• *SEM 2016 [best paper]</td>
</tr>
<tr>
<td>• *SEM 2017 (a)</td>
</tr>
<tr>
<td>• EACL 2017</td>
</tr>
<tr>
<td>• *SEM 2018</td>
</tr>
<tr>
<td>• CogALex 2016</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Lexical Composition</th>
</tr>
</thead>
<tbody>
<tr>
<td>• NAACL 2018</td>
</tr>
<tr>
<td>• ACL 2018 (a)</td>
</tr>
<tr>
<td>• MWE 2019</td>
</tr>
<tr>
<td>• TACL 2019</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Coreference Resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td>• *SEM 2017 (b)</td>
</tr>
<tr>
<td>• ACL 2019</td>
</tr>
<tr>
<td>• Findings of EMNLP 2020 (a)</td>
</tr>
</tbody>
</table>
Language

Lexical Semantics
- ACL 2016 [outstanding paper]
- *SEM 2016 [best paper]
- *SEM 2017 (a)
- EACL 2017
- *SEM 2018
- CogALex 2016

Lexical Composition
- NAACL 2018
- ACL 2018 (a)
- MWE 2019
- TACL 2019

Coreference Resolution
- *SEM 2017 (b)
- ACL 2019
- Findings of EMNLP 2020 (a)

Knowledge

Knowledge Representation
- LSDSem 2017
- CoNLL 2015

Knowledge Acquisition
- EMNLP 2020 (a)
Reasoning

Natural Language Inference
- *SEM 2015
- CoNLL 2019

Causal Reasoning
- IWCS 2017
- AAAI 2021 (a)

Nonmonotonic Reasoning
- EMNLP 2020 (b)
- Findings of EMNLP 2020 (b)
- AAAI 2021 (b)

Reasoning about Social Norms
- EMNLP 2020 (c)

Knowledge

Knowledge Representation
- LSDSem 2017
- CoNLL 2015

Knowledge Acquisition
- EMNLP 2020 (a)

Language

Lexical Semantics
- ACL 2016 [outstanding paper]
- *SEM 2016 [best paper]
- *SEM 2017 (a)
- EACL 2017
- *SEM 2018
- CogALex 2016

Lexical Composition
- NAACL 2018
- ACL 2018 (a)
- MWE 2019
- TACL 2019

Coreference Resolution
- *SEM 2017 (b)
- ACL 2019
- Findings of EMNLP 2020 (a)

Reasoning

*SEM 2015
- CoNLL 2019

Causal Reasoning
- IWCS 2017
- AAAI 2021 (a)

Nonmonotonic Reasoning
- EMNLP 2020 (b)
- Findings of EMNLP 2020 (b)
- AAAI 2021 (b)

Reasoning about Social Norms
- EMNLP 2020 (c)
Language

Lexical Semantics
- ACL 2016 [outstanding paper]
- *SEM 2016 [best paper]
- *SEM 2017 (a)
- EACL 2017
- *SEM 2018
- CogALex 2016

Lexical Composition
- NAACL 2018
- ACL 2018 (a)
- MWE 2019
- TACL 2019

Coreference Resolution
- *SEM 2017 (b)
- ACL 2019
- Findings of EMNLP 2020 (a)

Knowledge

Knowledge Representation
- LSDSem 2017
- CoNLL 2015

Knowledge Acquisition
- EMNLP 2020 (a)

Reasoning

Natural Language Inference
- *SEM 2015
- CoNLL 2019

Causal Reasoning
- IWCS 2017
- AAAI 2021 (a)

Nonmonotonic Reasoning
- EMNLP 2020 (b)
- Findings of EMNLP 2020 (b)
- AAAI 2021 (b)

Reasoning about Social Norms
- EMNLP 2020 (c)
Language

Knowledge

Reasoning
Introspective knowledge acquisition through asking questions and reasoning.
Introspective knowledge acquisition through asking questions and reasoning
Nonmonotonic reasoning in natural language

Introspective knowledge acquisition through asking questions and reasoning
1 Introspective **knowledge acquisition** through asking questions and **reasoning**

2 Nonmonotonic **reasoning** in natural **language**
Introspective **knowledge acquisition** through asking questions and **reasoning**

Nonmonotonic **reasoning** in natural **language**

Understanding **language** through generalizing existing world **knowledge**
1. Introspective knowledge acquisition through asking questions and reasoning
2. Nonmonotonic reasoning in natural language
3. Understanding language through generalizing existing world knowledge
Introspective **knowledge acquisition** through asking questions and **reasoning**

2. **Nonmonotonic reasoning** in natural **language**

3. Understanding **language** through generalizing existing world **knowledge**

4. Future directions
Introspective knowledge acquisition through asking questions and reasoning

Nonmonotonic reasoning in natural language

Understanding language through generalizing existing world knowledge

Future directions

Unsupervised Commonsense Question Answering with Self-Talk
Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula and Yejin Choi
EMNLP 2020
Children need to eat more vegetables because they are healthy.
Children need to eat more vegetables because they are healthy.
Reasoning with Implicit Knowledge

Children need to eat more vegetables because they are healthy.

1. Vegetables are healthy.
Reasoning with Implicit Knowledge

Children need to eat more vegetables because they are healthy

1. Vegetables are healthy.
2. Eating vegetables can make you healthier.
Reasoning with Implicit Knowledge

Children need to eat more vegetables because they are healthy.

1. Vegetables are healthy.
2. Eating vegetables can make you healthier.
3. People want to be healthy.
Reasoning with Implicit Knowledge

Children need to eat more vegetables because they are healthy

1. Vegetables are healthy.
2. Eating vegetables can make you healthier.
3. People want to be healthy.
Reasoning with Implicit Knowledge

Children need to eat more vegetables because they are healthy

1. Vegetables are healthy.
2. Eating vegetables can make you healthier.
3. People want to be healthy.

Research Questions:
Where to get knowledge?
How to incorporate this knowledge?
Discovery Learning (Bruner, 1961)

Children need to eat more vegetables because they are healthy
Discovery Learning (Bruner, 1961)

Children need to eat more vegetables because they are healthy

- children
- vegetables

Learner
Discovery Learning (Bruner, 1961)

Children need to eat more vegetables because they are healthy

Self-Inquiry: What are the properties of vegetables?
Discovery Learning (Bruner, 1961)

Children need to eat more vegetables because they are healthy

Self-Inquiry

What are the properties of vegetables?

Existing Knowledge

Vegetables are full of vitamins.

Learner

children
vegetables
Discovery Learning (Bruner, 1961)

Children need to eat more vegetables because they are healthy.

Self-Inquiry:
- What are the properties of vegetables?

Existing Knowledge:
- Vegetables are full of vitamins.

New Facts:
- children
- vegetables
The Self-Talk Paradigm

**Main Question**
Children need to eat more vegetables because they are healthy

**Nested QA**

**Self-Inquiry**
What are the properties of vegetables?

**Existing Knowledge**
Vegetables are full of vitamins.

**Neural Language Model**

**Main Answer**

*children* *vegetables*
Children need to eat more vegetables because they are healthy.

Answer choices:
children, vegetables

Predicted answer choice: vegetables
Children need to eat more vegetables because they are healthy.
Knowledge Discovery
Knowledge Discovery
Children need to eat more vegetables because they are healthy.
Children need to eat more vegetables because they are healthy.

Knowledge Discovery

<table>
<thead>
<tr>
<th>Instance</th>
<th>Nested Question Prefix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Children need to eat more vegetables because they are healthy.</td>
<td>What is the purpose of</td>
</tr>
<tr>
<td>Instance</td>
<td>Nested Question Prefix</td>
</tr>
<tr>
<td>-------------------------------------------------------------------------</td>
<td>------------------------</td>
</tr>
<tr>
<td>Children need to eat more vegetables because <strong>they</strong> are healthy.</td>
<td>What is the purpose of</td>
</tr>
<tr>
<td></td>
<td>vegetables?</td>
</tr>
</tbody>
</table>
Knowledge Discovery

Instance
Children need to eat more vegetables because they are healthy.

Nested Question Prefix
What is the purpose of vegetables?

Nested Answer Prefix
The purpose of ________ is
The purpose of vegetables is

Children need to eat more vegetables because they are healthy.

What is the purpose of vegetables?

The purpose of vegetables is
Children need to eat more vegetables because they are healthy.

What is the purpose of vegetables?

The purpose of vegetables is to provide a good base of nutrients and energy.
The purpose of vegetables is to provide a good base of nutrients and energy.
The purpose of vegetables is to provide a good base of nutrients and energy.
The definition of healthy is quality of life that is free of diseases.
The properties of being healthy are linked to the effects of exercise.
### Knowledge Discovery

<table>
<thead>
<tr>
<th>Instance</th>
<th>Nested Question Prefix</th>
<th>Nested Answer Prefix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Children need to eat more vegetables because they are healthy.</td>
<td>What is the purpose of vegetables?</td>
<td>The purpose of vegetables is to provide a good base of nutrients and energy.</td>
</tr>
</tbody>
</table>

**Nested Answers**
- The purpose of vegetables is to provide a good base of nutrients and energy.
- The definition of healthy is quality of life that is free of diseases.
- The properties of being healthy are linked to the effects of exercise.
<table>
<thead>
<tr>
<th>Instance</th>
<th>Nested Question Prefix</th>
<th>Nested Answer Prefix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Children need to eat more vegetables because they are healthy.</td>
<td>What is the purpose of</td>
<td>The purpose of vegetables is to provide a good base of nutrients and energy.</td>
</tr>
<tr>
<td></td>
<td>vegetables?</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>The definition of healthy is quality of life that is free of diseases.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>The properties of being healthy are linked to the effects of exercise.</td>
</tr>
</tbody>
</table>
Language Models
Text Generation

Language Model
Language Models
Text Generation

The properties of vegetables are
Language Models

Text Generation

The properties of vegetables are...
Language Models

Text Generation

The properties of vegetables are sampling p(next) that
Language Models

Text Generation

The properties of vegetables are sampling that they are full of vitamins.
Children need to eat more vegetables because they are healthy.

The purpose of vegetables is to provide a good base of nutrients and energy.

The properties of being healthy are linked to the effects of exercise.

The definition of healthy is quality of life that is free of diseases.

What is the definition of healthy?
Question Answering
Question Answering
Children need to eat more vegetables because **children** are healthy. The purpose of vegetables is to provide a good base of nutrients and energy.

...  

Children need to eat more vegetables because **children** are healthy. The definition of healthy is quality of life that is free of diseases.
## Question Answering

<table>
<thead>
<tr>
<th>children</th>
<th>vegetables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Children need to eat more vegetables because <em>children</em> are healthy. The purpose of vegetables is to provide a good base of nutrients and energy.</td>
<td>Children need to eat more vegetables because <em>vegetables</em> are healthy. The purpose of vegetables is to provide a good base of nutrients and energy.</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Children need to eat more vegetables because <em>children</em> are healthy. The definition of healthy is quality of life that is free of diseases.</td>
<td>Children need to eat more vegetables because <em>vegetables</em> are healthy. The definition of healthy is quality of life that is free of diseases.</td>
</tr>
</tbody>
</table>

### Most plausible statement

- Children need to eat more vegetables because *children* are healthy. The purpose of vegetables is to provide a good base of nutrients and energy.

- Children need to eat more vegetables because *vegetables* are healthy. The purpose of vegetables is to provide a good base of nutrients and energy.

- Children need to eat more vegetables because *children* are healthy. The definition of healthy is quality of life that is free of diseases.

- Children need to eat more vegetables because *vegetables* are healthy. The definition of healthy is quality of life that is free of diseases.
## Question Answering

<table>
<thead>
<tr>
<th>children</th>
<th>vegetables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Children need to eat more vegetables because <em>children</em> are healthy. The purpose of vegetables is to provide a good base of nutrients and energy.</td>
<td>Children need to eat more vegetables because <em>vegetables</em> are healthy. The purpose of vegetables is to provide a good base of nutrients and energy.</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Children need to eat more vegetables because <em>children</em> are healthy. The definition of healthy is quality of life that is free of diseases.</td>
<td>Children need to eat more vegetables because <em>vegetables</em> are healthy. The definition of healthy is quality of life that is free of diseases.</td>
</tr>
</tbody>
</table>

**Most plausible statement** → Statement with best language model score
### Question Answering

<table>
<thead>
<tr>
<th><strong>Most plausible statement</strong></th>
<th><strong>Statement with best language model score</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Children need to eat more vegetables because children are healthy. The purpose of vegetables is to provide a good base of nutrients and energy.</td>
<td>Children need to eat more vegetables because vegetables are healthy. The purpose of vegetables is to provide a good base of nutrients and energy.</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Children need to eat more vegetables because children are healthy. The definition of healthy is quality of life that is free of diseases.</td>
<td>Children need to eat more vegetables because vegetables are healthy. The definition of healthy is quality of life that is free of diseases.</td>
</tr>
</tbody>
</table>

**Language Model**
Most plausible statement

Children need to eat more vegetables because children are healthy. The purpose of vegetables is to provide a good base of nutrients and energy.

Statement with best language model score

$p(\text{need} | \text{Children}) \cdot p(\text{to} | \text{Children need}) \cdot \cdots \cdot p(\text{energy} | \cdots) \cdot p(<\text{eos}> | \cdots)$

Children

need

... and

energy

Language Model

Vegetables

Children

Vegetables are healthy. The purpose of vegetables is to provide a good base of nutrients and energy.

Children need to eat more vegetables because vegetables are healthy. The definition of healthy is quality of life that is free of diseases.
Question Answering

**Most plausible statement**

\[
\text{score} = - \frac{1}{n} \log \left( p(\text{need} | \text{Children}) \cdot p(\text{to} | \text{Children need}) \cdots p(\text{energy} | \ldots) \cdot p(<\text{eos}> | \ldots) \right)
\]

<table>
<thead>
<tr>
<th>children</th>
<th>vegetables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Children need to eat more vegetables because <em>children</em> are healthy. The purpose of vegetables is to provide a good base of nutrients and energy.</td>
<td>Children need to eat more vegetables because <em>vegetables</em> are healthy. The purpose of vegetables is to provide a good base of nutrients and energy.</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Children need to eat more vegetables because <em>children</em> are healthy. The definition of healthy is quality of life that is free of diseases.</td>
<td>Children need to eat more vegetables because <em>vegetables</em> are healthy. The definition of healthy is quality of life that is free of diseases.</td>
</tr>
</tbody>
</table>

**Statement with best language model score**
Children need to eat more vegetables because *children* are healthy. The purpose of vegetables is to provide a good base of nutrients and energy.

... 

Children need to eat more vegetables because *children* are healthy. The definition of healthy is quality of life that is free of diseases.

... 

Children need to eat more vegetables because *vegetables* are healthy. The purpose of vegetables is to provide a good base of nutrients and energy.

Children need to eat more vegetables because *vegetables* are healthy. The definition of healthy is quality of life that is free of diseases.
Children need to eat more vegetables because *children* are healthy. The purpose of vegetables is to provide a good base of nutrients and energy.

... 

Children need to eat more vegetables because *vegetables* are healthy. The purpose of vegetables is to provide a good base of nutrients and energy.

... 

Children need to eat more vegetables because *children* are healthy. The definition of healthy is quality of life that is free of diseases.

Children need to eat more vegetables because *vegetables* are healthy. The definition of healthy is quality of life that is free of diseases.
Children need to eat more vegetables because they are healthy.

Predicted answer choice: vegetables
Input
Context:
Children need to eat more vegetables because they are healthy.
Answer choices: children, vegetables

Output
Predicted answer choice: vegetables

Knowledge Discovery
What is the definition of healthy?

Question Answering
The purpose of vegetables is to provide a good base of nutrients and energy.
The properties of being healthy are linked to the effects of exercise.
The definition of healthy is quality of life that is free of diseases.

Answer with most plausible statement
Experiments

Tasks
- Social IQa
- Physical PIQA
- Temporal MC-TACO
- Causal COPA
- General CommonsenseQA WinoGrande

Baselines
- Children need to eat more vegetables because they are healthy.
- Vegetables are required for eating vegetables.
- Eating vegetables is motivated by being healthy.
- Vegetables are healthy.
- Because the children wanted to live longer.

Results

[Bar chart showing results]
Experiments

Tasks

<table>
<thead>
<tr>
<th>Social IQa</th>
<th>Causal</th>
</tr>
</thead>
<tbody>
<tr>
<td>SocialIQa</td>
<td>COPA</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Physical</th>
<th>General</th>
</tr>
</thead>
<tbody>
<tr>
<td>PIQA</td>
<td>CommonsenseQA</td>
</tr>
<tr>
<td></td>
<td>WinoGrande</td>
</tr>
</tbody>
</table>

Temporal

MC-TACO

Baselines

Children need to eat more vegetables because they are healthy.

Results

Vegetables are required for eating vegetables.
Eating vegetables is motivated by being healthy.
Vegetables are healthy.
Because the children wanted to live longer.
Experiments

- Social: SocialIQa
- Physical: PIQA
- Temporal: MC-TACO
- Causal: COPA
- General: CommonsenseQA, WinoGrande
MC-TACO
Temporal
SocialIQa
Physical
PIQA
Temporal
MC-TACO
Causal
COPA
General
CommonsenseQA
WinoGrande
Commonsense Question Answering Tasks

- Social
  - SocialIQa
- Causal
  - COPA
- Physical
  - PIQA
- General
  - CommonsenseQA
  - WinoGrande
- Temporal
  - MC-TACO
Commonsense Question Answering Tasks

**Social Interaction QA**

Although Aubrey was older and stronger, they lost to Alex in arm wrestling. How would Alex feel as a result?

1) they need to practice more.
2) ashamed.
3) **boastful.**
Commonsense Question Answering Tasks

Social Interaction QA

Although Aubrey was older and stronger, they lost to Alex in arm wrestling. How would Alex feel as a result?

1) they need to practice more.
2) ashamed.
3) boastful.

Choice of Plausible Alternatives

The man broke his toe. What was the cause?

1) He got a hole in his sock.
2) He dropped a hammer on his foot.
Experiments

Tasks

- Social IQa
  - SocialIQa
  - MC-TACO
- Causal
  - COPA
- Physical
  - PIQA
- General
  - CommonsenseQA
  - WinoGrande
- Temporal
  - MC-TACO

Baselines

- Children need to eat more vegetables because they are healthy.
  - vegetables
    - healthy
  - vegetables
    - eating vegetables
    - healthy
  - eating vegetables
    - motivated
    - vegetables
    - because the children wanted to live longer

Results

- Graph showing results for different baselines.
Baselines
Baselines

Children need to eat more vegetables because they are healthy.
Baselines

Children need to eat more vegetables because they are healthy.

∅

No Inquiry
Baselines

Children need to eat more vegetables because they are healthy.
Children need to eat more vegetables because they are healthy.
Baseline

Children need to eat more vegetables because they are healthy.

Vegetables are required for eating vegetables. Eating vegetables is motivated by being healthy.
Children need to eat more vegetables because they are healthy.
Children need to eat more vegetables because they are healthy.

Vegetables are required for eating vegetables.
Eating vegetables is motivated by being healthy.
Vegetables are healthy.
Because the children wanted to live longer.

No Inquiry

Expert Knowledge
Experiments

Tasks
- Social
  - SocialIQa
- Causal
  - COPA
- Physical
  - PIQA
- General
  - CommonsenseQA
  - WinoGrande
- Temporal
  - MC-TACO

Baselines

Results

Children need to eat more vegetables because they are healthy.

Vegetables are required for eating vegetables.
Eating vegetables is motivated by being healthy.
Vegetables are healthy.
Because the children wanted to live longer.
Results
Results

No Inquiry
Expert Knowledge
Self-Talk
Human

Results
1. Nested QA improves performance
1. Nested QA improves performance
2. Self-Talk performs similarly to models with expert knowledge
1. Nested QA improves performance
2. Self-Talk performs similarly to models with expert knowledge
3. Gap from human performance
Interactive QA Paradigm

Future Directions

Research Questions:
Where to get knowledge?
How to incorporate this knowledge?
Interactive QA Paradigm

Future Directions

Research Questions:
Where to get knowledge?
How to incorporate this knowledge?
What knowledge is missing?
Is a certain fact useful?
Interactive QA Paradigm

Future Directions

Research Questions:
Where to get knowledge?
How to incorporate this knowledge?

What knowledge is missing?
Is a certain fact useful?

Introspection:
Interactive QA Paradigm

Future Directions

Research Questions:
Where to get knowledge?
How to incorporate this knowledge?

- What knowledge is missing?
- Is a certain fact useful?

Introspection:
• Can I answer the main question without additional question?
Interactive QA Paradigm

Future Directions

Research Questions:
Where to get knowledge?
How to incorporate this knowledge?

What knowledge is missing?
Is a certain fact useful?

Introspection:
• Can I answer the main question without additional question?
• What should I ask about?
Interactive QA Paradigm

Future Directions

Research Questions:
Where to get knowledge?
How to incorporate this knowledge?

What knowledge is missing?
Is a certain fact useful?

Introspection:
- Can I answer the main question without additional question?
- What should I ask about?

Relevant but *unhelpful* knowledge may confuse the model!
Interactive QA Paradigm

Future Directions

Research Questions:
- Where to get knowledge?
- How to incorporate this knowledge?
- What knowledge is missing?
- Is a certain fact useful?

Introspection:
- Can I answer the main question without additional questions?
- What should I ask about?

Relevance, factual correctness and informativeness

Relevant but unhelpful knowledge may confuse the model!
Interactive QA Paradigm

Future Directions

**Research Questions:**
Where to get knowledge?
How to incorporate this knowledge?

What knowledge is missing?
Is a certain fact useful?

**Introspection:**
- Can I answer the main question without additional questions?
- What should I ask about?

Relevant but *unhelpful* knowledge may confuse the model!

Relevance, factual correctness and informativeness

Defined by people or by the model?
Understanding language through generalizing existing world knowledge

Nonmonotonic reasoning in natural language

Introspective knowledge acquisition through asking questions and reasoning

Future directions

**Back to the Future:**
Unsupervised Backprop-based Decoding for Counterfactual and Abductive Commonsense Reasoning
Lianhui (Karen) Qin, Vered Shwartz, Peter West, Chandra Bhagavatula, Jena Hwang, Ronan Le Bras, Antoine Bosselut, and Yejin Choi.
EMNLP 2020

**Thinking Like a Skeptic:**
Defeasible Inference in Natural Language
Rachel Rudinger, Vered Shwartz, Jena D. Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A. Smith, and Yejin Choi.
Findings of EMNLP 2020

**Learning to Rationalize for Nonmonotonic Reasoning with Distant Supervision**
Faeze Brahman, Vered Shwartz, Rachel Rudinger, and Yejin Choi.
AAAI 2021
Nonmonotonic Reasoning
Nonmonotonic Reasoning

Abductive reasoning

ART (Bhagavatula et al., 2020)

Most plausible explanation
Nonmonotonic Reasoning

Abductive reasoning
ART (Bhagavatula et al., 2020)
Most plausible explanation

Counterfactual reasoning
TimeTravel (Qin et al., 2019)
What if?
Nonmonotonic Reasoning

Abductive reasoning

ART (Bhagavatula et al., 2020)

Most plausible explanation

Counterfactual reasoning

TimeTravel (Qin et al., 2019)

What if?

Defeasible reasoning

δ-NLI (Rudinger et al., 2020)

Updating inferences with additional knowledge
Abductive Reasoning (Peirce, 1965)

Reason about the most plausible explanation for incomplete observations.
Abductive Reasoning (Peirce, 1965)

Reason about the most plausible explanation for incomplete observations.

Sara wanted to make dinner for some guests.

ART (Bhagavatula et al., 2020)
Abductive Reasoning (Peirce, 1965)

Reason about the most plausible explanation for incomplete observations.

Sara wanted to make dinner for some guests. She had to order pizza for her friends instead.

ART (Bhagavatula et al., 2020)
Abductive Reasoning (Peirce, 1965)

Reason about the most plausible explanation for incomplete observations.

Sara wanted to make dinner for some guests.

She had to order pizza for her friends instead.
Abductive Reasoning (Peirce, 1965)

Reason about the most plausible explanation for incomplete observations.

Sara wanted to make dinner for some guests. But she didn’t know how to cook. She had to order pizza for her friends instead.

ART (Bhagavatula et al., 2020)
Abductive Reasoning (Peirce, 1965)

Reason about the most plausible explanation for incomplete observations.

Sara wanted to make dinner for some guests. But she didn’t know how to cook. She had to order pizza for her friends instead.

ART (Bhagavatula et al., 2020)

Useful for filling in gaps in story understanding
Challenge: Language models are conditioned only on a past context

Sara wanted to make dinner for some guests.
Challenge: Language models are conditioned only on a past context

Sara wanted to make dinner for some guests. "I'm going to go grab some rice noodles," she says.
Challenge: Language models are conditioned only on a past context

Sara wanted to make dinner for some guests. "I'm going to go grab some rice noodles," she says.

Solution: compute loss w.r.t future constraints & backpropagate to the output
Challenge: Language models are conditioned only on a past context

Sara wanted to make dinner for some guests. "I'm going to go grab some rice noodles," she says.

Solution: compute loss w.r.t future constraints & backpropagate to the output

Inspiration:

Image Style Transfer (Gatys et al, 2016)
Input

X - past context
Sara wanted to make dinner for some guests.

Z - future constraints
She had to order pizza for her friends instead.
Sara wanted to make dinner for some guests. She had to order pizza for her friends instead.
Sara wanted to make dinner for some guests. She had to order pizza for her friends instead.

**Input**
- **X** - past context
  - Sara wanted to make dinner for some guests.
- **Z** - future constraints
  - She had to order pizza for her friends instead.

**Output**
- **Y** - continuation
  - Fluent continuation of X
  - Satisfies the constraints Z

---

**DeLorean**

**Initialization**

**Backward Pass**

**Forward Pass**
Initialization
Sara wanted to make dinner for some guests. She had to order pizza for her friends instead.
Sara wanted to make dinner for some guests. She had to order pizza for her friends instead.
Sara wanted to make dinner for some guests. She had to order pizza for her friends instead.
Sara wanted to make dinner for some guests. She had to order pizza for her friends instead.

**Input**
- **X** - past context
  Sara wanted to make dinner for some guests.
- **Z** - future constraints
  She had to order pizza for her friends instead.

**Output**
- **Y** - continuation
  - Fluent continuation of X
  - Satisfies the constraints Z

**Initialization**

\[
\hat{Y} = \begin{align*}
&\text{decode N tokens from} \\
&LM_{\text{forward}}(X)
\end{align*}
\]
Backward Pass
Backward Pass
Sara wanted to make dinner for some guests. She had to order pizza for her friends instead.
Sara wanted to make dinner for some guests.

She had to order pizza for her friends instead.
Sara wanted to make dinner for some guests. She had to order pizza for her friends instead.
Sara wanted to make dinner for some guests. She had to order pizza for her friends instead.
Sara wanted to make dinner for some guests.

She had to order pizza for her friends instead.

**Task-specific Loss Function**

Maximize the likelihood of LM to generate the future observation $Z$ following the past observation $X$ and the generated hypothesis $\tilde{Y}$

$$\mathcal{L}(X, \tilde{Y}, Z) := - \sum_{n=1}^{N_Z} \log P_{LM}(\tilde{z}_n | X, \tilde{Y}, Z_{1:n-1})$$
Sara wanted to make dinner for some guests. She had to order pizza for her friends instead.

**Input**
- \(X\) - past context
- \(Z\) - future constraints

**Output**
- \(Y\) - continuation
  - Fluent continuation of \(X\)
  - Satisfies the constraints \(Z\)

**Initialization**
\[
\tilde{Y} = \text{decode N tokens from } LM_{\text{forward}}(X)
\]

**Backward Pass**
Maximize the likelihood of \(LM\) to generate the future observation \(Z\) following the past observation \(X\) and the generated hypothesis \(\tilde{Y}\)

\[
\mathcal{L}(X, \tilde{Y}, Z) := -\sum_{n=1}^{N_Z} \log P_{LM}(z_n | X, \tilde{Y}, Z_{1:n-1})
\]
Forward Pass
Forward Pass
Sara wanted to make dinner for some guests. She had to order pizza for her friends instead.
Sara wanted to make dinner for some guests. She had to order pizza for her friends instead.
Sara wanted to make dinner for some guests. She had to order pizza for her friends instead.

DeLorean

**Input**
- X - past context
- Z - future constraints

**Output**
- Y - continuation
  - Fluent continuation of X
  - Satisfies the constraints Z

**Initialization**

\[ \hat{Y} = \text{decode N tokens from } LM_{\text{forward}}(X) \]

**Backward Pass**

Maximize the likelihood of LM to generate the future observation Z following the past observation X and the generated hypothesis \( \hat{Y} \)

\[ \mathcal{L}(X, \hat{Y}, Z) := - \sum_{n=1}^{N_z} \log P_{LM}(z_n | X, \hat{Y}, Z_{1:n-1}) \]

**Forward Pass**

\[ \tilde{Y} = \text{decode N tokens from } LM_{\text{forward}}(X) \]

And mix with backward logits
Sara wanted to make dinner for some guests.
She had to order pizza for her friends instead.

Initialization
\[ \bar{Y} = \text{decode N tokens from } LM_{\text{forward}}(X) \]

Backward Pass
Maximize the likelihood of LM to generate the future observation \( Z \) following the past observation \( X \) and the generated hypothesis \( \bar{Y} \)
\[ \mathcal{L}(X, \bar{Y}, Z) := - \sum_{n=1}^{N_Z} \log P_{LM}(z_n | X, \bar{Y}, Z^{n-1}) \]

Forward Pass
\[ \bar{Y} = \text{decode N tokens from } LM_{\text{forward}}(X) \]
And mix with backward logits

Output
\( Y \) - continuation
- Fluent continuation of \( X \)
- Satisfies the constraints \( Z \)
Generation
Sara wanted to make dinner for some guests.

She had to order pizza for her friends instead.
Sara wanted to make dinner for some guests. But she didn’t know how to cook. She had to order pizza for her friends instead.
Sara wanted to make dinner for some guests. She had to order pizza for her friends instead.

**Initialization**

\[ \tilde{Y} = \text{decode N tokens from } LM_{\text{forward}}(X) \]

**Backward Pass**

Maximize the likelihood of \( LM \) to generate the future observation \( Z \) following the past observation \( X \) and the generated hypothesis \( \tilde{Y} \)

\[ \mathcal{L}(X, \tilde{Y}, Z) := - \sum_{n=1}^{N_z} \log P_{LM}(\tilde{z}_n | X, \tilde{Y}, Z_{1:n-1}) \]

**Forward Pass**

\[ \tilde{Y} = \text{decode N tokens from } LM_{\text{forward}}(X) \]

And mix with backward logits

**DeLorean**

**Input**

- \( X \) - past context
- \( Z \) - future constraints

**Output**

- \( Y \) - continuation
  - Fluent continuation of \( X \)
  - Satisfies the constraints \( Z \)

Greedy decoding from \( \tilde{Y} \)
Sara wanted to make dinner for some guests. She had to order pizza for her friends instead.
Sara wanted to make dinner for some guests.
She had to order pizza for her friends instead.
Sara wanted to make dinner for some guests. She had to order pizza for her friends instead.
Select Best Y
Select Best Y
Select Best $Y$

Select $Y^{(i)}$ that is most likely to follow and precede its adjacent sentences
Select Best $Y$

Select $Y^{(i)}$ that is most likely to follow and precede its adjacent sentences

$$\text{score}(Y^{(i)}) = \text{BERT}_{\text{NSP}}(XY^{(i)}, Z) + \text{BERT}_{\text{NSP}}(X, Y^{(i)}Z)$$

P(She had to order pizza for her friends instead. | Sara wanted to make dinner for some guests. But she didn’t know how to cook.)

P(But she didn’t know how to cook. She had to order pizza for her friends instead. | Sara wanted to make dinner for some guests.)
Human Evaluation Results

Abductive Reasoning
Human Evaluation Results

Abductive Reasoning

Coherence

0
2.5
5
7.5
10

X-Y
Y-Z
XY-Z

DELOREAN
Unsupervised
Supervised
Human
Human Evaluation Results

Abductive Reasoning

Coherence

<table>
<thead>
<tr>
<th></th>
<th>DELOREAN</th>
<th>Unsupervised</th>
<th>Supervised</th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>XY</td>
<td>5.22</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YZ</td>
<td>5.25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XYZ</td>
<td>2.97</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Human Evaluation Results

Abductive Reasoning

- X-Y
- Y-Z
- X-Y-Z

1. Outperforms unsupervised models substantially
Human Evaluation Results

Abductive Reasoning

1. Outperforms unsupervised models substantially
2. Competitive with supervised models!
Human Evaluation Results

Abductive Reasoning

1. Outperforms unsupervised models substantially
2. Competitive with supervised models!
3. Large gap from human performance
Example Generations

Sara wanted to make dinner for some guests.

She had to order pizza for her friends instead.
Example Generations

Sara wanted to make dinner for some guests.

1. She was thinking about the best way.

She had to order pizza for her friends instead.
Example Generations

Sara wanted to make dinner for some guests.

1. She was thinking about the best way.
2. However, her cooking skills were the only thing that could make it a success.

Backward pass introduces: contrast!

She had to order pizza for her friends instead.
Example Generations

Sara wanted to make dinner for some guests.

1. She was thinking about the best way.
2. **However**, her cooking skills were the only thing that could make it a success.
3. **But** she couldn’t, because she was too busy with her work.

**Backward pass introduces: contrast!**

She had to order pizza for her friends instead.
Example Generations

Sara wanted to make dinner for some guests.

1. She was thinking about the best way.
2. **However**, her cooking skills were the only thing that could make it a success.
3. **But** she couldn’t, because she was too busy with her work.
4. **But** she didn’t have the money and she didn’t have her own kitchen.

Backward pass introduces: contrast!

She had to order pizza for her friends instead.
Sara wanted to make dinner for some guests.

1. She was thinking about the best way.
2. **However**, her cooking skills were the only thing that could make it a success.
3. **But** she couldn’t, because she was too busy with her work.
4. **But** she didn’t have the money and she didn’t have her own kitchen.
5. **But** she didn’t know how to cook.

Backward pass introduces: contrast!

She had to order pizza for her friends instead.
Counterfactual Reasoning (Goodman, 1947)

Reason about changes in outcomes given a change in conditions.
Counterfactual Reasoning (Goodman, 1947)

Reason about changes in outcomes given a change in conditions.

Useful for Argument Mining: If X would have happened, it would result in some unwanted outcome Y.
Counterfactual Reasoning (Goodman, 1947)

Reason about changes in outcomes given a change in conditions.

Useful for **Argument Mining:**
If $X$ would have happened, it would result in some unwanted outcome $Y$.

Useful for **Detecting Misinformation:**
Claim $X$ is false because it entails claim $Y$ which is known to be false.
Time Travel (Qin et al., 2019)

Original Story

Lisa was throwing a Halloween party.
All her friends were dressing up.
Lisa thought about being a wizard.
Then she decided on a scarier costume.
Lisa dressed up like a vampire.
Original Story
Lisa was throwing a Halloween party.
All her friends were dressing up.
Lisa thought about being a wizard.
Then she decided on a scarier costume.
Lisa dressed up like a vampire.

Counterfactual Beginning
Lisa was throwing a Halloween party.
All her friends were dressing up. It was a Game of Thrones themed party.
Time Travel (Qin et al., 2019)

Original Story
Lisa was throwing a Halloween party. All her friends were dressing up. Lisa thought about being a wizard. Then she decided on a scarier costume. Lisa dressed up like a vampire.

Counterfactual Beginning
Lisa was throwing a Halloween party. All her friends were dressing up. It was a Game of Thrones themed party.

Alternative Ending:
1. Adheres to the counterfactual beginning
2. Minimally edits the original ending
TimeTravel (Qin et al., 2019)

Original Story
Lisa was throwing a Halloween party.
All her friends were dressing up.
Lisa thought about being a wizard.
Then she decided on a scarier costume.
Lisa dressed up like a vampire.

Counterfactual Beginning
Lisa was throwing a Halloween party.
All her friends were dressing up. It was a Game of Thrones themed party.

Alternative Ending:
1. Adheres to the counterfactual beginning
2. Minimally edits the original ending

Lisa thought about being a wizard how she would dress up as a Lannister, but she didn’t want to look like a Lannister.
Then she decided on a scarier costume. She wanted to look like a Stark.
Lisa dressed up like a vampire Stark.
Counterfactual Reasoning

Input
X - counterfactual beginning
Lisa was throwing a Halloween party. All her friends were dressing up. It was a Game of Thrones themed party.

Z - original ending
Lisa thought about being a wizard. Then she decided on a scarier costume. Lisa dressed up like a vampire.

Output
Y - alternative ending
1. Adheres to the counterfactual story beginning
2. Minimally edits the original ending
**Counterfactual Reasoning**

**Input**

X - counterfactual beginning
Lisa was throwing a Halloween party. 
Underline: All her friends were dressing up.
It was a Game of Thrones themed party.

Z - original ending
Lisa thought about being a wizard.
Then she decided on a scarier costume.
Lisa dressed up like a vampire.

**Initialization**

**Output**

Y - alternative ending

1. Adheres to the counterfactual story beginning
2. Minimally edits the original ending
Lisa was throwing a Halloween party. All her friends were dressing up. It was a Game of Thrones themed party.

Input
X - counterfactual beginning
Lisa thought about being a wizard. Then she decided on a scarier costume. Lisa dressed up like a vampire.

Z - original ending
Lisa thought about being a wizard. Then she decided on a scarier costume. Lisa dressed up like a vampire.

Output
Y - alternative ending
1 Adheres to the counterfactual story beginning
2 Minimally edits the original ending

Backward Pass
Minimize the KL divergence between the original ending Z (one-hot representation) and generated ending \( \tilde{Y} \)

\[ \mathcal{L}(X, \tilde{Y}, Z) := KL (Z \| \text{softmax}(\tilde{Y}/\tau)) \]
Counterfactual Reasoning

**Input**
X - counterfactual beginning
Lisa was throwing a Halloween party.
*All her friends were dressing up.*
It was a Game of Thrones themed party.

Z - original ending
Lisa thought about being a wizard.
Then she decided on a scarier costume.
Lisa dressed up like a vampire.

**Output**
Y - alternative ending
1. Adheres to the counterfactual story beginning
2. Minimally edits the original ending

**Initialization**

**Backward Pass**
Minimize the KL divergence between the original ending \( Z \) (one-hot representation) and generated ending \( \tilde{Y} \)
\[
\mathcal{L}(X, \tilde{Y}, Z) := \text{KL}(Z || \text{softmax}(\tilde{Y}/\tau))
\]

**Forward Pass**
Generation + Select best \( Y \)
Counterfactual Reasoning

**Input**

X - counterfactual beginning
Lisa was throwing a Halloween party. All her friends were dressing up. It was a Game of Thrones themed party.

Z - original ending
Lisa thought about being a wizard. Then she decided on a scarier costume. Lisa dressed up like a vampire.

**Output**

Y - alternative ending

1. Adheres to the counterfactual story beginning
2. Minimally edits the original ending

**Backward Pass**

Minimize the KL divergence between the original ending $Z$ (one-hot representation) and generated ending $\tilde{Y}$

$$\mathcal{L}(X, \tilde{Y}, Z) := KL(Z||\text{softmax}(\tilde{Y}/\tau))$$

**Forward Pass**

Select best $Y$

DeLorean was the only method to achieve a good balance between the two requirements.
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 \textbf{(Reiter, 1980)}

Given premise $P$, a hypothesis $H$ is \textbf{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.
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.
- 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.

P: Tweety is a bird.
H: Tweety flies.
U: Tweety is a penguin.
Defeasible Inference in Natural Language

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 **strengthen**er.

P: Tweety is a bird.

H: Tweety flies.

Weakener: Tweety is a penguin.
An update $U$ is called a **weakened** 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.
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.
  - They are in a conference room. \( \rightarrow \) \( + \) Strengthener
  - They are in a library. \( \rightarrow \) \( - \) Weaken

**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. \( \rightarrow \) \( + \)
  - They are in a library. \( \rightarrow \) \( - \)
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.**
    - Strengthener
  - **They are in a library.**
    - Weaken

**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.**
    - +
  - **They are in a library.**
    - -

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

Learning to Rationalize for Nonmonotonic Reasoning with Distant Supervision

Faeze Brahman, Vered Shwartz, Rachel Rudinger, and Yejin Choi.
AAAI 2021
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 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.

Learning to Rationalize for Nonmonotonic Reasoning with Distant Supervision

Faeze Brahman, Vered Shwartz, Rachel Rudinger, and Yejin Choi.

AAAI 2021
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 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:

Learning to Rationalize for Nonmonotonic Reasoning with Distant Supervision
Faeze Brahman, Vered Shwartz, Rachel Rudinger, and Yejin Choi.
AAAI 2021
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 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:

The definition of a library is...

Learning to Rationalize for Nonmonotonic Reasoning with Distant Supervision

Faeze Brahman, Vered Shwartz, Rachel Rudinger, and Yejin Choi. AAAI 2021
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 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.
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.

**A conference room is where people have meetings at work.**

They are in a conference room.

**They are in a library.**

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.

**Trivially rephrasing the label! ("[+] implies that [H]")**

**Joint Prediction & Rationalization**  Predict the label (strengthen / weaken) 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.

**A conference room is where people have meetings at work.**

They are in a conference room.

**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 (strengthen / 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!
Introspective knowledge acquisition through asking questions and reasoning

Nonmonotonic reasoning in natural language

Understanding language through generalizing existing world knowledge

Future directions

Paraphrase to Explicate: Revealing Implicit Noun Compound Relations

Vered Shwartz and Ido Dagan
ACL 2018
Implicit Meaning in Noun Compounds
Implicit Meaning in Noun Compounds

oil made of olives
Implicit Meaning in Noun Compounds

Oil made of olives

If olive oil is made from olives...

Then that must mean baby oil is made from...

Olive Oil

QuickMeme.com
Implicit Meaning in Noun Compounds

Oil made of olives

Oil used for babies
Implicit Meaning in Noun Compounds

oil made of olives

oil used for babies

<table>
<thead>
<tr>
<th>ENGLISH - DETECTED</th>
<th>HEBREW</th>
<th>ITALIAN</th>
<th>HEBREW</th>
<th>ENGLISH</th>
</tr>
</thead>
<tbody>
<tr>
<td>olive oil</td>
<td></td>
<td>olio d'oliva = oil from olives</td>
<td></td>
<td></td>
</tr>
<tr>
<td>baby oil</td>
<td></td>
<td>olio per bambini = oil for babies</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Humans Easily Interpret Implicit Meaning

parsley cake
Humans Easily Interpret Implicit Meaning

I've baked a parsley cake, would you like a piece?
Humans Easily Interpret Implicit Meaning

I’ve baked a parsley cake, would you like a piece?

Cakes can be made from fruit.

Parsley is an herb.

Herbs are similar to fruit.

Cake made of parsley!
Humans Easily Interpret Implicit Meaning

I've baked a parsley cake, would you like a piece?

Hmm... no, thanks.

Parsley is an herb.

Herbs are similar to fruit.

Cake made of parsley!
parsley cake

I've baked a parsley cake, would you like a piece?
I've baked a parsley cake, would you like a piece?

Yes, please!
Noun Compound Paraphrasing

Produce a ranked list of paraphrases expressing the relationship between the constituents:

- oil used for babies
- oil for babies
- oil made for babies
- oil made of olives
- oil from olives
- oil extracted from olives

(Nakov and Hearst, 2006; Hendrickx et al., 2015)
Multi-task Learning for Noun Compound Paraphrasing

Input
\( w_1 = \text{olive}, w_2 = \text{oil} \)

Paraphrasing Model

Output
\{ [w_2] \text{ made of } [w_1], [w_2] \text{ extracted from } [w_1], \ldots \}
Paraphrasing Model
Paraphrasing Model
Paraphrasing Model

Instance: \(<w_1 = \text{olive}, w_2 = \text{oil}, p = [w_2] \text{ made of } [w_1]>\)
Paraphrasing Model

Instance: \(<w_1 = \textit{olive}, w_2 = \textit{oil}, p = [w_2] \text{ made of } [w_1]>\)

Main:

1. What is the relation between \textit{olive} and \textit{oil}?
Paraphrasing Model

Instance: $<w_1 = \text{olive}, w_2 = \text{oil}, p = [w_2] \text{ made of } [w_1]>$

Main:
1. What is the relation between olive and oil?
Paraphrasing Model

Instance: \(<w_1 = \text{oil}, \ w_2 = \text{oil}, \ p = [w_2] \text{ made of } [w_1]\)>

Main:

1. What is the relation between \text{oil} and \text{oil}?
Paraphrasing Model

Instance: \( w_1 = \text{olive}, w_2 = \text{oil}, p = [w_2] \text{ made of } [w_1] \)

Main:

1. What is the relation between \text{olive} and \text{oil}?
Paraphrasing Model

Instance: \(<w_1 = \text{olive}, w_2 = \text{oil}, p = [w_2] \text{ made of } [w_1]\)>

Main:

What is the relation between \text{olive} and \text{oil}?
Paraphrasing Model

Instance: \(<w_1 = \text{olive}, w_2 = \text{oil}, p = [w_2] \text{ made of } [w_1]\)>

Main:

1. What is the relation between \text{olive} and \text{oil}?
What is the relation between *olive* and *oil*?
What is the relation between olive and oil?

Paraphrasing Model

Instance: \(<w_1 = \text{olive}, w_2 = \text{oil}, p = [w_2] \text{ made of } [w_1]\)>

Main:

1. What is the relation between \text{olive} and \text{oil}?
What is the relation between olive and oil?

Instance: \( w_1 = \text{olive}, w_2 = \text{oil}, p = [w_2 \text{ made of } [w_1]] \)

Loss: \( \text{CE}(\hat{p}, p) \)
Paraphrasing Model

Instance: $<w_1 = {\textit{olive}}, w_2 = \textit{oil}, p = [w_2] \quad \text{made of} \quad [w_1]>$

Loss: $\text{CE}(\hat{p}, p)$

Main:

1. What is the relation between $\textit{olive}$ and $\textit{oil}$?

Auxiliary:

2. What can $\textit{oil}$ be made of? /
Paraphrasing Model

Instance: \(<w_1 = \text{olive}, w_2 = \text{oil}, p = [w_2] \text{ made of } [w_1]\)>
Loss: \(CE(\hat{p}, p)\)

Main:

1. What is the relation between \text{olive} and \text{oil}?

[Diagram showing the relationship between \text{olive}, \text{oil}, and \(\hat{p} = 78\) predicted by the MLP model.]

Auxiliary:

2. What can \text{oil} be made of? /
3. What can be made of \text{olive}?
Paraphrasing Model

Instance: \(<w_1 = \text{olive}, w_2 = \text{oil}, p = [w_2] \text{ made of } [w_1] >\)
Loss: \(CE(\hat{p}, p)\)

Main:
1. What is the relation between \textit{olive} and \textit{oil}?

Auxiliary:
2. What can \textit{oil} be made of? /
3. What can be made of \textit{olive}?

(28) made
(121) olive
(712) oil
(10) of

(1) \([w_1]\)
(2) \([w_2]\)
(3) \([p]\)

(78) \([w_2]\) containing \([w_1]\)
...

(28) made
(121) olive
(712) oil
(10) of

(1) \([w_1]\)
(2) \([w_2]\)
(3) \([p]\)
Paraphrasing Model

Instance: \(<w_1 =\textit{olive}, w_2 = \textit{oil}, p = [w_2] \text{ made of } [w_1]>\)

Loss: \(CE(\hat{p}, p) + CE(\hat{w}_1, w_1) + CE(\hat{w}_2, w_2)\)

**Main:**

1. What is the relation between \textit{olive} and \textit{oil}?

   \(\hat{p} = 78\ldots\)

   \(\text{MLP}_p,\)

   \(\text{oil}\)

   \([p]\)

   \(\text{olive}\)

   (28) made

   (121) olive

   (712) oil

   …

   (10) of

   (78) \([w_2]\) containing \([w_1]\)

   …

   (131) \([w_2]\) made of \([w_1]\)

   …

   (11) \([p]\)

**Auxiliary:**

2. What can \textit{oil} be made of? /

3. What can be made of \textit{olive}?

   \(\hat{w}_1 = 121\)

   \(\text{MLP}_w\)

   \(\text{oil}\)

   \(\text{made}\)

   \(\text{of}\)

   \([w_1]\)

   (28) made

   (121) olive

   (712) oil

   …

   (10) of

   (1) \([w_1]\)

   (2) \([w_2]\)

   (3) \([p]\)
Multi-task Learning for Noun Compound Paraphrasing

Instance: \(<w_1 = \text{olive}, w_2 = \text{oil}, p = [w_2] \text{ made of } [w_1]\)>

Loss: \(\text{CE}(\hat{p}, p) + \text{CE}(\hat{w}_1, w_1) + \text{CE}(\hat{w}_2, w_2)\)

Main:

1. What is the relation between \text{olive} and \text{oil}?

Generalizes similar noun compounds: \text{avocado oil} expected to predict similar paraphrases
Multi-task Learning for Noun Compound Paraphrasing

Instance: \(<w_1 = \text{oil}, w_2 = \text{oil}, p = \text{made of } [w_1]\)>
Loss: \(\text{CE}(\hat{p}, p) + \text{CE}(\hat{w}_1, w_1) + \text{CE}(\hat{w}_2, w_2)\)

2 Generalizes similar paraphrases:
\([w_2]\) from \([w_1]\) expected to predict the same constituent nouns

Auxiliary:
2 What can \text{oil} be made of? /
3 What can be made of \text{olive}?
Multi-task Learning for Noun Compound Paraphrasing

Instance: \( <w_1 = \text{olive}, w_2 = \text{oil}, p = [w_2] \text{ made of } [w_1]> \)

Loss: \( \text{CE}(\hat{p}, p) + \text{CE}(\hat{w}_1, w_1) + \text{CE}(\hat{w}_2, w_2) \)

Main:
1. What is the relation between \text{olive} and \text{oil}?

Auxiliary:
2. What can \text{oil} be made of? /
3. What can be made of \text{olive}?

Self-supervised with corpus co-occurrences!

\[ \text{MLP}_p \]

\[ \text{MLP}_w \]

\[ \hat{p} = 78 \]

\[ \hat{w}_1 = 121 \]
Multi-task Learning for Noun Compound Paraphrasing

Input
\( w_1 = \text{olive}, w_2 = \text{oil} \)

Output
\{ [w_2] \text{ made of } [w_1], [w_2] \text{ extracted from } [w_1], \ldots \}
Evaluation Results

SemEval 2015 Task 4 (Hendrickx et al., 2015)
Evaluation Results
SemEval 2013 Task 4 (Hendrickx et al., 2013)

Isomorphic Matching Score

100
75
50
25
0

MELODI  Prep Baseline  SFS  IIITH  MTL
Evaluation Results

SemEval 2013 Task 4 (Hendrickx et al., 2013)

Isomorphic Matching Score

<table>
<thead>
<tr>
<th>Method</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>MELODI</td>
<td>15</td>
</tr>
<tr>
<td>Prep Baseline</td>
<td>15.8</td>
</tr>
<tr>
<td>SFS</td>
<td>25.1</td>
</tr>
<tr>
<td>IIITH</td>
<td>25.8</td>
</tr>
<tr>
<td>MTL</td>
<td>83</td>
</tr>
</tbody>
</table>
Evaluation Results
SemEval 2015 Task 4 (Hendrickx et al., 2015)
Interpreting Non Compositional Noun Compounds
Interpreting Non Compositional Noun Compounds

Furious Meghan Markle says she won’t fall for dad’s crocodile tears after he claimed ‘she’d be better off if he were dead’
1. Introspective knowledge acquisition through asking questions and reasoning
2. Nonmonotonic reasoning in natural language
3. Understanding language through generalizing existing world knowledge
4. Future directions
Future Directions

Language

Knowledge

Reasoning
Future Directions

Language
Meaningful Text Representations
Reading between the Lines

Knowledge

Reasoning
Future Directions

Language
Meaningful Text Representations
Reading between the Lines

Knowledge
Acquiring Commonsense Knowledge

Reasoning
Future Directions

Language
Meaningful Text Representations
Reading between the Lines

Knowledge
Acquiring Commonsense Knowledge

Reasoning
Causal & Nonmonotonic Reasoning
Future Directions

Language
Meaningful Text Representations
Reading between the Lines

Knowledge
Acquiring Commonsense Knowledge

Reasoning
Causal & Nonmonotonic Reasoning

Reliable Evaluation
Reliable Evaluation

Language

Meaningful Text Representations
Reading between the Lines

Knowledge

Acquiring Commonsense Knowledge

Reasoning

Causal & Nonmonotonic Reasoning
Acquiring Commonsense Knowledge

Language
Meaningful Text Representations
Reading between the Lines

Knowledge

Reasoning
Causal & Nonmonotonic Reasoning
Acquiring Commonsense Knowledge

1 from people

2 from text
Acquiring Commonsense Knowledge

1. from people
2. from text

✗ Impossible to manually enumerate

$\text{\$\$\$}$
Acquiring Commonsense Knowledge

1. from people
   - Impossible to manually enumerate

2. from text
   - Reporting bias
     (Gordon and Van Durme, 2013)

- murdered + killed
- breathed + exhaled + inhaled
Acquiring Commonsense Knowledge

1. from people
   - Impossible to manually enumerate

2. from text
   - Reporting bias
     (Gordon and Van Durme, 2013)

3. from large-scale neural language models

- murdered + killed
- breathed + exhaled + inhaled
Acquiring Commonsense Knowledge from Large-scale Neural Language models
Acquiring Commonsense Knowledge from Large-scale Neural Language models

✓ Capture facts not explicitly mentioned in the corpus
(Petroni et al. 2019; Feldman et al. 2019, this talk)
Acquiring Commonsense Knowledge from Large-scale Neural Language models

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

- Not sensitive to negation (Kassner et al. 2020; Ettinger, 2020)
Acquiring Commonsense Knowledge from Large-scale Neural Language models

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

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

✗ Don’t differentiate constant vs. contingent facts

Zebras are black and white. My shirt is blue / red.
Acquiring Commonsense Knowledge from Large-scale Neural Language models

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

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

- Don’t differentiate constant vs. contingent facts

Don’t differentiate generic facts from grounded knowledge about named entities

"You are grounded!": Latent Name Artifacts in Pre-trained Language Models. Vered Shwartz, Rachel Rudinger, and Oyvind Tafjord. EMNLP 2020.
Acquiring Commonsense Knowledge from Large-scale Neural Language models

- Capture facts not explicitly mentioned in the corpus (Petroni et al. 2019; Feldman et al. 2019, this talk)
- Not sensitive to negation (Kassner et al. 2020; Ettinger, 2020)
- Don’t differentiate constant vs. contingent facts

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

Don’t differentiate generic facts from grounded knowledge about named entities

Richard has a bad

"You are grounded!": Latent Name Artifacts in Pre-trained Language Models. Vered Shwartz, Rachel Rudinger, and Oyvind Tafjord. EMNLP 2020.
Acquiring Commonsense Knowledge from Large-scale Neural Language models

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

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

- Don’t differentiate constant vs. contingent facts

Richard has a bad habit of saying things that are not true.

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

"You are grounded!": Latent Name Artifacts in Pre-trained Language Models. Vered Shwartz, Rachel Rudinger, and Oyvind Tafjord. EMNLP 2020.
Acquiring Commonsense Knowledge from Large-scale Neural Language models

- Capture facts not explicitly mentioned in the corpus (Petroni et al. 2019; Feldman et al. 2019, this talk)
- Not sensitive to negation (Kassner et al. 2020; Ettinger, 2020)
- Don’t differentiate constant vs. contingent facts

Don’t differentiate generic facts from grounded knowledge about named entities

Richard has a bad habit of saying things that are not true.

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

"You are grounded!": Latent Name Artifacts in Pre-trained Language Models. Vered Shwartz, Rachel Rudinger, and Oyvind Tafjord. EMNLP 2020.
Acquiring Commonsense Knowledge from Large-scale Neural Language models

- Capture facts not explicitly mentioned in the corpus ([Petroni et al. 2019; Feldman et al. 2019, this talk)]
- Not sensitive to negation ([Kassner et al. 2020; Ettinger, 2020])
- Don’t differentiate constant vs. contingent facts

Don’t differentiate generic facts from grounded knowledge about named entities

Richard has a bad habit of saying things that are not true.

Donald has a bad

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

"You are grounded!": Latent Name Artifacts in Pre-trained Language Models. Vered Shwartz, Rachel Rudinger, and Oyvind Tafjord. EMNLP 2020.
Acquiring Commonsense Knowledge from Large-scale Neural Language models

- Capture facts not explicitly mentioned in the corpus (Petroni et al. 2019; Feldman et al. 2019, this talk)
- Not sensitive to negation (Kassner et al. 2020; Ettinger, 2020)
- Don’t differentiate constant vs. contingent facts

- Don’t differentiate generic facts from grounded knowledge about named entities

Acquiring Commonsense Knowledge from Large-scale Neural Language models

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

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

- Don't differentiate generic facts from grounded knowledge about named entities

- Don't differentiate constant vs. contingent facts

"You are grounded!": Latent Name Artifacts in Pre-trained Language Models. Vered Shwartz, Rachel Rudinger, and Oyvind Tafjord. EMNLP 2020.
Acquiring Commonsense Knowledge from Large-scale Neural Language models

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

- Don’t differentiate constant vs. contingent facts

- Don’t differentiate generic facts from grounded knowledge about named entities

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

- Don’t completely overcome reporting bias

"You are grounded!": Latent Name Artifacts in Pre-trained Language Models. Vered Shwartz, Rachel Rudinger, and Oyvind Tafjord. EMNLP 2020.

Acquiring Commonsense Knowledge from Large-scale Neural Language Models

- Capture facts not explicitly mentioned in the corpus (Petroni et al. 2019; Feldman et al. 2019, this talk)
- Not sensitive to negation (Kassner et al. 2020; Ettinger, 2020)
- Don’t differentiate constant vs. contingent facts

- Don’t differentiate generic facts from grounded knowledge about named entities

- Don’t completely overcome reporting bias

"You are grounded!": Latent Name Artifacts in Pre-trained Language Models. Vered Shwartz, Rachel Rudinger, and Oyvind Tafjord. EMNLP 2020.

Acquiring Commonsense Knowledge from Large-scale Neural Language models

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

- Don't completely overcome reporting bias
  \[\text{The man turned on the faucet. As a result,}\]
  \[\text{the man's blood was sprayed everywhere.}\]

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

- Don't differentiate constant vs. contingent facts
  \[\text{Zebras are black and white. My shirt is blue / red.}\]

- Don't differentiate generic facts from grounded knowledge about named entities
  \[\begin{array}{ll}
  \text{Richard has a bad habit of saying things that are not true.} \\
  \text{Donald has a bad reputation for being a racist.}
  \end{array}\]

"You are grounded!": Latent Name Artifacts in Pre-trained Language Models. Vered Shwartz, Rachel Rudinger, and Oyvind Tafjord. EMNLP 2020.

Does GPT-3 have commonsense?

Examples by Gary Marcus. GPT-3 is not publicly available.
Does GPT-3 have commonsense?

You need flour to bake bread.
You have a sack of flour in the garage.
When you get there, you find that it got thoroughly soaked in a heavy rain last night.

So you have to
Does GPT-3 have commonsense?

You need flour to bake bread.
You have a sack of flour in the garage.
When you get there, you find that it got thoroughly soaked in a heavy rain last night.

So you have to dry it out before you can use it. You can do this by spreading it out on a table and putting a fan on it.

Examples by Gary Marcus. GPT-3 is not publicly available.
Does GPT-3 have commonsense?

You need flour to bake bread.
You have a sack of flour in the garage.
When you get there, you find that it got thoroughly soaked in a heavy rain last night.

So you have to **dry it out before you can use it. You can do this by spreading it out on a table and putting a fan on it.**

Examples by Gary Marcus. GPT-3 is not publicly available.
Does GPT-3 have commonsense?

You need flour to bake bread.
You have a sack of flour in the garage.
When you get there, you find that it got thoroughly soaked in a heavy rain last night.

So you have to dry it out before you can use it. You can do this by spreading it out on a table and putting a fan on it.

A cat was sitting next to a mouse hole in the kitchen wall waiting for the mouse, but it was too cautious, so...
Does GPT-3 have commonsense?

You need flour to bake bread.
You have a sack of flour in the garage.
When you get there, you find that it got thoroughly soaked in a heavy rain last night.

So you have to **dry it out before you can use it. You can do this by spreading it out on a table and putting a fan on it.**

A cat was sitting next to a mouse hole in the kitchen wall waiting for the mouse,
but it was too cautious, so **the cat was waiting for a long time. The cat was getting hungry, so it was thinking of going to the market to buy some food.**

Examples by Gary Marcus. GPT-3 is not publicly available.
Does GPT-3 have commonsense?

You need flour to bake bread.
You have a sack of flour in the garage.
When you get there, you find that it got thoroughly soaked in a heavy rain last night.

So you have to **dry it out before you can use it. You can do this by spreading it out on a table and putting a fan on it.**

A cat was sitting next to a mouse hole in the kitchen wall waiting for the mouse,
but it was too cautious, so **the cat was waiting for a long time. The cat was getting hungry, so it was thinking of going to the market to buy some food.**

Examples by Gary Marcus. GPT-3 is not publicly available.
Does GPT-3 have commonsense?

Examples by Gary Marcus. GPT-3 is not publicly available.

- You need flour to bake bread.
  - You have a sack of flour in the garage.
  - When you get there, you find that it got thoroughly soaked in a heavy rain last night.
  - So you have to dry it out before you can use it. You can do this by spreading it out on a table and putting a fan on it.

- A cat was sitting next to a mouse hole in the kitchen wall waiting for the mouse, but it was too cautious, so the cat was getting hungry, so it was thinking of going to the market to buy some food. 

---

<table>
<thead>
<tr>
<th>Rank</th>
<th>Name</th>
<th>Model</th>
<th>URL Score</th>
<th>BoolQ</th>
<th>CB COPA</th>
<th>MultiRC</th>
<th>ReCoRD</th>
<th>RTE</th>
<th>WIC</th>
<th>WSC</th>
<th>AX-b</th>
<th>AX-g</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DeBERTa Team - Microsoft</td>
<td>DeBERTa / TuringNLRv4</td>
<td>93.3</td>
<td>90.4</td>
<td>95.7/97.6</td>
<td>98.4/88.2/83.7</td>
<td>94.5/94.1</td>
<td>93.2</td>
<td>77.5</td>
<td>95.9</td>
<td>66.7</td>
<td>93.3/93.8</td>
</tr>
<tr>
<td>2</td>
<td>Zinui Wang</td>
<td>T5 + Meena, Single Model (Meena Team - Google Brain)</td>
<td>90.2</td>
<td>91.3/95.8/97.6</td>
<td>97.4/86.3/83.0</td>
<td>94.2/93.5</td>
<td>92.7</td>
<td>77.9</td>
<td>95.9</td>
<td>66.5</td>
<td>86.8/89.9</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>SuperGLUE Human Baselines</td>
<td>SuperGLUE Human Baselines</td>
<td>89.6</td>
<td>89.9/85.8/89.9</td>
<td>100.0/91.8/91.9</td>
<td>91.7/91.3</td>
<td>93.6</td>
<td>80.0</td>
<td>100.0</td>
<td>76.0</td>
<td>99.3/99.7</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>5 T5 Team - Google</td>
<td>T5</td>
<td>89.3</td>
<td>91.2/93.3/96.8</td>
<td>94.8/88.1/86.3</td>
<td>94.1/80.4</td>
<td>92.5</td>
<td>76.9</td>
<td>90.8</td>
<td>65.9</td>
<td>92.7/91.9</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Huawei Noah's Ark Lab</td>
<td>NIEZHA-Plus</td>
<td>88.7</td>
<td>87.8/94.4/98.0</td>
<td>93.9/84.6/55.1</td>
<td>90.1/89.6</td>
<td>89.1</td>
<td>74.8</td>
<td>90.2</td>
<td>58.0</td>
<td>87.1/74.4</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Alibaba PAI/ICBU</td>
<td>PAI Albert</td>
<td>88.1</td>
<td>88.1/92.4/98.4</td>
<td>91.8/84.6/54.7</td>
<td>89.0/88.3</td>
<td>88.8</td>
<td>74.1</td>
<td>90.2</td>
<td>75.6</td>
<td>98.3/99.2</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Tencent Jarvis Lab</td>
<td>RoBERTa (ensemble)</td>
<td>85.9</td>
<td>86.2/92.5/95.0</td>
<td>90.8/84.4/53.4</td>
<td>91.5/91.0</td>
<td>87.9</td>
<td>74.1</td>
<td>91.0</td>
<td>57.0</td>
<td>89.3/75.6</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Zhuyi Technology</td>
<td>RoBERTa-mtl-tech</td>
<td>85.7</td>
<td>87.1/92.4/95.0</td>
<td>91.2/85.1/54.3</td>
<td>91.7/81.3</td>
<td>88.1</td>
<td>72.1</td>
<td>91.0</td>
<td>58.3</td>
<td>91.0/76.1</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Infosys : DAWN : AI Research RoBERTa-ICETS</td>
<td>85.0</td>
<td>86.2/93.2/95.2</td>
<td>91.2/84.6/53.4</td>
<td>89.9/88.3</td>
<td>88.5</td>
<td>72.1</td>
<td>90.0</td>
<td>35.0</td>
<td>93.8/68.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Facebook AI</td>
<td>RoBERTa</td>
<td>84.6</td>
<td>87.1/90.5/93.5</td>
<td>90.6/84.4/52.5</td>
<td>90.8/86.0</td>
<td>88.2</td>
<td>69.3</td>
<td>89.0</td>
<td>87.0</td>
<td>91.0/78.1</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Thilo Schlick</td>
<td>IPET (ALBERT) - Few-Shot (32 Examples)</td>
<td>75.4</td>
<td>81.2/79.9/88.8</td>
<td>89.7/84.1/31.7</td>
<td>85.9/88.4</td>
<td>70.8</td>
<td>49.3</td>
<td>88.4</td>
<td>36.2</td>
<td>97.8/57.9</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Adrian de Wijter</td>
<td>Bort (Alexa AI)</td>
<td>74.1</td>
<td>83.7/81.9/86.4</td>
<td>89.6/83.7/54.4</td>
<td>49.8/49.0</td>
<td>81.2</td>
<td>70.1</td>
<td>65.8</td>
<td>48.0</td>
<td>96.1/61.5</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>IBM Research AI</td>
<td>BERT-ml</td>
<td>73.5</td>
<td>84.8/89.6/94.0</td>
<td>73.6/73.2/30.5</td>
<td>74.6/74.0</td>
<td>84.1</td>
<td>60.2</td>
<td>61.0</td>
<td>29.0</td>
<td>97.8/57.3</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Ben Mann</td>
<td>GPT-3 few-shot - OpenAI</td>
<td>71.9</td>
<td>76.4/52.0/75.6</td>
<td>92.0/75.4/30.5</td>
<td>91.1/95.2</td>
<td>69.0</td>
<td>49.4</td>
<td>80.1</td>
<td>21.1</td>
<td>96.4/55.3</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>SuperGLUE Baselines</td>
<td>BERTa++</td>
<td>71.5</td>
<td>75.0/84.8/90.4</td>
<td>73.8/71.9/71.3</td>
<td>73.0</td>
<td>69.6</td>
<td>64.4</td>
<td>38.0</td>
<td>93.4/51.4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Acquiring Commonsense Knowledge

Learning Commonsense Knowledge from Text
Acquiring Commonsense Knowledge

Learning Commonsense Knowledge from Text

Reporting Bias

from Text, Images and Videos
Acquiring Commonsense Knowledge

Learning Commonsense Knowledge from Text

from Text, Images and Videos

last row ⇒ standing

front row ⇒ cross-legged
Acquiring Commonsense Knowledge

Learning Commonsense Knowledge from Text

<<

from Text, Images and Videos

Hanging up the phone without saying goodbye

Reporting Bias!
Acquiring Commonsense Knowledge

Learning Commonsense Knowledge from Text

Reporting Bias!

Hanging up the phone without saying goodbye

Reporting Bias!
Language

Meaningful Text Representations
Reading between the Lines

Knowledge

Acquiring Commonsense Knowledge

Reasoning

Causal & Nonmonotonic Reasoning

Reliable Evaluation
Meaningful Text Representations

Language

Knowledge

Reasoning

Reading between the Lines

Acquiring Commonsense Knowledge

Causal & Nonmonotonic Reasoning
Meaningful Text Representations

Relatedness

Meaningful Text Representations

Relatedness → Finer-grained Semantic Relations

I believe he would come here.

I doubt he would come here.

Meaningful Text Representations

Relatedness → Finer-grained Semantic Relations

Language
Meaningful Text Representations
Reading between the Lines

Knowledge
Acquiring Commonsense Knowledge

Reasoning
Causal & Nonmonotonic Reasoning

Reliable Evaluation
Reading between the Lines

Language

Meaningful Text Representations

Knowledge

Acquiring Commonsense Knowledge

Reasoning

Causal & Nonmonotonic Reasoning

Reliable Evaluation
Reading between the Lines

Required language understanding skills:
Reading between the Lines

Required language understanding skills:

(i) Implicit meaning
Reading between the Lines

I didn’t eat anything since the morning.

Me too. I could eat a horse!

Required language understanding skills:
(1) Implicit meaning
(2) Non-literal meaning
Reading between the Lines

I didn’t eat anything since the morning.

Me too. I could eat a horse!

Do you have some food here?

Required language understanding skills:

(1) Implicit meaning
(2) Non-literal meaning
(3) Pragmatics
Reading between the Lines

Required language understanding skills:
(i) Implicit meaning
(2) Non-literal meaning
(3) Pragmatics
(4) Common background

**Reading between the Lines**

**Required language understanding skills:**

1. Implicit meaning
2. Non-literal meaning
3. Pragmatics
4. Common background
Reliable Evaluation

Language
- Meaningful Text Representations
- Reading between the Lines

Knowledge
- Acquiring Commonsense Knowledge

Reasoning
- Causal & Nonmonotonic Reasoning
Reliable Evaluation
Reliable Evaluation

Discriminative tasks:
Reliable Evaluation

Discriminative tasks:

✅ Easy to evaluate
Reliable Evaluation

Discriminative tasks:

✅ Easy to evaluate

❌ Models are right for the wrong reasons
Reliable Evaluation

**Discriminative tasks:**

- Easy to evaluate
- Models are right for the wrong reasons

**Generative tasks:**
Reliable Evaluation

Discriminative tasks:
- Easy to evaluate
- Models are right for the wrong reasons

Generative tasks:
- More nuanced & flexible than pre-defined labels
Reliable Evaluation

Discriminative tasks:

- Easy to evaluate
- Models are right for the wrong reasons

Generative tasks:

- More nuanced & flexible than pre-defined labels
- More similar to human reasoning process (no “answer choices”)
Reliable Evaluation

**Discriminative tasks:**
- Easy to evaluate
- Models are right for the wrong reasons

**Generative tasks:**
- More nuanced & flexible than pre-defined labels
- More similar to human reasoning process (no “answer choices”)
- Infinite answer space (no “guessing” of correct answer)
Reliable Evaluation

**Discriminative tasks:**
- ✔ Easy to evaluate
- ✗ Models are right for the wrong reasons

**Generative tasks:**
- ✔ More nuanced & flexible than pre-defined labels
- ✔ More similar to human reasoning process (no “answer choices”)
- ✔ Infinite answer space (no “guessing” of correct answer)
- ✗ No reliable automatic evaluation metric
Sara wanted to make dinner for some guests.

But she didn’t know how to cook.

She had to order pizza for her friends instead.
Sara wanted to make dinner for some guests. But she didn’t know how to cook. She had to order pizza for her friends instead.

Desiderata:
Sara wanted to make dinner for some guests. But she didn’t know how to cook. She had to order pizza for her friends instead.

Desiderata:

1 Reward correct answers that are different from the reference. Right before the guests arrived she tasted the food and it tasted bad.
Desiderata:

1. Reward **correct** answers that are **different** from the reference.
2. Penalize **incorrect** answers that are **similar** to the reference.

Sara wanted to make dinner for some guests. But she didn’t know how to cook. She had to order pizza for her friends instead.

Right before the guests arrived she tasted the food and it tasted bad. She didn’t know how to cook meat.
Desiderata:

1. Reward correct answers that are different from the reference.
2. Penalize incorrect answers that are similar to the reference.
Desiderata:

1. **Reward correct** answers that are **different** from the reference.
2. **Penalize incorrect** answers that are **similar** to the reference.

**Lexical Overlap Metrics:** BLEU, ROUGE, METEOR, CIDEr

- **x1** lexical variation  
- **x2**

Weak correlation with human judgement (**Novikova et al., 2017**).
# Reliable Evaluation

## Generative Evaluation

### Desiderata:

1. **Reward** correct answers that are **different** from the reference.
2. **Penalize** incorrect answers that are **similar** to the reference.

<table>
<thead>
<tr>
<th><strong>Lexical Overlap Metrics:</strong> BLEU, ROUGE, METEOR, CIDEr</th>
</tr>
</thead>
<tbody>
<tr>
<td>✔️ 1 lexical variation</td>
</tr>
</tbody>
</table>

Weak correlation with human judgement ([Novikova et al., 2017](#)).

<table>
<thead>
<tr>
<th><strong>Semantic Similarity Based Metrics:</strong> SPICE, BERTScore, BLEURT</th>
</tr>
</thead>
<tbody>
<tr>
<td>✔️ 1</td>
</tr>
</tbody>
</table>
Desiderata:
1. Reward correct answers that are different from the reference.
2. Penalize incorrect answers that are similar to the reference.

Lexical Overlap Metrics: BLEU, ROUGE, METEOR, CIDEr

Weak correlation with human judgement (Novikova et al., 2017).

Semantic Similarity Based Metrics: SPICE, BERTScore, BLEURT

Relatedness / similarity is very fuzzy!
Reliable Evaluation

Generative Evaluation

Desiderata:
1. Reward correct answers that are different from the reference.
2. Penalize incorrect answers that are similar to the reference.

Lexical Overlap Metrics: BLEU, ROUGE, METEOR, CIDEr

- Weak correlation with human judgement (Novikova et al., 2017).

Semantic Similarity Based Metrics: SPICE, BERTScore, BLEURT

- Relatedness / similarity is very fuzzy!
Desiderata:

1. Reward correct answers that are different from the reference.
2. Penalize incorrect answers that are similar to the reference.

**Lexical Overlap Metrics:** BLEU, ROUGE, METEOR, CIDEr

- **✓** 1 lexical variation
- **×** 2

Weak correlation with human judgement (Novikova et al., 2017).

**Semantic Similarity Based Metrics:** SPICE, BERTScore, BLEURT

- **✓** 1
- **×** 2 Relatedness / similarity is very fuzzy!

Combine metrics
Extrinsic evaluation
Task-specific learned metric (Chen et al., 2020)
Desiderata:

1. Reward correct answers that are different from the reference.
2. Penalize incorrect answers that are similar to the reference.

**Lexical Overlap Metrics:** BLEU, ROUGE, METEOR, CIDEr

- ✗ 1 lexical variation
- ✔ 2

Weak correlation with human judgement (Novikova et al., 2017).

**Semantic Similarity Based Metrics:** SPICE, BERTScore, BLEURT

- ✔ 1
- ✗ 2 Relatedness / similarity is very fuzzy!

Takeaways
Future Directions

Language
Meaningful Text Representations
Reading between the Lines

Knowledge
Acquiring Commonsense Knowledge

Reasoning
Causal & Nonmonotonic Reasoning
Reliable Evaluation

Takeaways
Future Directions

Language
Meaningful Text Representations
Reading between the Lines

Knowledge
Acquiring Commonsense Knowledge

Reasoning

Thank you! Questions?

vereds@allenai.org
@VeredShwartz
References (1)


(17) Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. going on a vacation takes longer than going for a walk: A study of temporal commonsense understanding. EMNLP 2019.

References (2)


(34) Robyn Speer and Catherine Havasi. Representing general relational knowledge in ConceptNET 5. LREC 2012.
References (3)

(56) Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference. EMNLP 2018.


