Unsupervised Methods for Commonsense Reasoning

Vered Shwartz
Oct 8, 2020
Machine Commonsense Reasoning
Machine Commonsense Reasoning

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

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

Bumping into people annoys them
Machine Commonsense Reasoning

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

- Bumping into people annoys them
- Raining makes the road slippery
Machine Commonsense Reasoning

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

- Bumping into people annoys them
- Raining makes the road slippery
- Commenting on someone’s weight is impolite
Machine Commonsense Reasoning

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

- Bumping into people annoys them
- Raining makes the road slippery
- Commenting on someone's weight is impolite
- Do not cut the branch on which you sit!
Machine Commonsense Reasoning

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

- Bumping into people annoys them
- Raining makes the road slippery
- Commenting on someone’s weight is impolite
- Do not cut the branch on which you sit!

Commonsense is essential for humans to live and interact with each other in a reasonable and safe way.
Machine Commonsense Reasoning

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

- Bumping into people annoys them
- Raining makes the road slippery
- Commenting on someone’s weight is impolite
- Do not cut the branch on which you sit!

Commonsense is essential for humans to live and interact with each other in a reasonable and safe way.

And for AI to understand human needs and actions better!
Why unsupervised?
Why unsupervised?

- We (the NLP community) have already trained expensive models - now is the time to leverage them as much as possible!
Why unsupervised?

- We (the NLP community) have already trained expensive models - now is the time to leverage them as much as possible!

- Supervised models often learn to solve a *dataset*, rather than the *task*.
Increasing number of parameters in NLP models

- **Mar 2018**: 93.6M
- **May 2018**: 100M
- **Aug 2018**: 340M
- **Nov 2018**: 665M
- **Feb 2019**: 1.5B
- **May 2019**: 8.5B
- **Aug 2019**: 11B
- **Nov 2019**: 355M
- **Feb 2020**: 340M
- **May 2020**: 1.5B
- **Aug 2020**: 175B

Models:
- ELMo
- GPT
- BERT
- GPT-2
- XLM
- XLNet
- RoBERTa
- T5
- Bart
- GPT-3
- RoBERTa
- XLNet
- Megatron-LM
- T-NLG
- GShard
Increasing cost of training NLP models
Spurious correlations & Annotation artifacts

“Hypothesis-only Baseline” for NLI

Two dogs run outside

A white dog and a brown dog are running on grass

Two cats running on grass

(Gururangan, Swayamdipta, et al., 2018; Poliak et al., 2018)
Spurious correlations & Annotation artifacts

“Hypothesis-only Baseline” for NLI

Two dogs run outside

A white dog and a brown dog are running on grass

Two cats running on grass

(Gururangan, Swayamdipta, et al., 2018; Poliak et al., 2018)
Spurious correlations & Annotation artifacts

“Hypothesis-only Baseline” for NLI

Two dogs run outside

A white dog and a brown dog are running on grass

Two cats running on grass

Outside is a general word → Entailment

Long sentence → Neutral

(Gururangan, Swayamdipta, et al., 2018; Poliak et al., 2018)
Spurious correlations & Annotation artifacts

“Hypothesis-only Baseline” for NLI

Two dogs run outside

A white dog and a brown dog are running on grass

Two cats running on grass

Outside is a general word

Entailment

Long sentence

Neutral

Cat is the “opposite” of dog which is common in the premise

Contradiction

(Gururangan, Swayamdipta, et al., 2018; Poliak et al., 2018)
In this talk

Unsupervised Commonsense Question Answering with Self-Talk

Backpropagation-based Decoding for Unsupervised Counterfactual and Abductive Reasoning
Unsupervised Commonsense Question Answering with Self-Talk

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula and Yejin Choi
EMNLP 2020
Unsupervised Commonsense Question Answering with Self-Talk

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula and Yejin Choi
EMNLP 2020
Commonsense Question Answering
COPA: Choice of Plausible Alternatives

Context: The man broke his toe.
Question: What was the cause?
Choices:
1) He got a hole in his sock.
2) He dropped a hammer on his foot.
<table>
<thead>
<tr>
<th>COPA: Choice of Plausible Alternatives</th>
<th>CommonsenseQA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Context:</strong> The man broke his toe.</td>
<td><strong>Question:</strong> Where on a river can you hold a cup upright to catch water on a sunny day?</td>
</tr>
<tr>
<td><strong>Question:</strong> What was the cause?</td>
<td><strong>Choices:</strong></td>
</tr>
<tr>
<td><strong>Choices:</strong></td>
<td>1) waterfall, 2) bridge, 3) valley, 4) pebble, 5) mountain</td>
</tr>
<tr>
<td>1) He got a hole in his sock.</td>
<td></td>
</tr>
<tr>
<td>2) <strong>He dropped a hammer on his foot.</strong></td>
<td></td>
</tr>
</tbody>
</table>
## Commonsense Question Answering

<table>
<thead>
<tr>
<th>COPA: Choice of Plausible Alternatives</th>
<th>CommonsenseQA</th>
<th>WinoGrande</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Context:</strong> The man broke his toe.</td>
<td><strong>Question:</strong> Where on a river can you hold a cup upright to catch water on a sunny day?</td>
<td><strong>Context:</strong> Katrina had the financial means to afford a new car while Monica did not, since ____ had a high paying job.</td>
</tr>
<tr>
<td><strong>Question:</strong> What was the cause?</td>
<td><strong>Choices:</strong> 1) waterfall, 2) bridge, 3) valley, 4) pebble, 5) mountain</td>
<td><strong>Choices:</strong> 1) Katrina 2) Monica</td>
</tr>
<tr>
<td><strong>Choices:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1) He got a hole in his sock.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2) He dropped a hammer on his foot.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Commonsense Question Answering

**COPA: Choice of Plausible Alternatives**

**Context:** The man broke his toe.

**Question:** What was the cause?

**Choices:**
1) He got a hole in his sock.
2) **He dropped a hammer on his foot.**

**CommonsenseQA**

**Question:** Where on a river can you hold a cup upright to catch water on a sunny day?

**Choices:**
1) **waterfall**, 2) bridge, 3) valley, 4) pebble, 5) mountain

**WinoGrande**

**Context:** Katrina had the financial means to afford a new car while Monica did not, since ____ had a high paying job.

**Choices:** 1) Katrina 2) Monica

**MC-TACO:** Multiple Choice Temporal Commonsense

**Context:** [...] dream of becoming a judge.

**Question:** How many years did it take for Mark to become a judge?

**Choices:** 1) 63 years 2) 7 weeks 3) **7 years** 4) 7 seconds 5) 7 hours
COPA: Choice of Plausible Alternatives

**Context:** The man broke his toe.
**Question:** What was the cause?
**Choices:**
1) He got a hole in his sock.
2) He dropped a hammer on his foot.

CommonsenseQA

**Question:** Where on a river can you hold a cup upright to catch water on a sunny day?
**Choices:**
1) waterfall, 2) bridge, 3) valley, 4) pebble, 5) mountain

WinoGrande

**Context:** Katrina had the financial means to afford a new car while Monica did not, since ____ had a high paying job.
**Choices:** 1) Katrina 2) Monica

MC-TACO: Multiple Choice Temporal Commonsense

**Context:** [...] dream of becoming a judge.
**Question:** How many years did it take for Mark to become a judge?
**Choices:**
1) 63 years 2) 7 weeks 3) 7 years 4) 7 seconds 5) 7 hours

SocialIQa: Social Interaction QA

**Context:** In the school play, Robin played a hero in the struggle [...] angry villain.
**Question:** How would others feel as a result?
**Choices:**
1) sorry for the villain.
2) hopeful that Robin will succeed.
3) like Robin should lose the fight.
### COPA: Choice of Plausible Alternatives

**Context:** The man broke his toe.
**Question:** What was the cause?
**Choices:**
1) He got a hole in his sock.
2) He dropped a hammer on his foot.

### CommonsenseQA

**Question:** Where on a river can you hold a cup upright to catch water on a sunny day?
**Choices:**
1) waterfall, 2) bridge, 3) valley, 4) pebble, 5) mountain

### WinoGrande

**Context:** Katrina had the financial means to afford a new car while Monica did not, since ____ had a high paying job.
**Choices:** 1) Katrina 2) Monica

### MC-TACO: Multiple Choice Temporal Commonsense

**Context:** [...] dream of becoming a judge.
**Question:** How many years did it take for Mark to become a judge?
**Choices:** 1) 63 years 2) 7 weeks 3) 7 years 4) 7 seconds 5) 7 hours

### SocialIQa: Social Interaction QA

**Context:** In the school play, Robin played a hero in the struggle [...] angry villain.
**Question:** How would others feel as a result?
**Choices:**
1) sorry for the villain.
2) hopeful that Robin will succeed.
3) like Robin should lose the fight.

### PIQA: Physical Interaction QA

**Question:** To separate egg whites from the yolk using a water bottle, you should
**Choices:**
1) [...] Release, which creates suction and lifts the yolk.
2) [...] Keep pushing, which creates suction and lifts the yolk.
Unsupervised
Unsupervised

$P_{LM}(\text{The answer is } \text{answer}_1)$
$P_{LM}(\text{The answer is } \text{answer}_2)$
...
$P_{LM}(\text{The answer is } \text{answer}_k)$
Unsupervised

$P_{LM}(\text{The answer is } \text{answer}_1)$

$P_{LM}(\text{The answer is } \text{answer}_2)$

$\ldots$

$P_{LM}(\text{The answer is } \text{answer}_k)$

Predict most “plausible” answer choice
Because Brett found an internship while in college but Ian was unable to, **Brett** found a job less quickly after graduation.

Because Brett found an internship while in college but Ian was unable to, **Ian** found a job less quickly after graduation.
Self-Talk
Self-Talk

instance $\text{LM}_1$
Self-Talk

instance

clarification question

LM$_1$
Self-Talk

instance

clarification

question

LM$_1$
Self-Talk

instance

LM₁

clarification

question

clarification

LM₂
Self-Talk
Can we use LMs to generate missing or implicit knowledge?*
Can we use LMs to generate missing or implicit knowledge?*

* Acknowledging limited coverage (Gordon and Van Durme, 2015) and limited precision (e.g. Kassner and Suchutze, 2019).
1. Generating Clarifications
Example #1: WinoGrande
Example #1: WinoGrande

Question Generation:
Because Brett found an internship while in college but Ian was unable to, ___ found a job less quickly after graduation.
Example #1: WinoGrande

Question Generation:
Because Brett found an internship while in college but Ian was unable to, ____ found a job less quickly after graduation.

What is the purpose of ___
Because Brett found an internship while in college but Ian was unable to, ____ found a job less quickly after graduation.

Question Generation:

What is the purpose of

The purpose of ____ is

Question & Answer Prefixes
Example #1: WinoGrande

Question Generation:
Because Brett found an internship while in college but Ian was unable to, ___ found a job less quickly after graduation.
What is the purpose of ___

Answer Generation:
Because Brett found an internship while in college but Ian was unable to, ___ found a job less quickly after graduation.
What is the purpose of the internship?
Example #1: WinoGrande

Question Generation:
Because Brett found an internship while in college but Ian was unable to, ___ found a job less quickly after graduation.
What is the purpose of ___?

Answer Generation:
Because Brett found an internship while in college but Ian was unable to, ___ found a job less quickly after graduation.
What is the purpose of ___?
The purpose of ___ is ___.

Question & Answer Prefixes
What is the purpose of ___?
The purpose of ___ is ___.
Example #1: WinoGrande

**Question Generation:**
Because Brett found an internship while in college but Ian was unable to, ___ found a job less quickly after graduation.

**Answer Generation:**
The purpose of the internship is to help people find jobs.
Example #1: WinoGrande

Question Generation:
Because Brett found an internship while in college but Ian was unable to, ___ found a job less quickly after graduation.
What is the purpose of ___?

Answer Generation:
The purpose of ___ is to help people find jobs.

Because Brett found an internship while in college but Ian was unable to, ___ found a job less quickly after graduation.
What is the purpose of ___?
The purpose of ___ is to help people find jobs.
Example #2: CommonsenseQA
Example #2: CommonsenseQA

Question Generation:

What do professors primarily do?
Example #2: CommonsenseQA

Question Generation:

What do professors primarily do?
What is the main function of

What is the main function of
The main function of ____ is
Example #2: CommonsenseQA

Question Generation:

What do professors primarily do?
What is the main function of

What is the main function of
The main function of ____ is

a professor’s teaching career?
Example #2: CommonsenseQA

**Question Generation:**

What do professors primarily do?
What is the main function of

**Answer Generation:**

What do professors primarily do?
What is the main function of

The main function of ____ is

(a professor’s teaching career)
Example #2: CommonsenseQA

Question Generation:
What do professors primarily do?
What is the main function of

Answer Generation:
What do professors primarily do?
What is the main function of
The main function of is
Example #2: CommonsenseQA

**Question Generation:**

What do professors primarily do?
What is the main function of

**Answer Generation:**

What do professors primarily do?
What is the main function of

The main function of a professor’s teaching career is to teach students how they can improve their knowledge.
Example #2: CommonsenseQA

Question Generation:
What do professors primarily do?
What is the main function of

Answer Generation:
What do professors primarily do?
What is the main function of

The main function of a professor’s teaching career is to teach students how they can improve their knowledge.
## Question & Answer Prefixes

### PIQA

<table>
<thead>
<tr>
<th>Question Prefix</th>
<th>Answer Prefix</th>
</tr>
</thead>
<tbody>
<tr>
<td>How to</td>
<td>The way to do ___ is</td>
</tr>
<tr>
<td>How do you</td>
<td>The way you do ___ is</td>
</tr>
<tr>
<td>How can one</td>
<td>One can ___ by</td>
</tr>
<tr>
<td>What can be used for</td>
<td>___ can be used for</td>
</tr>
<tr>
<td>What can one do in order to</td>
<td>In order to __, one can</td>
</tr>
<tr>
<td>What should you use for</td>
<td>For __, you should use</td>
</tr>
<tr>
<td>What is the definition of</td>
<td>The definition of ___ is</td>
</tr>
<tr>
<td>What are the properties of a</td>
<td>The properties of a ___ are</td>
</tr>
<tr>
<td>What is</td>
<td>A ___ is</td>
</tr>
</tbody>
</table>
# Question & Answer Prefixes

## PIQA

<table>
<thead>
<tr>
<th>Question Prefix</th>
<th>Answer Prefix</th>
</tr>
</thead>
<tbody>
<tr>
<td>How to</td>
<td>The way to do __ is</td>
</tr>
<tr>
<td>How do you</td>
<td>The way you do __ is</td>
</tr>
<tr>
<td>How can one</td>
<td>One can __ by</td>
</tr>
<tr>
<td>What can be used for</td>
<td>__ can be used for</td>
</tr>
<tr>
<td>What can one do in order to</td>
<td>In order to __, one can</td>
</tr>
<tr>
<td>What should you use for</td>
<td>For __, you should use</td>
</tr>
<tr>
<td>What is the definition of</td>
<td>The definition of __ is</td>
</tr>
<tr>
<td>What are the properties of a</td>
<td>The properties of a __ are</td>
</tr>
<tr>
<td>What is</td>
<td>A __ is</td>
</tr>
</tbody>
</table>

## WinoGrande

<table>
<thead>
<tr>
<th>Question Prefix</th>
<th>Answer Prefix</th>
</tr>
</thead>
<tbody>
<tr>
<td>What is the definition of</td>
<td>The definition of __ is</td>
</tr>
<tr>
<td>What are the properties of a</td>
<td>The properties of a __ are</td>
</tr>
<tr>
<td>What is</td>
<td>A __ is</td>
</tr>
<tr>
<td>What is the purpose of</td>
<td>The purpose of __ is</td>
</tr>
<tr>
<td>What is the main function of</td>
<td>The main function of __ is</td>
</tr>
<tr>
<td>What does mean to</td>
<td>__ Means</td>
</tr>
</tbody>
</table>
2. Predicting the Correct Answer Choice
Because Brett found an internship while in college but Ian was unable to, **Brett** found a job less quickly after graduation. The purpose of the internship is to help people find jobs.

Because Brett found an internship while in college but Ian was unable to, **Ian** found a job less quickly after graduation. The purpose of the internship is to help people find jobs.

Because Brett found an internship while in college but Ian was unable to, **Brett** found a job less quickly after graduation. The definition of “job” is to be employed by someone.

Because Brett found an internship while in college but Ian was unable to, **Ian** found a job less quickly after graduation. The definition of “job” is to be employed by someone.
Baselines
Because Brett found an internship while in college but Ian was unable to, **Brett** found a job less quickly after graduation.

Because Brett found an internship while in college but Ian was unable to, **Ian** found a job less quickly after graduation.
Knowledge-informed Models

Examples: Social IQa

Taylor was doing her job so she put the money in the drawer. What will Taylor do next?
Taylor was doing her job so she put the money in the drawer. What will Taylor do next?
Knowledge-informed Models

Examples: Social IQa

Taylor was doing her job so she put the money in the drawer.

What will Taylor do next?
Taylor was doing her job so she put the money in the drawer.

What will Taylor do next?

Job is a type of work. You would work because you want money.
Taylor was doing her job so she put the money in the drawer. What will Taylor do next?

Job is a type of work. You would work because you want money.
Taylor was doing her job so she put the money in the drawer. What will Taylor do next?

Job is a type of work. You would work because you want money. Job to earn money.
Taylor was doing her job so she put the money in the drawer.

Job is a type of work. You would work because you want money.

Job to earn money.

What will Taylor do next?

Want
Taylor was doing her job so she put the money in the drawer.

What will Taylor do next?

Job is a type of work. You would work because you want money. Job to earn money. As a result, Taylor wants to keep the money in the drawer.
Results
Results

Accuracy on the validation set
Results

Accuracy on the validation set
Results

Accuracy on the validation set

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Majority</th>
<th>Baseline</th>
<th>Knowledge-informed</th>
<th>Self-talk</th>
<th>Best supervised</th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>COPA</td>
<td>55</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MC-TACO</td>
<td>40.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social IQa</td>
<td>33.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PIQA</td>
<td>50.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Winogrande</td>
<td>50.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CommonsenseQA</td>
<td>20.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Results

Accuracy on the validation set

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Majority</th>
<th>Baseline</th>
<th>Knowledge-informed</th>
<th>Self-talk</th>
<th>Best supervised</th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>COPA</td>
<td>55</td>
<td>53</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MC-TACO</td>
<td>40.3</td>
<td>53.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social IQa</td>
<td>33.6</td>
<td>41.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PIQA</td>
<td>50.5</td>
<td>62.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Winogrande</td>
<td>50.4</td>
<td>54.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CommonsenseQA</td>
<td>20.9</td>
<td>37.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

24
Results

Accuracy on the validation set

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Majority</th>
<th>Baseline</th>
<th>Knowledge-informed</th>
<th>Self-talk</th>
<th>Best supervised</th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>COPA</td>
<td>69</td>
<td>53</td>
<td>55</td>
<td>53</td>
<td>55</td>
<td>55</td>
</tr>
<tr>
<td>MC-TACO</td>
<td>58.8</td>
<td>53.1</td>
<td>53.1</td>
<td>47.5</td>
<td>55</td>
<td>55</td>
</tr>
<tr>
<td>Social IQa</td>
<td>62.6</td>
<td>47.5</td>
<td>47.5</td>
<td>47.5</td>
<td>53.1</td>
<td>53.1</td>
</tr>
<tr>
<td>PIQA</td>
<td>69.6</td>
<td>55.4</td>
<td>55.4</td>
<td>55.4</td>
<td>55.4</td>
<td>55.4</td>
</tr>
<tr>
<td>Winogrande</td>
<td>54.8</td>
<td>54.8</td>
<td>54.8</td>
<td>54.8</td>
<td>54.8</td>
<td>54.8</td>
</tr>
<tr>
<td>CommonsenseQA</td>
<td>39.7</td>
<td>37.2</td>
<td>37.2</td>
<td>37.2</td>
<td>37.2</td>
<td>37.2</td>
</tr>
</tbody>
</table>
Results

Accuracy on the validation set

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Majority</th>
<th>Baseline</th>
<th>Knowledge-informed</th>
<th>Self-talk</th>
<th>Best supervised</th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>COPA</td>
<td>55</td>
<td>53</td>
<td>58.8</td>
<td>59.9</td>
<td>54.7</td>
<td>69.6</td>
</tr>
<tr>
<td>MC-TACO</td>
<td>40.3</td>
<td>53.1</td>
<td>58.8</td>
<td>59.9</td>
<td>54.7</td>
<td>69.6</td>
</tr>
<tr>
<td>Social IQa</td>
<td>33.6</td>
<td>41.1</td>
<td>47.5</td>
<td>46.2</td>
<td>50.5</td>
<td>54.7</td>
</tr>
<tr>
<td>PIQA</td>
<td>62.6</td>
<td>69.6</td>
<td>70.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Winogrande</td>
<td>50.4</td>
<td>54.5</td>
<td>55.4</td>
<td>54.7</td>
<td>50.5</td>
<td>54.7</td>
</tr>
<tr>
<td>CommonsenseQA</td>
<td>20.9</td>
<td>37.2</td>
<td>39.7</td>
<td>32.4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Results

Accuracy on the validation set

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Majority</th>
<th>Baseline</th>
<th>Knowledge-informed</th>
<th>Self-talk</th>
<th>Best supervised</th>
</tr>
</thead>
<tbody>
<tr>
<td>COPA</td>
<td>55</td>
<td>53</td>
<td>58.8</td>
<td>59.9</td>
<td>94.8</td>
</tr>
<tr>
<td>MC-TACO</td>
<td>69</td>
<td>66</td>
<td>53.1</td>
<td>58.8</td>
<td>69.6</td>
</tr>
<tr>
<td>Social IQa</td>
<td>40.3</td>
<td>41.1</td>
<td>47.5</td>
<td>46.2</td>
<td>76.6</td>
</tr>
<tr>
<td>PIQA</td>
<td>33.6</td>
<td>35.6</td>
<td>50.6</td>
<td>67.2</td>
<td>79.2</td>
</tr>
<tr>
<td>Winogrande</td>
<td>50.4</td>
<td>54.8</td>
<td>55.1</td>
<td>54.7</td>
<td>79.3</td>
</tr>
<tr>
<td>CommonsenseQA</td>
<td>20.9</td>
<td>37.2</td>
<td>39.7</td>
<td>32.4</td>
<td>83.7</td>
</tr>
</tbody>
</table>
Results

Accuracy on the validation set

<table>
<thead>
<tr>
<th>Test Set</th>
<th>Majority</th>
<th>Baseline</th>
<th>Knowledge-informed</th>
<th>Self-talk</th>
<th>Best supervised</th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>COPA</td>
<td>55</td>
<td>53</td>
<td>66</td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MC-TACO</td>
<td>53.1</td>
<td>58.8</td>
<td>59.9</td>
<td>94.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social IQa</td>
<td>33.6</td>
<td>41.1</td>
<td>46.2</td>
<td>76.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PIQA</td>
<td>62.6</td>
<td>69.6</td>
<td>70.2</td>
<td>94.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Winogrande</td>
<td>50.4</td>
<td>54.8</td>
<td>55.4</td>
<td>54.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CommonsenseQA</td>
<td>20.9</td>
<td>37.2</td>
<td>39.7</td>
<td>32.4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Human Evaluation
Human Evaluation

- Preliminary experiments: high ratio of noisy clarifications
  - But the model only uses \( \leq 1 \) clarification per instance
What do people think of useful clarifications?

Passage: Kendall was being careless and accidentally set fire to Sydney. Sydney had to go to the hospital because of the burns.
Main question: How would Kendall feel afterwards?

Evaluate the following clarification question asked about the passage/main question and its answer.

Question: What is the relationship between 'careless' and 'reckless'?  
☐ The question is completely gibberish, I can't understand it at all.  
☐ The question is not perfectly grammatical, but I think I can understand it.  
☒ The question is grammatical.  
☐ The question is on topic with respect to the given passage.

Answer: Careless is similar to reckless  
☐ The answer is completely gibberish, I can't understand it at all.  
☐ The answer is not perfectly grammatical, but I think I can understand it.  
☒ The answer is grammatical.  
☐ The answer is on topic with respect to the given passage and/or main question.  
☐ The answer is factually correct or likely true.  
☐ The answer provides useful information for answering the main question.

* Clarifications that change baseline prediction from incorrect to correct
What do people think of useful clarifications?

Most clarifications are grammatical or at least understandable:

- Grammatical: 87.2%
- Understandable: 10.1%
- Gibberish: 2.7%
What do people think of useful clarifications?

Most clarifications are grammatical or at least understandable:

- Grammatical: 87.2%
- Understandable: 10.1%
- Gibberish: 2.7%
What do people think of useful clarifications?

Clarifications judged as often relevant and factually correct but less frequently as helpful

- Relevant: 65%
- Correct: 60%
- Helpful: 41%
What do people think of useful clarifications?

Example: MC-TACO
What are the errors in Harmful clarifications?

* Clarifications that change baseline prediction from correct to incorrect
What are the errors in harmful clarifications?

The children were not vaccinated, which was fine with Betty but annoyed Mary. ______ believed they made kids autistic.

What does it mean to be “autistic”?

Be “autistic” means to have problems in social interaction and communication skills.

* Clarifications that change baseline prediction from correct to incorrect
Takeaways
Takeaways

- Generating knowledge with LMs improves upon the baseline and performs similarly to knowledge-informed models.
Takeaways

- Generating knowledge with LMs improves upon the baseline and performs similarly to knowledge-informed models.
- Generated clarifications don’t align with what humans consider helpful → different “reasoning process”? 
Takeaways

- Generating knowledge with LMs improves upon the baseline and performs similarly to knowledge-informed models.
- Generated clarifications don’t align with what humans consider helpful → different “reasoning process”?

Future directions:
- Multiple hops
- Introspection
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
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
Nonmonotonic Reasoning

A core human reasoning ability in which some conclusions can be invalidated by adding more knowledge.
Nonmonotonic Reasoning

A core human reasoning ability in which some conclusions can be invalidated by adding more knowledge.

What if something happened differently?
Nonmonotonic Reasoning

A core human reasoning ability in which some conclusions can be invalidated by adding more knowledge.

What if something happened differently?

What might have caused current events?
Nonmonotonic Reasoning

A core human reasoning ability in which some conclusions can be invalidated by adding more knowledge.

What if something happened differently?

What might have caused current events?

Nothing will change my mind! Unless...
Nonmonotonic Reasoning

A core human reasoning ability in which some conclusions can be invalidated by adding more knowledge.

What if something happened differently?

What might have caused current events?

Nothing will change my mind! Unless...
Abductive Reasoning

Reason about the most plausible explanation for incomplete observations.

Dataset: ART (Bhagavatula et al., 2020)
Abductive Reasoning

Reason about the most plausible explanation for incomplete observations.

Sara wanted to make dinner for some guests.

Dataset: ART (Bhagavatula et al., 2020)
Abductive Reasoning

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.

Dataset: **ART** (Bhagavatula et al., 2020)
Abductive Reasoning

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.

Dataset: ART (Bhagavatula et al., 2020)
Abductive Reasoning

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.

Dataset: ART (Bhagavatula et al., 2020)
Challenge:
Causal language models are conditioned only on a past context.
DELOREAN:
DEcoding for nonmonotonic LOGical REAsoNing

- Using pre-trained language models
- Unsupervised decoding strategy, based on backpropagation:

\[ \text{context } X \xrightarrow{\text{forward}} \text{continuation } Y = y_1, \ldots, y_n \xleftarrow{\text{backward}} \text{constraints } Z \]

Y is a fluent continuation of the context X
Y satisfies the task-specific constraints Z
DELOREAN: DEcoding for nonmonotonic LOgical REAsoNing

- Using pre-trained language models
- Unsupervised decoding strategy, based on backpropagation:

\[
\begin{align*}
\text{context } X & \quad \xrightarrow{\text{forward}} \quad \text{continuation } Y = y_1, \ldots, y_n \\
Y & \text{ is a fluent continuation of the context } X \\
\text{constraints } Z & \quad \xrightarrow{\text{backward}} \quad Y \text{ satisfies the task-specific constraints } Z
\end{align*}
\]
Abductive Reasoning

$X$  First observation  Sara wanted to make dinner for some guests.

$Z$  Second observation  She had to order pizza for her friends instead.

$Ý$  Generated hypothesis  But she didn’t know how to cook.
Constraints

Abductive Reasoning

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

Abductive Reasoning

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

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

Abductive Reasoning

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

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

Sara wanted to make dinner for some guests.
Constraints

Abductive Reasoning

Maximize the likelihood of LM to generate the second observation $Z$ following the first observation and the generated hypothesis $XY$

$$\mathcal{L}(X, \tilde{Y}, Z) := - \sum_{n=1}^{N_Z} \log P_{LM}(\tilde{z}_n \mid X, \tilde{Y}, Z_{1:n-1})$$

Sara wanted to make dinner for some guests. But she didn’t know how to cook.
Constraints

Abductive Reasoning

Maximize the likelihood of LM to generate the second observation $Z$ following the first observation and the generated hypothesis $X\bar{Y}$

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

Sara wanted to make dinner for some guests. But she didn’t know how to cook. She had to order.)
Constraints

Abductive Reasoning

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

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

$P(\text{pizza}| \text{Sara wanted to make dinner for some guests. But she didn’t know how to cook. She had to order})$
Constraints

Abductive Reasoning

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

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

$P(\text{pizza} | \text{Sara wanted to make dinner for some guests. But she didn’t know how to cook. She had to order})$

Loss is computed on a prefix of $\tilde{Y}$, which might be an incomplete sentence!
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.
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. But she didn’t know how to cook. She had to order pizza for her friends instead.
We repeat this process $T=20$ times.

Then we rank the $T$ candidate generations $Y_s = \{Y^{(t)} | t = 1,...,T\}$ by overall coherence and choose the best one.

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.
Ranking
Abductive Reasoning

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

$$\text{ranking\_score}(Y^{(t)}) = \text{BERT}_{\text{NSP}}(XY^{(t)}, Z) + \text{BERT}_{\text{NSP}}(X, Y^{(t)}Z)$$
Ranking

Abductive Reasoning

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

\[
\text{ranking\_score}(Y^{(t)}) = \text{BERT}_{\text{NSP}}(XY^{(t)}, Z) + \text{BERT}_{\text{NSP}}(X, Y^{(t)}Z)
\]

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.)
Ranking
Abductive Reasoning

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

\[
\text{ranking\_score}(Y^{(t)}) = BERT_{NSP}(XY^{(t)}, Z) + BERT_{NSP}(X, Y^{(t)}Z)
\]

P(He 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

<table>
<thead>
<tr>
<th>Coherence</th>
<th>X-Y</th>
<th>Y-Z</th>
<th>X-Y-Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Zero-Shot X,Z
- Zero-Shot X + Ranking
- Zero-Shot X,Z + Ranking
- Supervised
- Supervised + COMeT
- DELOREAN
- Human
Human Evaluation Results

Abductive Reasoning

Coherence

<table>
<thead>
<tr>
<th></th>
<th>X-Y</th>
<th>Y-Z</th>
<th>X-Y-Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zero-Shot X,Z</td>
<td>2.33</td>
<td>1.05</td>
<td>1.08</td>
</tr>
<tr>
<td>Zero-Shot X + Ranking</td>
<td>4.78</td>
<td>2.08</td>
<td>1.95</td>
</tr>
<tr>
<td>Zero-Shot X,Z + Ranking</td>
<td>4.74</td>
<td>2.58</td>
<td>2.56</td>
</tr>
<tr>
<td>Supervised</td>
<td>2.36</td>
<td>1.08</td>
<td>1.03</td>
</tr>
<tr>
<td>Supervised + COMeT</td>
<td>1.95</td>
<td>2.08</td>
<td>1.03</td>
</tr>
<tr>
<td>DELOREAN</td>
<td>4.78</td>
<td>2.58</td>
<td>2.56</td>
</tr>
<tr>
<td>Human</td>
<td>2.36</td>
<td>1.05</td>
<td>1.08</td>
</tr>
</tbody>
</table>
Human Evaluation Results

Abductive Reasoning

<table>
<thead>
<tr>
<th>Method</th>
<th>X-Y</th>
<th>Y-Z</th>
<th>X-Y-Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zero-Shot X,Z</td>
<td>2.33</td>
<td>1.08</td>
<td>1.95</td>
</tr>
<tr>
<td>Zero-Shot X + Ranking</td>
<td>4.78</td>
<td>2.08</td>
<td>2.56</td>
</tr>
<tr>
<td>Zero-Shot X,Z + Ranking</td>
<td>4.74</td>
<td>5.75</td>
<td>5.14</td>
</tr>
<tr>
<td>Supervised</td>
<td>5.1</td>
<td>3.42</td>
<td>2.86</td>
</tr>
<tr>
<td>Supervised + COMeT</td>
<td>4.66</td>
<td>5.1</td>
<td>5.1</td>
</tr>
<tr>
<td>DELOREAN</td>
<td>2.33</td>
<td>1.05</td>
<td>1.03</td>
</tr>
<tr>
<td>Human</td>
<td>2.55</td>
<td>2.38</td>
<td>2.33</td>
</tr>
</tbody>
</table>
Human Evaluation Results

Abductive Reasoning

Coherence

<table>
<thead>
<tr>
<th>Method</th>
<th>X-Y</th>
<th>Y-Z</th>
<th>X-Y-Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zero-Shot X,Z</td>
<td>4.74</td>
<td>5.1</td>
<td>5.22</td>
</tr>
<tr>
<td>Zero-Shot X + Ranking</td>
<td>4.66</td>
<td>5.75</td>
<td>5.25</td>
</tr>
<tr>
<td>Zero-Shot X,Z + Ranking</td>
<td>3.42</td>
<td>5.1</td>
<td>2.86</td>
</tr>
<tr>
<td>Supervised</td>
<td>2.55</td>
<td>2.08</td>
<td>1.95</td>
</tr>
<tr>
<td>Supervised + COMeT</td>
<td>1.03</td>
<td>1.08</td>
<td>1.03</td>
</tr>
<tr>
<td>DELOREAN</td>
<td>2.33</td>
<td>2.38</td>
<td>2.36</td>
</tr>
<tr>
<td>Human</td>
<td>5.22</td>
<td>5.25</td>
<td>2.97</td>
</tr>
</tbody>
</table>
Human Evaluation Results

Abductive Reasoning

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Coherence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zero-Shot X,Z</td>
<td>4.78, 5.1, 5.22</td>
</tr>
<tr>
<td>Zero-Shot X + Ranking</td>
<td>5.1, 4.66, 5.22</td>
</tr>
<tr>
<td>Zero-Shot X,Z + Ranking</td>
<td>5.22, 5.75, 6.25</td>
</tr>
<tr>
<td>Supervised</td>
<td>4.75, 5.25, 5.75</td>
</tr>
<tr>
<td>Supervised + COMeT</td>
<td>5.1, 5.65, 6.25</td>
</tr>
<tr>
<td>DELOREAN</td>
<td>5.1, 5.65, 6.25</td>
</tr>
<tr>
<td>Human</td>
<td>5.1, 5.65, 6.25</td>
</tr>
</tbody>
</table>

Legend:
- Blue: Zero-Shot X,Z
- Green: Zero-Shot X + Ranking
- Yellow: Zero-Shot X,Z + Ranking
- Red: Supervised
- Orange: Supervised + COMeT
- Purple: DELOREAN
- Gray: Human
Human Evaluation Results

Abductive Reasoning

Outperforms zero-shot models substantially

Competitive with supervised methods!
Example Generations
Abductive Reasoning

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

<table>
<thead>
<tr>
<th>t</th>
<th>Generated Y</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>She was thinking about the best way.</td>
<td>2.95</td>
</tr>
<tr>
<td>2</td>
<td>However, her cooking skills were the only thing that could make it a success.</td>
<td>7.73</td>
</tr>
<tr>
<td>3</td>
<td>But she couldn’t, because she was too busy with her work.</td>
<td>8.81</td>
</tr>
<tr>
<td>4</td>
<td>But she didn’t have the money and she didn’t have her own kitchen.</td>
<td>8.08</td>
</tr>
<tr>
<td>5</td>
<td>But she didn’t know how to cook.</td>
<td><strong>9.05</strong></td>
</tr>
</tbody>
</table>
**Example Generations**

**Abductive Reasoning**

<table>
<thead>
<tr>
<th>t</th>
<th>Generated Y</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>She was thinking about the best way.</td>
<td>2.95</td>
</tr>
<tr>
<td>2</td>
<td>However, her cooking skills were the only thing that could make it a success.</td>
<td>7.73</td>
</tr>
<tr>
<td>3</td>
<td>But she couldn’t, because she was too busy with her work.</td>
<td>8.81</td>
</tr>
<tr>
<td>4</td>
<td>But she didn’t have the money and she didn’t have her own kitchen.</td>
<td>8.08</td>
</tr>
<tr>
<td>5</td>
<td>But she didn’t know how to cook.</td>
<td><strong>9.05</strong></td>
</tr>
</tbody>
</table>

Backward pass introduces: contrast!
Counterfactual Reasoning

Reason about the causal change in future events given a change in condition.

Dataset: *TIME TRAVEL* (Qin et al., 2019)
Counterfactual Reasoning

Reason about the causal change in future events given a change in condition.

S1: Lisa was throwing a Halloween party.
S2: All her friends were dressing up.

Dataset: TIME TRAVEL (Qin et al., 2019)
Counterfactual Reasoning

Reason about the causal change in future events given a change in condition.

S1: Lisa was throwing a Halloween party.
S2: All her friends were dressing up.
S3: Lisa thought about being a wizard.
S4: Then she decided on a scarier costume.
S5: Lisa dressed up like a vampire.

Dataset: *TIME TRAVEL* (Qin et al., 2019)
Counterfactual Reasoning

Reason about the causal change in future events given a change in condition.

$S_1$: Lisa was throwing a Halloween party.

$S_2$: All her friends were dressing up.

$S_3$: Lisa thought about being a wizard.

$S_4$: Then she decided on a scarier costume.

$S_5$: Lisa dressed up like a vampire.

$S_2'$: It was a Game of Thrones themed party, so everyone had to dress as one of the main characters!

Dataset: *TIME TRAVEL* (Qin et al., 2019)
Counterfactual Reasoning

Reason about the causal change in future events given a change in condition.

$S_1$: Lisa was throwing a Halloween party.
$S_2$: All her friends were dressing up.
$S_3$: Lisa thought about being a wizard.
$S_4$: Then she decided on a scarier costume.
$S_5$: Lisa dressed up like a vampire.

$S_2'$: It was a Game of Thrones themed party, so everyone had to dress as one of the main characters!
$S_3'$: Lisa thought about how she would dress up as a Lannister, but she didn’t want to look like a Lannister.
$S_4'$: She wanted to look like a Stark.
$S_5'$: Lisa dressed up like a Stark.

Dataset: $\textit{TIME TRAVEL}$ (Qin et al., 2019)
Counterfactual Reasoning

Reason about the causal change in future events given a change in condition.

$S1$: Lisa was throwing a Halloween party.

$S2$: All her friends were dressing up.

$S3$: Lisa thought about being a wizard.

$S4$: Then she decided on a scarier costume.

$S5$: Lisa dressed up like a vampire.

$S2'$: It was a Game of Thrones themed party, so everyone had to dress as one of the main characters!

$S3'$: Lisa thought about how she would dress up as a Lannister, but she didn’t want to look like a Lannister.

$S4'$: She wanted to look like a Stark.

$S5'$: Lisa dressed up like a Stark.

Dataset: *TIME TRAVEL* (Qin et al., 2019)
Counterfactual Reasoning

X  Revised story beginning
Lisa was throwing a Halloween party.
It was a Game of Thrones themed party, so everyone had to dress as one of the main characters!

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

Ŷ  Generated ending
Lisa thought about how she would dress up as a Lannister, but she didn’t want to look like a Lannister.
She wanted to look like a Stark.
Lisa dressed up like a Stark.
Constraints

Counterfactual Reasoning

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

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

Counterfactual Reasoning

\[ \text{ranking\_score}(Y^{(T)}) = \text{BERT}_{\text{NSP}}(X, Y^{(T)}) + \sum_{s=1}^{S-1} \text{BERT}_{\text{NSP}}(Y^{(T)}_s, Y^{(T)}_{s+1}) \]
Ranking

Counterfactual Reasoning

Select $Y^{(t)}$ that is most likely to follow the revised beginning $X$

$$\text{ranking\_score}(Y^{(T)}) = \text{BERT}_{NSP}(X, Y^{(T)}) + \sum_{s=1}^{S-1} \text{BERT}_{NSP}(Y^{(T)}_s, Y^{(T)}_{s+1})$$
Ranking

Counterfactual Reasoning

Select $Y^{(t)}$ that is most likely to follow the revised beginning $X$

$$\text{ranking\_score}(Y^{(T)}) = \text{BERT}_{\text{NSP}}(X, Y^{(T)}) + \sum_{s=1}^{S-1} \text{BERT}_{\text{NSP}}(Y^{(T)}_s, Y^{(T)}_{s+1})$$

And also has internal consistency in its sentences
Human Evaluation Results

Counterfactual Reasoning

$F_\beta$

Only coherence $\beta$

Coherence and min edit are equal
Human Evaluation Results

Counterfactual Reasoning

2
1.75
1.5
1.25
1

$F_\beta$

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

Only coherence

$\beta$

Coherence and min edit are equal

- Zero-shot
- Zero-shot + Ranking
- Fine-tuned on original stories
- Fine-tuned on original stories + CF sentence
- Reconstructing original ending + CF sentence
- DELOREAN
Human Evaluation Results

Counterfactual Reasoning

- Zero-shot
- Zero-shot + Ranking
- Fine-tuned on original stories
- Fine-tuned on original stories + CF sentence
- Reconstructing original ending + CF sentence
- DELOREAN

$F_\beta$ values:

- $F_\beta = 2$ for Only coherence
- $F_\beta = 1.75$ for $\beta = 0$
- $F_\beta = 1.5$ for $\beta = 0.1$
- $F_\beta = 1.25$ for $\beta = 0.2$
- $F_\beta = 1$ for $\beta = 0.3$
- $F_\beta$ values drop as $\beta$ increases from 0.3 to 1

Coherence and min edit are equal for $\beta = 0$. The $\beta$ values range from 0 to 1.
Human Evaluation Results

Counterfactual Reasoning

- Zero-shot
- Zero-shot + Ranking
- Fine-tuned on original stories
- Fine-tuned on original stories + CF sentence
- Reconstructing original ending + CF sentence
- DELOREAN

The diagram shows the performance of different methods in Counterfactual Reasoning with varying values of $\beta$. The $F_\beta$ metric is plotted against $\beta$ values ranging from 0 to 1. The methods are compared based on their $F_\beta$ scores, with higher scores indicating better performance. The methods include:

- Only coherence
- Coherence and min edit are equal
Human Evaluation Results

Counterfactual Reasoning

- Zero-shot
- Zero-shot + Ranking
- Fine-tuned on original stories
- Fine-tuned on original stories + CF sentence
- Reconstructing original ending + CF sentence
- DELOREAN

$F_\beta$ vs $\beta$

Only coherence and min edit are equal
Human Evaluation Results

Counterfactual Reasoning

Coherence and min edit are equal

Only coherence

$F_\beta$
Human Evaluation Results

Counterfactual Reasoning

Most models are coherent with X but deviate too much from the original ending.

- Zero-shot
- Zero-shot + Ranking
- Fine-tuned on original stories
- Fine-tuned on original stories + CF sentence
- Reconstructing original ending + CF sentence
- DELOREAN

The graph shows the $F_\beta$ values for different methods as $\beta$ varies from 0 to 1. The $F_\beta$ values range from 2 to 1, with a trend that most methods perform better as $\beta$ decreases, indicating a preference for coherence. The $\beta$ axis represents the trade-off between coherence and min edit, where higher $\beta$ values favor min edit and lower $\beta$ values favor coherence.
Human Evaluation Results

Counterfactual Reasoning

Most models are coherent with X but deviate too much from the original ending.

\[ F_\beta \]

- Zero-shot
- Zero-shot + Ranking
- Fine-tuned on original stories
- Fine-tuned on original stories + CF sentence
- Reconstructing original ending + CF sentence
- DELOREAN

\( F_\beta \) vs. \( \beta \):

- Only coherence
- Coherence and min edit are equal
Human Evaluation Results

Counterfactual Reasoning

Most models are coherent with X but deviate too much from the original ending.

Recon+CF is very good at sticking to the original plot!... but it copies the original ending 84% of the times ¯\(_-\_\)/¯

Only coherence

Coherence and min edit are equal
**Human Evaluation Results**

**Counterfactual Reasoning**

- Zero-shot
- Zero-shot + Ranking
- Fine-tuned on original stories
- Fine-tuned on original stories + CF sentence
- Reconstructing original ending + CF sentence
- DELOREAN

Most models are coherent with X but deviate too much from the original ending.

Recon+CF is very good at sticking to the original plot... but it copies the original ending 84% of the times ¬¬

---

**Graph Details:**

- **Fβ** values range from 1.25 to 2.00.
- **β** values range from 0.0 to 1.0.
- Coherence and min edit are equal.
Human Evaluation Results

Counterfactual Reasoning

Most models are coherent with X but deviate too much from the original ending.

Recon+CF is very good at sticking to the original plot... but it copies the original ending 84% of the times ¯\_(ツ)_/¯

DELOREAN maintains a good balance!
Example Generations
Counterfactual Reasoning

<table>
<thead>
<tr>
<th>t</th>
<th>Generated Y: s3', s4', s5'</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>I was looking for a dog and I found this one. I was so happy to find a dog that I could take care of. I love him and he is so sweet.</td>
<td>4.04</td>
</tr>
<tr>
<td>2</td>
<td>I was looking for a dog and I found this one. I went to the vet and he said it was a good dog. I was so happy.</td>
<td>4.02</td>
</tr>
<tr>
<td>3</td>
<td>I was looking around on facebook and saw a dog. I went to the pet store and bought the dog. I was so excited to get my dog.</td>
<td>5.08</td>
</tr>
<tr>
<td>4</td>
<td>I was looking around on the internet for a pet and I found this website. I went to pick her up. I was so excited to see her.</td>
<td>2.75</td>
</tr>
<tr>
<td>5</td>
<td>I was looking around on the internet for a pet. I found a dog named ”Bubba”. I was so excited to have a dog.</td>
<td>5.07</td>
</tr>
</tbody>
</table>
Takeaways
Takeaways

- Unsupervised LM-based approach to generate text conditioned on past and future constraints.
Takeaways

- Unsupervised LM-based approach to generate text conditioned on past and future constraints.
- Effective for generative nonmonotonic reasoning tasks.
Takeaways

- Unsupervised LM-based approach to generate text conditioned on past and future constraints.
- Effective for generative nonmonotonic reasoning tasks.
- General approach that can be easily adapted for other generative reasoning tasks.
Takeaways

☐ Unsupervised LM-based approach to generate text conditioned on past and future constraints.

☐ Effective for generative nonmonotonic reasoning tasks.

☐ General approach that can be easily adapted for other generative reasoning tasks.

Thank you! Questions?

@VeredShwartz  vereds@allenai.org
Collaborators
References


[8] 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 2018.


