A Simple Method for Commonsense Reasoning

Trieu and Quoc
Beyond the Turing Test

Gary Marcus, Francesca Rossi, Manuela Veloso

On our best behaviour

Hector J. Levesque
Dept. of Computer Science
University of Toronto
Toronto, Ontario
Canada M5S 3A6
hector@cs.toronto.edu
• The **trophy** cannot fit in the **suitcase** because *it* is too big.
• The **trophy** cannot fit in the **suitcase** because *it* is too big.
The *trophy* cannot fit in the *suitcase* because *it* is too big.
Winograd Schema Challenge

- The **trophy** cannot fit in the **suitcase** because *it* is too big.
Winograd Schema Challenge

- The **trophy** cannot fit in the **suitcase** because *it* is too big.
Winograd Schema Challenge

The racecar zoomed by the school bus because [it] was going so fast.
Winograd Schema Challenge

The **racecar** zoomed by the **school bus** because [**it**] was going so fast.

**Comment:** From (Levesque 2009); deliberately created as a non-example. "Fast" is associated with racecars.
Winograd Schema Challenge

The racecar zoomed by the school bus because [it] was going so fast.

Comment: From (Levesque 2009); deliberately created as a non-example. "Fast" is associated with racecars.

Frank was pleased when Bill said [he] was the winner of the competition.

Comment: From (Levesque 2009); deliberately created as a non-example. The version with "pleased" is genuinely ambiguous (i.e. to the human reader). Frank might well be pleased on learning that Bill was the winner.
Winograd Schema Challenge

- The **trophy** cannot fit in the **suitcase** because **it** is too big.

On our best behaviour

Hector J. Levesque
Dept. of Computer Science
University of Toronto
Winograd Schema Challenge

- The *trophy* cannot fit in the *suitcase* because *it* is too big.

Expensive to construct commonsensereasoning.org < 300 samples.
Winograd Schema Challenge

- The **trophy** cannot fit in the **suitcase** because *it* is too big.

Random Guess: ~50%

Human: ~90%
Winograd Schema Challenge

- The **trophy** cannot fit in the **suitcase** because *it* is too big.

Random Guess: ~50%

SOTA: ~53%

Human: ~90%
• The **trophy** cannot fit in the **suitcase** because *it* is too big.

**Winograd Schema Challenge**

- Random Guess: ~50%
- SOTA: ~53%
  - Rule-based reasoning, Hand-crafted features.
  - Wordnet(1995)
  - ConceptNet(2004)-17m
  - Cyc(1984)
  - Google Search API
- Human: ~90%
The trophy cannot fit in the suitcase because it is too big.
The trophy cannot fit in the suitcase because it is too big.

Winograd Schema Challenge

Random Guess: ~50%

SOTA: ~53%
Word2Vec + Supervised DeepNN

Human: ~90%

Wordnet(1995)
ConceptNet(2004)-17m
Cyc(1984)
Google Search API
The trophy cannot fit in the suitcase because it is too big.

Winograd Schema Challenge

- Random Guess: ~50%
- SOTA: ~53%
- Human: ~90%

**Word2Vec + Supervised DeepNN**

- Wordnet (1995)
- ConceptNet (2004) - 17m
- Cyc (1984)
- Google Search API
The **trophy** cannot fit in the **suitcase** because *it* is too big.

**Winograd Schema Challenge**

- **Random Guess:** ~50%
- **SOTA:** ~53%
- **Ours LM:** ~64%
- **Human:** ~90%
Method

- The **trophy** cannot fit in the **suitcase** because *it* is too big.
Method

- The **trophy** cannot fit in the **suitcase** because *it* is too big.

The **troph**y cannot fit in the **suitcase** because **the trophy** is too big.
Method

- The **trophy** cannot fit in the **suitcase** because *it* is too big.

The **trophy** cannot fit in the **suitcase** because **the trophy** is too big.

The **trophy** cannot fit in the **suitcase** because **the suitcase** is too big.
Method

- The **trophy** cannot fit in the **suitcase** because *it* is too big.

More probable?

The **trophy** cannot fit in the **suitcase** because **the trophy** is too big.

The **trophy** cannot fit in the **suitcase** because **the suitcase** is too big.
Method

- The **trophy** cannot fit in the **suitcase** because *it* is too big.

2. Then use *big data*: search all the English text on the web to determine which is the more common pattern:

   - *x does not fit in y* + *x is so small* vs.
   - *x does not fit in y* + *y is so small*

The **trophy** cannot fit in the **suitcase** because **the suitcase** is too big.
Method

- The trophy cannot fit in the suitcase because it is too big.

Fred is the only man alive who still remembers my father as an infant. When Fred first saw my father, he was twelve years old. Who was twelve years old?
- Fred
- my father (Special=years; other=months)

The trophy cannot fit in the suitcase because the suitcase is too big.
Method

- The **trophy** cannot fit in the **suitcase** because *it* is too big.

More probable?

The **trophy** cannot fit in the **suitcase** because the **trophy** is too big.

The **trophy** cannot fit in the **suitcase** because the **suitcase** is too big.
Method

- The **trophy** cannot fit in the **suitcase** because *it* is too big.

\[P(\text{substitution}|\text{Human knowledge})\]

The **trophy** cannot fit in the **suitcase** because **the trophy** is too big.

The **trophy** cannot fit in the **suitcase** because **the suitcase** is too big.
Method

- The **trophy** cannot fit in the **suitcase** because *it* is too big.

\[
P(\text{substitution}|\text{Human knowledge})
\]

The **trophy** cannot fit in the **suitcase** because **the trophy** is too big.

The **trophy** cannot fit in the **suitcase** because **the suitcase** is too big.
Method

- The trophy cannot fit in the suitcase because it is too big.

Language Model

\[ P(substitution|\hat{\theta}) \]

The trophy cannot fit in the suitcase because the trophy is too big.

The trophy cannot fit in the suitcase because the suitcase is too big.
The trophy cannot fit in the suitcase because *it* is too big.

The trophy cannot fit in the suitcase because *the trophy* is too big.

The trophy cannot fit in the suitcase because *the suitcase* is too big.
Method

- The **trophy** cannot fit in the **suitcase** because *it* is too big.

The **trophy** cannot fit in the **suitcase** because **the trophy** is too big.

The **trophy** cannot fit in the **suitcase** because **the suitcase** is too big.
Method

- The *trophy* cannot fit in the *suitcase* because *it* is too big.

Language Model

The *trophy* cannot fit in the *suitcase* because *the trophy* is too big.

The *trophy* cannot fit in the *suitcase* because *the suitcase* is too big.
Method

- The **trophy** cannot fit in the **suitcase** because *it* is too big.

Language Model

The **trophy** cannot fit in the **suitcase** because *the trophy* is too big.

The **trophy** cannot fit in the **suitcase** because *the suitcase* is too big.
Method

- The **trophy** cannot fit in the **suitcase** because **it** is too big.

The **trophy** cannot fit in the **suitcase** because **the trophy** is too big.

The **trophy** cannot fit in the **suitcase** because **the suitcase** is too big.
Method

- The **trophy** cannot fit in the **suitcase** because **it** is too big.

Language Model

The **trophy** cannot fit in the **suitcase** because **the trophy** is too big.

The **trophy** cannot fit in the **suitcase** because **the suitcase** is too big.
Method

- The **tmmhoy** cannot fit in the **suitcase** because *it* is too big.

<table>
<thead>
<tr>
<th>Data name</th>
<th>Wrong prediction</th>
<th>Corrected</th>
<th>Correction percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDP-60</td>
<td>30</td>
<td>10</td>
<td>33.3%</td>
</tr>
<tr>
<td>PDP-122</td>
<td>55</td>
<td>33</td>
<td>60.0%</td>
</tr>
<tr>
<td>WSC-273</td>
<td>102</td>
<td>64</td>
<td>62.7%</td>
</tr>
</tbody>
</table>

The **tmmhoy** cannot fit in the **suitcase** because **the suitcase** is too big.
Method

- The trophy cannot fit in the suitcase because it is too big.

The trophy cannot fit in the suitcase because the trophy is too big.

The trophy cannot fit in the suitcase because the suitcase is too big.
Method

- The **trophy** cannot fit in the **suitcase** because *it* is too big.

---

**Language Model**

The **trophy** cannot fit in the **suitcase** because **the trophy** is too big.

The **trophy** cannot fit in the **suitcase** because **the suitcase** is too big.
The trophy cannot fit in the suitcase because it is too big.

Language Model

The trophy cannot fit in the suitcase because the trophy is too big.

The trophy cannot fit in the suitcase because the suitcase is too big.

Method

- The **trophy** cannot fit in the **suitcase** because **it** is too big.
The *trophy* cannot fit in the *suitcase* because *it* is too big.

The *trophy* cannot fit in the *suitcase* because *the trophy* is too big.

The *trophy* cannot fit in the *suitcase* because *the suitcase* is too big.
Method

- The **trophy** cannot fit in the **suitcase** because *it* is too big.

Language Model

The **trophy** cannot fit in the **suitcase** because **the trophy** is too big.

The **trophy** cannot fit in the **suitcase** because **the suitcase** is too big.
The trophy cannot fit in the suitcase because it is too small.

The trophy cannot fit in the suitcase because the trophy is too big.

The trophy cannot fit in the suitcase because the suitcase is too big.
Method

- The **trophy** cannot fit in the **suitcase** because *it* is too **small**.
- The **trophy** cannot fit in the **suitcase** because *it* is too **big**.

**Language Model**

The **trophy** cannot fit in the **suitcase** because the **trophy** is too **big**.

The **trophy** cannot fit in the **suitcase** because the **suitcase** is too **big**.

$\% = \frac{\text{area of red box}}{\text{area of green box}}$
The **trophy** cannot fit in the **suitcase** because *it* is too big.

The **trophy** cannot fit in the **suitcase** because *it* is too small.
Capturing special words

Full: Paul tried to call George on the phone, but (Paul/ George*) wasn’t [available].
Partial: Paul tried to call George on the phone, but (Paul/ George*) wasn’t [available].
Capturing special words

Full: Paul tried to call George on the phone, but (Paul/ George*) wasn't [available].
Partial: Paul tried to call George on the phone, but (Paul/ George*) wasn't [available].

Full: The drain is clogged with hair. The (drain*/ hair) has to be [cleaned].
Partial: The drain is clogged with hair. The (drain*/ hair) has to be [cleaned].

Full: Sam took French classes from Adam, because (Sam*/ Adam) was [eager] to speak it fluently.
Partial: Sam took French classes from Adam, because (Sam*/ Adam) was [eager] to speak it fluently.

Full: Jim [yelled at] Kevin because (Jim*/ Kevin) was so upset.
Partial: Jim [yelled at] Kevin because (Jim*/ Kevin) was so upset.

- Vocabulary of 800K words, including common names
Capturing special words

<table>
<thead>
<tr>
<th>Full:</th>
<th>Partial:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full:</td>
<td>Paul tried to call George on the phone, but <em>(Paul/ George</em>) wasn't [available].</td>
</tr>
<tr>
<td>Full:</td>
<td>The drain is clogged with hair. The <em>(drain</em>/ hair) has to be [cleaned].</td>
</tr>
<tr>
<td>Full:</td>
<td>Sam took French classes from Adam, because <em>(Sam</em>/ Adam) was [eager] to speak it fluently.</td>
</tr>
<tr>
<td>Full:</td>
<td>Jim [yelled at] Kevin because <em>(Jim</em>/ Kevin) was so upset.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Special word retrieved</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forward scoring</td>
</tr>
<tr>
<td>Backward scoring</td>
</tr>
</tbody>
</table>
## Results

### Table 4: Accuracy on Winograd Schema Challenge

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random guess</td>
<td>50.0%</td>
</tr>
<tr>
<td>USSM + Knowledge Base</td>
<td>52.0 %</td>
</tr>
<tr>
<td>USSM + Supervised DeepNet + Knowledge Base</td>
<td>52.8 %</td>
</tr>
<tr>
<td>Char-LM</td>
<td>51.3%</td>
</tr>
<tr>
<td>Word-LM</td>
<td>56.4%</td>
</tr>
<tr>
<td><strong>Ensemble of 10 Unsupervised LMs</strong></td>
<td><strong>61.5 %</strong></td>
</tr>
</tbody>
</table>
# Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random guess</td>
<td>50.0%</td>
</tr>
<tr>
<td>USSM + Knowledge Base</td>
<td>52.0%</td>
</tr>
<tr>
<td>USSM + Supervised DeepNet + Knowledge Base</td>
<td>52.8%</td>
</tr>
<tr>
<td>Char-LM</td>
<td>51.3%</td>
</tr>
<tr>
<td>Word-LM</td>
<td>56.4%</td>
</tr>
<tr>
<td><strong>Ensemble of 10 Unsupervised LMs</strong></td>
<td><strong>61.5%</strong></td>
</tr>
</tbody>
</table>
Results

STORIES = CommonCrawl documents w/ largest overlapping n-grams

One day when John and I had been out on some business of our master's, and were returning gently on a long, straight road, at some distance we saw a boy trying to leap a pony over a gate; the pony would not take the leap, and the boy cut him with the whip, but he only turned off on one side. He whipped him again, but the pony turned off on the other side. Then the boy got off and gave him a hard thrashing, and knocked him about the head...
Results

STORIES = CommonCrawl documents w/ largest overlapping n-grams

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>USSM + Supervised DeepNet + Knowledge Base</td>
<td>52.8 %</td>
</tr>
<tr>
<td>Char-LM-<em>partial</em></td>
<td>57.9 %</td>
</tr>
<tr>
<td>Word-LM-<em>partial</em></td>
<td>62.6 %</td>
</tr>
<tr>
<td>Ensemble of 14 LMs</td>
<td>63.7 %</td>
</tr>
</tbody>
</table>

Gutenberg Books
LM-1-Billion
CommonCrawl
SQuAD
+ STORIES
Diversity of Training Data

Ensemble on WSC

- LM1b
- CommCrawl
- SQuAD
- Gutenberg
- STORIES
- All

Scores:
- LM1b: 0.55
- CommCrawl: 0.57
- SQuAD: 0.59
- Gutenberg: 0.61
- STORIES: 0.63
- All: 0.62
Preprint and Code for Reproducing

A Simple Method for Commonsense Reasoning

Trieu H. Trinh*  
Google Brain  
thtrieu@google.com

Quoc V. Le  
Google Brain  
qvl@google.com

Abstract

Commonsense reasoning is a long-standing challenge for deep learning. For example, it is difficult to use neural networks to tackle the Winograd Schema dataset [1]. In this paper, we present a simple method for commonsense reasoning with neural networks, using unsupervised learning. Key to our method is the use of language models, trained on a massive amount of unlabeled data, to score multiple choice questions posed by commonsense reasoning tests. On both Pronoun Disambiguation and Winograd Schema challenges, our models outperform previous state-of-the-art methods by a large margin, without using expensive annotated knowledge bases or hand-engineered features. We train an array of large RNN language models that operate at word or character level on LM-1-Billion, CommonCrawl, SQuAD, Gutenberg Books, and a customized corpus for this task and show that diversity of training data plays an important role in test performance. Further analysis also shows that our system successfully discovers important features of the context that decide the correct answer, indicating a good grasp of commonsense knowledge.
Takeaway

● Deep NN can capture Commonsense.

● Commonsense representation does not have to be Graphs or Tuples, but might as well be vectors.
Thank you!

Q & A