Computational Multilingualism in the Era of Large Language Models

Freda Shi
Toyota Technological Institute at Chicago
University of Waterloo & Vector Institute (starting July 2024)
My Research in the Context of Science

Computation

Language
Computational Linguistics

Language Model
Sciences of Language Models
This Talk: Computational Multilingualism

Computation

Language Model

Sciences of Language Models

Language Computational Linguistics

3
Research Question:

- How much translation exists in monolingual corpora? Can we do unsupervised translation?
- How many mutually translatable word pairs can we find given only monolingual corpora, and optionally a small seed lexicon?

<table>
<thead>
<tr>
<th>English</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>bank</td>
<td>orilla</td>
</tr>
<tr>
<td>shore</td>
<td>orilla</td>
</tr>
<tr>
<td>bank</td>
<td>banco</td>
</tr>
</tbody>
</table>
Background: Pretrained Multilingual Models

[Figure credit: Sebastian Ruder]
Background: Pretrained Multilingual Models

In the vector space, sentences with similar meanings are closer, regardless of their language or writing system.

Hypotheses on the reason: code-switching data, shared vocabulary, ...

We can encode sentences and retrieve pseudo bitext in the vector space.
Background: Cross-Lingual Word Aligner

Tokens are subwords in practice. Many-to-one alignment is possible.

No supervision is required!

☑️: Alignment
(horizontal and vertical maximum)

[SimAlign: Sabet et al., 2020]
Statistics from Language Models & Pseudo Bitext

Aligned (Pseudo) Bitext

Das ist eine Katze .
This is a cat .

Guten Morgen .
Good morning .

Guten Abend .
Good evening .

Danke .
Thank you .

Features for a bilingual pair \( s, t \):

\#Cooccurrence(\( s, t \)) -- How many times \( s \) and \( t \) appear in a pair of (pseudo) bitext.

\#align-1(\( s, t \)) -- How many times \( s \) and \( t \) are matched in one-to-one alignments.

\#align-many(\( s, t \)) -- How many times \( s \) and \( t \) are matched in one-to-many alignments.

Cosine similarity, inner product, \#count(\( s \)), \#count(\( t \)), …
Learning with Simple Statistics

Weakly Supervised Induction
A few bilingual lexicon entries are available.

Positive Examples:
\[ \mathcal{D}_+ = \{ \langle s, t \rangle \in \text{Lexicon} \} \]

Negative Examples:
\[ \mathcal{D}_- = \{ \langle s, t \rangle \notin \text{Lexicon, cooccurrence > 0} \} \]

\[ \max_\Theta \sum_{\langle s, t \rangle \in \mathcal{D}_+} \log P(s, t) + \sum_{\langle s', t' \rangle \in \mathcal{D}_-} \log(1 - P(s', t')) \]

Statistical Features between \( \langle s, t \rangle \)

Multi-Layer Perceptron (\( \Theta \))

\[ P(s, t) = P(\langle s, t \rangle \in \text{Lexicon}) \]
Results on the BUCC 2020 Shared Task

$F_1$ score: Harmonic mean between the precision and recall of the predicted lexicon. Averaged across DE-FR, EN-FR, EN-DE, EN-ES, EN-ZH, EN-RU and their reversed language pairs.

<table>
<thead>
<tr>
<th>Method</th>
<th>$F_1$ Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best BUCC</td>
<td>58.4</td>
</tr>
<tr>
<td>VecMap</td>
<td>41.5</td>
</tr>
<tr>
<td>WikiMatrix</td>
<td>72.0</td>
</tr>
<tr>
<td>Unsupervised Bitext</td>
<td>73.3</td>
</tr>
</tbody>
</table>
Looking into the “False Positive” Cases

<table>
<thead>
<tr>
<th>Ours</th>
<th>VecMap</th>
</tr>
</thead>
<tbody>
<tr>
<td>倉庫 depot</td>
<td>申明 endorsing declare</td>
</tr>
<tr>
<td>浪費 wasting</td>
<td>條件 preconditions condition</td>
</tr>
<tr>
<td>背面 reverse</td>
<td>天津 shanghai tianjin</td>
</tr>
<tr>
<td>嘴巴 mouths</td>
<td>個案 cases</td>
</tr>
<tr>
<td>可笑 laughable</td>
<td>百合 peony lily</td>
</tr>
<tr>
<td>隱藏 conceal</td>
<td>申報 filing</td>
</tr>
<tr>
<td>虔誠 devout</td>
<td>車廂 carriages</td>
</tr>
<tr>
<td>純淨 purified</td>
<td>海草 seaweed</td>
</tr>
<tr>
<td>截止 deadline</td>
<td>收容所 asylums</td>
</tr>
<tr>
<td>鍾 clocks</td>
<td>開幕 soft-opened opening</td>
</tr>
</tbody>
</table>

-green: Acceptable
-orange: Unacceptable
-yellow: Acceptable (in a certain context)
Takeaways

Q: How much translation exists in **monolingual corpora**? Can we do unsupervised translation?

A: A large portion of word-level translations may exist in unsupervised data.

Contextualized representations may help on (seemingly) non-contextual tasks. The simple MLP with statistical feature approach also improves word alignment.
This Talk: Computational Multilingualism

- Computation
- Multiple Languages
  - Bilingual Lexicon Induction
- Multilingual Model
  - Multilingual Math Reasoning
Multilingual Math Reasoning

Research Question:

❑ How well can language models do on reasoning with different languages?

We extend GSM8K to the multilingual grade-school math (MGSM) benchmark. MGSM covers 10 languages: Bengali, Mandarin Chinese, French, German, Japanese, Russian, Spanish, Swahili, Telugu and Thai.

Question: Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now?

Frage: Shawn hat fünf Spielzeuge. Zu Weihnachten hat er von seiner Mama und seinem Papa jeweils zwei Spielzeuge bekommen. Wie viele Spielzeuge hat er jetzt?
Native Chains of Thought

Frage: Shawn hat fünf Spielzeuge. Zu Weihnachten hat er von seiner Mama und seinem Papa jeweils zwei Spielzeuge bekommen. Wie viele Spielzeuge hat er jetzt?


Frage: Roger hat 5 Tennisbälle. Er kauft noch 2 Dosen Tennisbälle. In jeder Dose sind 3 Tennisbälle. Wie viele Tennisbälle hat er jetzt?

Schritt-für-Schritt-Antwort:

English Chains of Thought

Frage: Shawn hat fünf Spielzeuge. Zu Weihnachten hat er von seiner Mama und seinem Papa jeweils zwei Spielzeuge bekommen. Wie viele Spielzeuge hat er jetzt?

Step-by-Step Answer: He has 5 toys. He got 2 from mom, so after that he has 5 + 2 = 7 toys. Then he got 2 more from dad, so in total he has 7 + 2 = 9 toys. The answer is 9.

Frage: Roger hat 5 Tennisbälle. Er kauft noch 2 Dosen Tennisbälle. In jeder Dose sind 3 Tennisbälle. Wie viele Tennisbälle hat er jetzt?

Step-by-Step Answer:
Experiment 1: Native vs. English Reasoning Steps

Problem solution accuracy (%) on MGSM.

English, as the CoT solution language, generally outperforms the native language.
### English-Only Exemplars

**Question**: Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now?

**Step-by-Step Answer**: He has 5 toys. He got 2 from mom, so after that he has $5 + 2 = 7$ toys. Then he got 2 more from dad, so in total he has $7 + 2 = 9$ toys. The answer is 9.

... (5 more English examples)

### Multilingual Exemplars

**Question**: Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now?

**Step-by-Step Answer**: He has 5 toys. He got 2 from mom, so after that he has $5 + 2 = 7$ toys. Then he got 2 more from dad, so in total he has $7 + 2 = 9$ toys. The answer is 9.

... (5 examples, questions in DE, FR, ES, RU, ZH, respectively; solutions are in English.)
Experiment 2: English vs. Multilingual Exemplars

Multilingual exemplars work better than English for all languages except English.
What’s Next

• People use different prepositions to describe the same spatial relations [Feist & Gentner, 2003].

<table>
<thead>
<tr>
<th>Language</th>
<th>Preposition</th>
<th>English</th>
<th>Spanish</th>
<th>Japanese</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>over</td>
<td>sobre</td>
<td>ue</td>
</tr>
<tr>
<td></td>
<td></td>
<td>on</td>
<td>en</td>
<td>ue</td>
</tr>
<tr>
<td></td>
<td></td>
<td>in</td>
<td>en</td>
<td>naka</td>
</tr>
</tbody>
</table>
What’s Next: Computation for Language Sciences

• People use different prepositions to describe the same spatial relations [Feist & Gentner, 2003].
• Corpus, computation, and large language models provide new toolkits to better discover and understand these phenomena.
What’s Next

Language may affect how humans view the world.

Do language models trained on these languages have similar biases?  
Can we explain language model behaviors with machine learning theory and/or interpretability techniques?  
Can the results shed light on explaining human language acquisition? ...
Human experiments are usually expensive.
Language models are, to some extent, models of humans.

Experiments on faithful language models may even serve as the (pilot) pilot study for human experiments.
Thanks!