Common Sense Beyond English: Evaluating and Improving Multilingual Language Models for Commonsense Reasoning

Bill Yuchen Lin  Seyeon Lee  Xiaoyang Qiao  Xiang Ren
{yuchen.lin, seyeonle, xiaoyanq, xiangren}@usc.edu
Department of Computer Science and Information Sciences Institute,
University of Southern California
Common Sense Beyond English and En-LMs

- Common Sense research has been limited to English and En-LMs.
  - **Probing** (e.g., LAMA)
  - **Reasoning** (e.g., CSQA, CODAH)

- What about other languages and multilingual LMs (ML-LMs)?
  - Common Sense Beyond En and En-LMs.
Background: LAMA Probing and Its Limitations

• LAMA probing (EMNLP’19) as a zero-shot token-ranking task.

• Limitations
  • Only token-level predictions.
    • Multi-token words can be very common in other languages, e.g., Chinese.
    • Sentence-level common sense is not explicitly tested.
  • Evaluation metric.
    • Different langs have totally different vocabulary.
    • False negative is problematic because of corpus bias.
Mickey Probe: Multilingual Commonsense Knowledge Probing

The effect of reading the news is lying about the world.
... of interviewing the deceased is learning about the world.
... of tracking the dragon is learning about the world.
... of reading the news is learning about the world. \( M^\text{en}_4 \)
... of reading the news is saying about the world.

\[
M = \{ M^\text{en}, \ldots, M^\text{zh} \}
\]

Input:
- A set of (K=5) statements.
- There are \(|\mathcal{L}|\) different versions in testing languages.
  \[
  \mathcal{L} = \{ \text{en, fr, ru, zh, \ldots} \}
  \]

Output:
- A sorted list of sentences in each lang’s version.
- The top-1 should have highest commonsense plausibility.

Probing w/ PLLs (Salazar; ACL’20):

\[
\log P(s) = \text{PLL}(s) := \sum_{i=1}^{s} \log P(w_i | s_{\setminus i})
\]
Constructing the Corpus for Mickey Probe

- OMCS corpus: a collection of commonsense assertions in English.
  - Step 1: generate distractors via masked-attacking.
  - Step 2: machine translation with MarianMT for En-X.
  - Step 3: back-translation (En-X-EN) and filter with SentenceBERT.

561k sentences in total.

$T = \# \text{ probes.}$

$K = \# \text{ options (K-1 distractors)}$

$|L| = \# \text{ languages.}$

$T = 10.2k, \ K=5, \ |L|=11$

More implementation details are in our paper.
Mickey Probing Results

Q1: Do different ML-LMs have similar language preferences? **No**

Q2: Does length influence PLL-based ranking? **Not much.** See the “Shortest” row.

Corpus size and translation quality are not **major factors** in lang preference of ML-LMs.

**Further implications:** common sense can be a bridge for the alignment within ML-LMs.
X-CSR Datasets and Evaluation

### Zero-Shot Cross-Lingual Transfer

- Initialize pre-trained ML-MLMs
- Fine-tune on English Training data.
- Validate and test on all languages.

**Shared setup (and langs) w/ XNLI.**

{en, zh, de, es, fr, it, jap, nl, pl, pt, ru, ar, vi, hi, **sw, ur**}

---

**Table: X-CSQA vs X-CODAH**

<table>
<thead>
<tr>
<th>Stat. ↓ Dataset →</th>
<th>X-CSQA</th>
<th>X-CODAH</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task Format</td>
<td>QA 15 + en</td>
<td>SceneComp. 15 + en</td>
</tr>
<tr>
<td># Languages</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td># Options per Example</td>
<td></td>
<td></td>
</tr>
<tr>
<td># Training (en)</td>
<td>8,888</td>
<td>8,476</td>
</tr>
<tr>
<td># Dev per Lang.</td>
<td>1,000</td>
<td>300</td>
</tr>
<tr>
<td># Test per Lang.</td>
<td>1,074</td>
<td>1,000</td>
</tr>
<tr>
<td># Total Instances</td>
<td>80,550</td>
<td>60,000</td>
</tr>
</tbody>
</table>
MCP: Multilingual Contrastive Pre-training of ML-LMs.

• Background: most ML-LMs only have token-level objectives during pre-training as follows:
  • Masked-LM (masked-word-prediction inside a single language)
  • Translation-LM (a cross-lingual version in a pair of parallel sentences.)

• Motivation:
  • The key to NLU tasks is strong sentence representation ([CLS]).
  • However, it is not well pre-trained.
  • Can we design a multilingual, sentence-level pre-training task?
**MCP: Multilingual Contrastive Pre-training of ML-LMs.**

- **Goal:** reuse the Mickey corpus to create a multilingual task.
- **Start:** a MickeyProbe instance.
  - Step 1): we randomly pick a lang as the anchor lang.
  - Step 2): use the correct assertion in the anchor lang as the target sentence.
  - Step 3): randomly select distractors from other languages.
- **End:** a MCP task instance.
  - **MCP input:** a set of sampled sentences in different languages, where only one is a commonsense assertion and others are distractors.
  - **MCP output:** the label to the correct commonsense assertion (via [CLS] representation).
An example of MCP instance.

```json
{
    "id": "1472b1b7350f4fcb",
    "truth_id": 1, # the id of the correct assertion.
    "langs": ["bg", "en", "zh", "ru", "hi", "fr", "vi", "de"],
    "probes": [
        "You would visit other countries because you want to experience other cultures.",
        "你会去其他国家，因为你想体验其他文化。",
        "Вы бы посетили другие страны, потому что хотите испытать другие таланты.",
        "आप दूसरे स्तरों पर जाते क्योंकि आप अन्य संस्कृतियों पर विज्ञापन का भय रखते हैं।",
        "Vous visiteriez d'autres personnes parce que vous voulez faire l'expérience d'autres cultures.",
        "Bạn sẽ đi thăm những con đường khác bởi vì bạn sợ để phá vỡ khác chủng tộc.",
        "Sie würden andere Personen besuchen, weil Sie andere Kulturen erleben möchten."
    ]
}
```
## X-CSR Experimental Results

<table>
<thead>
<tr>
<th></th>
<th>en</th>
<th>de</th>
<th>it</th>
<th>es</th>
<th>fr</th>
<th>nl</th>
<th>ru</th>
<th>vi</th>
<th>zh</th>
<th>hi</th>
<th>pl</th>
<th>ar</th>
<th>ja</th>
<th>pt</th>
<th>sw</th>
<th>uk</th>
<th>avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC-size (GB)</td>
<td>300.8</td>
<td>66.6</td>
<td>30.2</td>
<td>53.3</td>
<td>56.8</td>
<td>29.3</td>
<td>278.0</td>
<td>137.3</td>
<td>46.9</td>
<td>20.2</td>
<td>44.6</td>
<td>28.0</td>
<td>69.3</td>
<td>49.1</td>
<td>1.6</td>
<td>5.7</td>
<td>76.10</td>
</tr>
</tbody>
</table>

### X-CODAH
*Task: Scene Completion; Random Guess: 25.0; RoBERTa_L for en: 81.6*

<table>
<thead>
<tr>
<th>Model</th>
<th>en</th>
<th>de</th>
<th>it</th>
<th>es</th>
<th>fr</th>
<th>nl</th>
<th>ru</th>
<th>vi</th>
<th>zh</th>
<th>hi</th>
<th>pl</th>
<th>ar</th>
<th>ja</th>
<th>pt</th>
<th>sw</th>
<th>uk</th>
</tr>
</thead>
<tbody>
<tr>
<td>mBERT</td>
<td>42.9</td>
<td>33.1</td>
<td>33.5</td>
<td>33.8</td>
<td>35.2</td>
<td>33.7</td>
<td>31.9</td>
<td>22.8</td>
<td>38.0</td>
<td>26.5</td>
<td>31.0</td>
<td>34.8</td>
<td>34.0</td>
<td>37.2</td>
<td>30.8</td>
<td>31.5</td>
</tr>
<tr>
<td>XLM-100</td>
<td>42.7</td>
<td>31.5</td>
<td>32.2</td>
<td>30.7</td>
<td>34.9</td>
<td>32.6</td>
<td>30.9</td>
<td>24.7</td>
<td>31.4</td>
<td>26.8</td>
<td>27.0</td>
<td>30.0</td>
<td>27.4</td>
<td>33.2</td>
<td>25.3</td>
<td>24.9</td>
</tr>
<tr>
<td>XLM-R-B</td>
<td>50.1</td>
<td>45.8</td>
<td>44.4</td>
<td>44.2</td>
<td>45.2</td>
<td>42.0</td>
<td>44.1</td>
<td>43.2</td>
<td>44.6</td>
<td>38.1</td>
<td>41.9</td>
<td>37.8</td>
<td>42.0</td>
<td>44.1</td>
<td>35.6</td>
<td>34.6</td>
</tr>
<tr>
<td>XLM-R-L</td>
<td>66.4</td>
<td>59.6</td>
<td>59.9</td>
<td>60.9</td>
<td>60.1</td>
<td>59.3</td>
<td>56.3</td>
<td>57.4</td>
<td>57.3</td>
<td>49.1</td>
<td>57.5</td>
<td>51.2</td>
<td>53.8</td>
<td>58.2</td>
<td>42.2</td>
<td>46.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>en</th>
<th>de</th>
<th>it</th>
<th>es</th>
<th>fr</th>
<th>nl</th>
<th>ru</th>
<th>vi</th>
<th>zh</th>
<th>hi</th>
<th>pl</th>
<th>ar</th>
<th>ja</th>
<th>pt</th>
<th>sw</th>
<th>uk</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCP(XLM-R_B)</td>
<td>52.2</td>
<td>47.6</td>
<td>46.2</td>
<td>44.4</td>
<td>48.1</td>
<td>44.8</td>
<td>42.9</td>
<td>43.2</td>
<td>45.7</td>
<td>37.8</td>
<td>41.8</td>
<td>41.8</td>
<td>42.9</td>
<td>44.7</td>
<td>37.2</td>
<td>36.4</td>
</tr>
<tr>
<td>MCP(XLM-R_L)</td>
<td>69.9</td>
<td>60.7</td>
<td>61.9</td>
<td>60.7</td>
<td>61.4</td>
<td>60.7</td>
<td>58.6</td>
<td>62.3</td>
<td>61.9</td>
<td>53.7</td>
<td>59.0</td>
<td>54.1</td>
<td>54.7</td>
<td>60.8</td>
<td>44.6</td>
<td>48.0</td>
</tr>
</tbody>
</table>

### X-CSQA
*Task: Question Answering; Random Guess: 20.0; RoBERTa_L for en: 70.4*

<table>
<thead>
<tr>
<th>Model</th>
<th>en</th>
<th>de</th>
<th>it</th>
<th>es</th>
<th>fr</th>
<th>nl</th>
<th>ru</th>
<th>vi</th>
<th>zh</th>
<th>hi</th>
<th>pl</th>
<th>ar</th>
<th>ja</th>
<th>pt</th>
<th>sw</th>
<th>uk</th>
</tr>
</thead>
<tbody>
<tr>
<td>mBERT</td>
<td>38.8</td>
<td>29.6</td>
<td>36.4</td>
<td>35.3</td>
<td>33.8</td>
<td>32.6</td>
<td>32.7</td>
<td>22.2</td>
<td>37.8</td>
<td>21.1</td>
<td>27.2</td>
<td>27.7</td>
<td>31.4</td>
<td>34.1</td>
<td>21.8</td>
<td>23.7</td>
</tr>
<tr>
<td>XLM-100</td>
<td>34.3</td>
<td>26.7</td>
<td>28.5</td>
<td>29.3</td>
<td>28.3</td>
<td>27.2</td>
<td>29.9</td>
<td>21.1</td>
<td>28.6</td>
<td>22.1</td>
<td>26.6</td>
<td>26.3</td>
<td>25.1</td>
<td>30.9</td>
<td>20.1</td>
<td>21.7</td>
</tr>
<tr>
<td>XLM-R_B</td>
<td>51.5</td>
<td>44.1</td>
<td>42.1</td>
<td>44.8</td>
<td>44.0</td>
<td>43.3</td>
<td>39.5</td>
<td>42.6</td>
<td>40.6</td>
<td>34.6</td>
<td>40.2</td>
<td>38.4</td>
<td>37.5</td>
<td>43.4</td>
<td>29.6</td>
<td>33.0</td>
</tr>
<tr>
<td>XLM-R_L</td>
<td>66.7</td>
<td>56.1</td>
<td>58.2</td>
<td>59.5</td>
<td>60.3</td>
<td>56.8</td>
<td>52.1</td>
<td>51.4</td>
<td>52.7</td>
<td>48.7</td>
<td>53.9</td>
<td>48.4</td>
<td>50.0</td>
<td>59.9</td>
<td>41.6</td>
<td>45.2</td>
</tr>
</tbody>
</table>

### Unseen langs in MCP.
- Low-resource.

<table>
<thead>
<tr>
<th>Model</th>
<th>en</th>
<th>de</th>
<th>it</th>
<th>es</th>
<th>fr</th>
<th>nl</th>
<th>ru</th>
<th>vi</th>
<th>zh</th>
<th>hi</th>
<th>pl</th>
<th>ar</th>
<th>ja</th>
<th>pt</th>
<th>sw</th>
<th>uk</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCP(XLM-R_B)</td>
<td>52.1</td>
<td>46.2</td>
<td>45.6</td>
<td>44.3</td>
<td>44.7</td>
<td>45.3</td>
<td>42.8</td>
<td>45.3</td>
<td>44.3</td>
<td>36.8</td>
<td>41.4</td>
<td>36.8</td>
<td>37.5</td>
<td>44.9</td>
<td>28.1</td>
<td>37.4</td>
</tr>
<tr>
<td>MCP(XLM-R_L)</td>
<td>69.5</td>
<td>59.3</td>
<td>60.3</td>
<td>61.4</td>
<td>60.0</td>
<td>61.1</td>
<td>57.5</td>
<td>55.7</td>
<td>56.7</td>
<td>51.3</td>
<td>56.1</td>
<td>52.3</td>
<td>50.2</td>
<td>60.7</td>
<td>43.3</td>
<td>48.8</td>
</tr>
</tbody>
</table>

### Δ(XLM-R_L)
- +2.8 +3.3 +2.2 +1.9 -0.4 +4.3 +5.4 +4.3 +4.0 +2.6 +2.1 +3.9 +0.2 +0.8 +1.7 +3.6 +2.7
Take-home Messages

- **MickeyProbe** task can analyze commonsense knowledge of ML-LMs.
  - Multilingual, sentence-level extension of LAMA probing.
  - It can compare performance across ML-LMs & across languages.
- **MickeyCorpus** is a large multilingual corpus for studying common sense beyond English and English LMs. (561k sentences, in 11 languages.)
- **X-CSR** is a benchmark with two datasets (X-CSQA & X-CODAH) for zero-shot x-l transfer.
- **MCP** is a method to improve sentence rep. via multilingual contrastive pre-training.
  - Reusing the Mickey corpus for a multilingual sentence selection task.
  - Largely improve the X-CSR performance, even for unseen langs and low-resource langs.
Thank you very much for listening!
Please check our paper and project website for more info.

http://inklab.usc.edu/XCSR/