What Kind of Language Is Hard to Language-Model?

ACL 2019

Sebastian J. Mielke and Ryan Cotterell, Kyle Gorman, Brian Roark, Jason Eisner

Johns Hopkins University // City University of New York Graduate Center // Google
sjmielke@jhu.edu

Twitter: @sjmielke – paper and thread pinned!
0. Do current language models do equally well on all languages?
Questions and answers

0. Do current language models do equally well on all languages? No.

1. Which one do they struggle more with: German or English? German.


4. Is Translationese easier? It's different, but not actually easier!
0. Do current language models do equally well on all languages? No.

1. Which one do they struggle more with: German or English?
0. Do current language models do equally well on all languages? No.

1. Which one do they struggle more with: German or English? German.
0. Do current language models do equally well on all languages? No.

1. Which one do they struggle more with: German or English? German.

2. What about non-Indo-European languages, say Chinese?
0. Do current language models do equally well on all languages? No.

1. Which one do they struggle more with: German or English? German.

Questions and answers

0. Do current language models do equally well on all languages? 
   No.

1. Which one do they struggle more with: German or English? 
   German.

2. What about non-Indo-European languages, say Chinese? 
   It depends.

3. What makes a language harder to model?
Questions and answers

0. Do current language models do equally well on all languages? No.

1. Which one do they struggle more with: German or English? German.


Questions and answers

0. Do current language models do equally well on all languages?  
   No.

1. Which one do they struggle more with: German or English?  
   German.

2. What about non-Indo-European languages, say Chinese?  
   It depends.

3. What makes a language harder to model?  
   Actually, rather technical factors.

4. Is Translationese easier?
Questions and answers

0. Do current language models do equally well on all languages?  

No.

1. Which one do they struggle more with: German or English?  

German.

2. What about non-Indo-European languages, say Chinese?  

It depends.

3. What makes a language harder to model?  

Actually, rather technical factors.

4. Is Translationese easier?  

It’s different, but not actually easier!
“Difficulty”
Outline

“Difficulty”

Models and languages
Outline

“Difficulty”

Models and languages

What correlates with difficulty?
“Difficulty”

Models and languages

What correlates with difficulty?

And... is Translationese really easier?
How to measure “difficulty”? Language models measure surprisal/information content (NLL; $-\log p(\cdot)$):

\[
p(\cdot) \Rightarrow \text{NLL}
\]

<table>
<thead>
<tr>
<th>Language</th>
<th>Sentence</th>
<th>NLL</th>
<th>Bits</th>
</tr>
</thead>
<tbody>
<tr>
<td>en</td>
<td>I love Florence!</td>
<td>0.03</td>
<td>5 bits</td>
</tr>
<tr>
<td>de</td>
<td>Ich grüße meine Oma und die Familie dahein.</td>
<td>0.008</td>
<td>7 bits</td>
</tr>
<tr>
<td>nl</td>
<td>Alle mensen worden vrij en gelijk in waardigheid en rechten geboren.</td>
<td>0.0004</td>
<td>11 bits</td>
</tr>
</tbody>
</table>

Issue 1: Different topics/styles/content
Solution: train and test on translations!

Europarl: 21 languages share ~40M chars
Bibles: 62 languages share ~4M chars

and this one takes a big ILP to solve, which is really fun

Gurobi

\[
\sum_{69 \text{ languages}}^{13 \text{ language families}}
\]
How to measure “difficulty”?

Language models measure surprisal/information content (NLL; $-\log p(\cdot)$):

\[
p(\cdot) \Rightarrow \text{NLL}
\]

<table>
<thead>
<tr>
<th>Language</th>
<th>Sentence</th>
<th>NLL</th>
<th>Bits</th>
</tr>
</thead>
<tbody>
<tr>
<td>en</td>
<td>I love Florence!</td>
<td>0.03</td>
<td>5 bits</td>
</tr>
<tr>
<td>de</td>
<td>Ich grüße meine Oma und die Familie dahein.</td>
<td>0.008</td>
<td>7 bits</td>
</tr>
</tbody>
</table>

Issue 1: Different topics/styles/content
Solution: train and test on translations!

- Europarl: 21 languages share ~40M chars
- Bibles: 62 languages share ~4M chars

This one takes a big ILP to solve, which is really fun using Gurobi.

Issue 2: Comparing scores
Use total bits of an open-vocabulary model.

Why?
How to measure “difficulty”?

Language models measure surprisal/information content (NLL; $-\log p(\cdot)$):

\[
p(\cdot) \Rightarrow \text{NLL}
\]

<table>
<thead>
<tr>
<th>Language</th>
<th>Sentence</th>
<th>NLL</th>
<th>Bits</th>
</tr>
</thead>
<tbody>
<tr>
<td>en</td>
<td>I love Florence!</td>
<td>0.03</td>
<td>5</td>
</tr>
<tr>
<td>de</td>
<td>Ich grüße meine Oma und die Familie dahein.</td>
<td>0.008</td>
<td>7</td>
</tr>
<tr>
<td>nl</td>
<td>Alle mensen worden vrij en gelijk in waardigheid en rechten geboren.</td>
<td>0.0004</td>
<td>11</td>
</tr>
</tbody>
</table>
How to measure “difficulty”?

Language models measure surprisal/information content (NLL; $-\log p(\cdot)$):

\[
p(\cdot) \Rightarrow \text{NLL}
\]

<table>
<thead>
<tr>
<th>Language</th>
<th>Sentence</th>
<th>NLL</th>
<th>Bits</th>
</tr>
</thead>
<tbody>
<tr>
<td>en</td>
<td>I love Florence!</td>
<td>0.03</td>
<td>5</td>
</tr>
<tr>
<td>de</td>
<td>Ich grüße meine Oma und die Familie dahein.</td>
<td>0.008</td>
<td>7</td>
</tr>
<tr>
<td>nl</td>
<td>Alle mensen worden vrij en gelijk in waardigheid en rechten geboren.</td>
<td>0.0004</td>
<td>11</td>
</tr>
</tbody>
</table>

**Issue 1: Different topics/styles/content**
How to measure “difficulty”?  

Language models measure surprisal/information content (NLL; $-\log p(\cdot)$):

\[
p(\cdot) \Rightarrow \text{NLL}
\]

<table>
<thead>
<tr>
<th>Language</th>
<th>Translation</th>
<th>NLL</th>
<th>Bits</th>
</tr>
</thead>
<tbody>
<tr>
<td>en</td>
<td>Resumption of the session.</td>
<td>0.013</td>
<td>6.5</td>
</tr>
<tr>
<td>de</td>
<td>Wiederaufnahme der Sitzung.</td>
<td>0.011</td>
<td>6.3</td>
</tr>
<tr>
<td>nl</td>
<td>Hervatting van de sessie.</td>
<td>0.012</td>
<td>6.4</td>
</tr>
</tbody>
</table>

**Issue 1: Different topics/styles/content**

Solution: train and test on translations!

- Europarl: 21 languages share ~40M chars
- Bibles: 62 languages share ~4M chars
How to measure “difficulty”?

Language models measure surprisal/information content (NLL; $-\log p(\cdot)$):

\[
p(\cdot) \Rightarrow \text{NLL}
\]

<table>
<thead>
<tr>
<th>Language</th>
<th>Translation</th>
<th>NLL</th>
<th>Bits</th>
</tr>
</thead>
<tbody>
<tr>
<td>en</td>
<td>Resumption of the session.</td>
<td>0.013</td>
<td>6.5</td>
</tr>
<tr>
<td>de</td>
<td>Wiederaufnahme der Sitzung.</td>
<td>0.011</td>
<td>6.3</td>
</tr>
<tr>
<td>nl</td>
<td>Hervatting van de sessie.</td>
<td>0.012</td>
<td>6.4</td>
</tr>
</tbody>
</table>

**Issue 1: Different topics/styles/content**

Solution: train and test on translations!

Europarl: 21 languages share ~40M chars
Bibles: 62 languages share ~4M chars

\[\Rightarrow\] and this one takes a big ILP to solve, which is really fun.

Gurobi
How to measure “difficulty”? 

Language models measure surprisal/information content (NLL; $-\log p(\cdot)$): 

$$p(\cdot) \Rightarrow \text{NLL}$$

<table>
<thead>
<tr>
<th>Language</th>
<th>Sentence</th>
<th>Probability</th>
<th>NLL (bits)</th>
</tr>
</thead>
<tbody>
<tr>
<td>en</td>
<td>Resumption of the session.</td>
<td>0.013</td>
<td>6.5</td>
</tr>
<tr>
<td>de</td>
<td>Wiederaufnahme der Sitzung.</td>
<td>0.011</td>
<td>6.3</td>
</tr>
<tr>
<td>nl</td>
<td>Hervatting van de sessie.</td>
<td>0.012</td>
<td>6.4</td>
</tr>
</tbody>
</table>

**Issue 1: Different topics/styles/content**

Solution: train and test on translations!

- Europarl: 21 languages, 40M words
- Bibles: 62 languages, 4M words

\[\sum 69 \text{ languages} \quad 13 \text{ language families}\]

It takes a big ILP to solve, Gurobi, which is really fun.
How to measure “difficulty”?

Language models measure surprisal/information content (NLL; \(-\log p(\cdot)\)):

\[
p(\cdot) \Rightarrow \text{NLL}
\]

<table>
<thead>
<tr>
<th>Language</th>
<th>Description</th>
<th>NLL</th>
<th>Bits</th>
</tr>
</thead>
<tbody>
<tr>
<td>en</td>
<td>Resumption of the session.</td>
<td>0.013</td>
<td>6.5</td>
</tr>
<tr>
<td>de</td>
<td>Wiederaufnahme der Sitzung.</td>
<td>0.011</td>
<td>6.3</td>
</tr>
<tr>
<td>nl</td>
<td>Hervatting van de sessie.</td>
<td>0.012</td>
<td>6.4</td>
</tr>
</tbody>
</table>

Issue 1: Different topics/styles/content

Solution: train and test on translations!

Europarl: 21 languages share ~40M chars
Bibles: 62 languages share ~4M chars

\[\sum 69 \text{ languages} \quad \sum 13 \text{ language families}\]

Issue 2: Comparing scores

Use total bits of an open-vocabulary model. Why?
How to measure “difficulty”? 

Language models measure surprisal/information content (NLL; $-\log p(\cdot)$): 

$$ p(\cdot) \Rightarrow \text{NLL} $$

<table>
<thead>
<tr>
<th>Language</th>
<th>Text</th>
<th>NLL</th>
<th>Bits</th>
</tr>
</thead>
<tbody>
<tr>
<td>en</td>
<td>Resumption of the session.</td>
<td>0.013</td>
<td>6.5 bits</td>
</tr>
<tr>
<td>de</td>
<td>Wiederaufnahme der Sitzung.</td>
<td>0.011</td>
<td>6.3 bits</td>
</tr>
<tr>
<td>nl</td>
<td>Hervatting van de sessie.</td>
<td>0.012</td>
<td>6.4 bits</td>
</tr>
</tbody>
</table>

**Issue 1: Different topics/styles/content**

Solution: train and test on translations!

Europarl: 21 languages, 40M characters
Bibles: 62 languages, 4M characters

This takes a big ILP to solve, which is really fun with Gurobi.

**Issue 2: Comparing scores**

Use total bits of an open-vocabulary model.

Why?
How to compare your language models across languages

1. We need to be open-vocabulary – no UNKs.
1. **We need to be open-vocabulary – no UNKs.**
   Every UNK is “cheating” – morphologically rich languages have more UNKs, unfairly advantaging them.
1. **We need to be open-vocabulary – no UNKs.**
   Every UNK is “cheating” – morphologically rich languages have more UNKs, unfairly advantaging them.

2. **We can’t normalize per word or even per character in languages individually.**
How to compare your language models across languages

1. **We need to be open-vocabulary – no UNKs.**
   Every UNK is “cheating” – morphologically rich languages have more UNKs, unfairly advantaging them.

2. **We can’t normalize per word or even per character in languages individually.**
   Example: if puč_{cz} and Putsch_{de} are equally likely, they should be equally “difficult.”

⇒ just use overall bits (i.e., surprisal/NLL) of an aligned sentence

\[ \text{note: total easily obtainable from BPC or perplexity by multiplying with total chars/words} \]
1. **We need to be open-vocabulary – no UNKs.**
   Every UNK is “cheating” – morphologically rich languages have more UNKs, unfairly advantaging them.

2. **We can’t normalize per word or even per character in languages individually.**
   Example: if puč\_cz and Putsch\_de are equally likely, they should be equally “difficult.”

⇒ just use overall bits (i.e., surprisal/NLL) of an aligned sentence
1. **We need to be open-vocabulary – no UNKs.**
   Every UNK is “cheating” – morphologically rich languages have more UNKs, unfairly advantaging them.

2. **We can’t normalize per word or even per character in languages individually.**
   Example: if puč\textsubscript{cz} and Putsch\textsubscript{de} are equally likely, they should be equally “difficult.”

⇒ **Just use overall bits (i.e., surprisal/NLL) of an aligned sentence**
   [note: total easily obtainable from BPC or perplexity by multiplying with total chars/words]
How to aggregate multiple intents’ surprisals into “difficulties”?

For fully parallel corpora...

<table>
<thead>
<tr>
<th>en</th>
<th>de</th>
<th>bg</th>
</tr>
</thead>
</table>
| Resumption of the session | Wiederaufnahme der ... | Възобновяване на се- ...
| The peace that ... | Der gestern verein- ... | Мирът, който бе... |
| Although we were not al- ... | Obwohl wir nicht ... | Макар че не бяхме ... |
| Now we can fi- nally ... | Jetzt ist die Zeit ... | Накрая всички можем ... |

aligned multi-text
How to aggregate multiple intents’ surprisals into “difficulties”?

For fully parallel corpora...

```
<table>
<thead>
<tr>
<th>en</th>
<th>de</th>
<th>bg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resumption of the</td>
<td>Wiederaufnahme</td>
<td>Въобновяване на</td>
</tr>
<tr>
<td>session</td>
<td>der</td>
<td>се-</td>
</tr>
<tr>
<td>The peace that</td>
<td>Der</td>
<td>Мирът, който</td>
</tr>
<tr>
<td>...</td>
<td>gestern</td>
<td>беше</td>
</tr>
<tr>
<td>Although we were</td>
<td>Obwohl</td>
<td>Макар че не</td>
</tr>
<tr>
<td>not all</td>
<td>wir</td>
<td>бяхме</td>
</tr>
<tr>
<td>...</td>
<td>nicht</td>
<td>...</td>
</tr>
<tr>
<td>Now we can finally</td>
<td>Jetzt ist die Zeit</td>
<td>Накрая всички можем</td>
</tr>
<tr>
<td>...</td>
<td>Zeit</td>
<td>...</td>
</tr>
</tbody>
</table>
```

aligned multi-text

LM surprisals/NLLs

```
<table>
<thead>
<tr>
<th></th>
<th>en</th>
<th>de</th>
<th>bg</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$Y_1$</td>
<td>$Y_1$</td>
<td>$Y_1$</td>
</tr>
<tr>
<td></td>
<td>$Y_1$</td>
<td>$Y_1$</td>
<td>$Y_1$</td>
</tr>
<tr>
<td></td>
<td>$Y_2$</td>
<td>$Y_2$</td>
<td>$Y_2$</td>
</tr>
<tr>
<td></td>
<td>$Y_2$</td>
<td>$Y_2$</td>
<td>$Y_2$</td>
</tr>
<tr>
<td></td>
<td>$Y_3$</td>
<td>$Y_3$</td>
<td>$Y_3$</td>
</tr>
<tr>
<td></td>
<td>$Y_3$</td>
<td>$Y_3$</td>
<td>$Y_3$</td>
</tr>
<tr>
<td></td>
<td>$Y_4$</td>
<td>$Y_4$</td>
<td>$Y_4$</td>
</tr>
<tr>
<td></td>
<td>$Y_4$</td>
<td>$Y_4$</td>
<td>$Y_4$</td>
</tr>
</tbody>
</table>
```

This is a probabilistic model we can perform inference in!
How to aggregate multiple intents’ surprisals into “difficulties”?

For fully parallel corpora... we can just sum everything up and compare – that is *fair*.
How to aggregate multiple intents’ surprisals into “difficulties”?  

But what if there’s missing data? Or we want robustness?

<table>
<thead>
<tr>
<th>en</th>
<th>de</th>
<th>bg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resumption of the session</td>
<td>Wiederaufnahme der ...</td>
<td>...</td>
</tr>
<tr>
<td>The peace that ...</td>
<td>Der gestern verein ...</td>
<td>Мирът, който бяхме ...</td>
</tr>
<tr>
<td>Obwohl wir nicht ...</td>
<td>Макарче не бяхме ...</td>
<td>...</td>
</tr>
<tr>
<td>Now we can finally ...</td>
<td>Накрая всички можем ...</td>
<td>...</td>
</tr>
</tbody>
</table>

LM surprisals/NLLs

\[ \sum_{en} y_{1, en} \sum_{de} y_{1, de} \]

\[ \sum_{en} y_{2, en} \sum_{de} y_{2, de} \sum_{bg} y_{2, bg} \]

\[ \sum_{en} y_{3, de} \sum_{de} y_{3, bg} \]

\[ \sum_{en} y_{4, en} \sum_{bg} y_{4, bg} \]

aligned multi-text
How to aggregate multiple intents’ surprisals into “difficulties”?

But what if there’s missing data? Or we want robustness?

aligned multi-text

<table>
<thead>
<tr>
<th></th>
<th>en</th>
<th>de</th>
<th>bg</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Resumption of the session</td>
<td>Wieder- aufnahme der ...</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>The peace that ...</td>
<td>Der gestern verein- ...</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Obwohl wir nicht ...</td>
<td>Макарче небяхме ...</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Now we can finally ...</td>
<td>Накрая всички можем ...</td>
<td></td>
</tr>
</tbody>
</table>

LM surprisals/NLLs

\[ y_{1,\text{en}} \quad y_{1,\text{de}} \quad \Rightarrow n_1 \]

\[ y_{2,\text{en}} \quad y_{2,\text{de}} \quad y_{2,\text{bg}} \quad \Rightarrow n_2 \]

\[ y_{3,\text{de}} \quad y_{3,\text{bg}} \quad \Rightarrow n_3 \]

\[ y_{4,\text{en}} \quad y_{4,\text{bg}} \quad \Rightarrow n_4 \]

\[ d_{\text{en}} \quad d_{\text{de}} \quad d_{\text{bg}} \]
How to aggregate multiple intents’ surprisals into “difficulties”?

But what if there’s missing data? Or we want robustness?

<table>
<thead>
<tr>
<th>en</th>
<th>de</th>
<th>bg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resumption of the session</td>
<td>Wieder-annahme der ...</td>
<td></td>
</tr>
<tr>
<td>The peace that ...</td>
<td>Der gestern verein...</td>
<td>Мирът, който бяхме ...</td>
</tr>
<tr>
<td>Obwohl wir nicht ...</td>
<td>Макар че не бяхме ...</td>
<td></td>
</tr>
<tr>
<td>Now we can finally ...</td>
<td>Накрая всички можем ...</td>
<td></td>
</tr>
</tbody>
</table>

aligned multi-text

LM surprisals/NLLs

\[
\sum_{y_1\in\text{en}} n_1 \\
\sum_{y_2\in\text{en}} n_2 \\
\sum_{y_3\in\text{en}} n_3 \\
\sum_{y_4\in\text{en}} n_4 \\
\sum_{d_{\text{en}}} d_{\text{en}} \\
\sum_{d_{\text{de}}} d_{\text{de}} \\
\sum_{d_{\text{bg}}} d_{\text{bg}}
\]

\[
y_{2,\text{de}} \Rightarrow n_2 \\
y_{3,\text{de}} \Rightarrow n_3 \\
y_{4,\text{de}} \Rightarrow n_4
\]
How to aggregate multiple intents’ surprisals into “difficulties”?

But what if there’s missing data? Or we want robustness?

\[ \sum_{\text{en}} \sum_{\text{de}} \sum_{\text{bg}} \Rightarrow n_1 \]

\[ y_{2,\text{de}} = n_2 \]

\[ y_{2,\text{en}} y_{2,\text{de}} y_{2,\text{bg}} \Rightarrow n_2 \]

\[ y_{3,\text{de}} y_{3,\text{bg}} \Rightarrow n_3 \]

\[ y_{4,\text{en}} y_{4,\text{bg}} \Rightarrow n_4 \]

\[ d_{\text{en}} d_{\text{de}} d_{\text{bg}} \]
How to aggregate multiple intents’ surprisals into “difficulties”?

But what if there’s missing data? Or we want robustness?

```
resumption of the session

the peace that ...

obwohl wir nicht ...

now we can finally ...
```

```
y_{1,en} \quad y_{1,de} \quad \Rightarrow n_1

\quad y_{2,en} \quad y_{2,de} \quad y_{2,bg} \quad \Rightarrow n_2

\quad y_{3,de} \quad y_{3,bg} \quad \Rightarrow n_3

\quad y_{4,en} \quad y_{4,bg} \quad \Rightarrow n_4
```

```
d_{en} \quad d_{de} \quad d_{bg}
```

This is a probabilistic model we can perform inference in!

```
y_{i,j} = n_i \cdot \exp(d_{j})
```

```
\sigma_{i}^{2} = \ln(\epsilon_{i} + \exp(\sigma_{i}^{2})) - 1
```

```
\epsilon_{i,j} \sim \mathcal{N}(\sigma_{i}^{2} - \sigma_{i}^{2}, \sigma_{i}^{2})
```

This is a heteroscedastic model.
How to aggregate multiple intents’ surprisals into “difficulties”?

But what if there’s missing data? Or we want robustness?

This is a probabilistic model we can perform inference in!

### Aligned Multi-Text

<table>
<thead>
<tr>
<th>en</th>
<th>de</th>
<th>bg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resumption of the session</td>
<td>Wieder-</td>
<td>...</td>
</tr>
<tr>
<td>The peace that ...</td>
<td>gesteil-</td>
<td>...</td>
</tr>
<tr>
<td>Obwohl wir nicht ...</td>
<td>Мирт, който бяхме ...</td>
<td></td>
</tr>
<tr>
<td>Now we can finally ...</td>
<td>Накрая всички можем ...</td>
<td></td>
</tr>
</tbody>
</table>

### LM Surprisals/NLLs

\[
\begin{align*}
  y_1,_{en} & \Rightarrow n_1 \\
  y_1,_{de} & \\
  y_2,_{en} & \Rightarrow n_2 \\
  y_2,_{de} & \\
  y_2,_{bg} & \\
  y_2,_{bg} & \\
  y_3,_{en} & \Rightarrow n_3 \\
  y_3,_{de} & \\
  y_3,_{bg} & \\
  y_3,_{bg} & \\
  y_4,_{en} & \Rightarrow n_4 \\
  y_4,_{de} & \\
  y_4,_{bg} & \\
  d_{en} & \\
  d_{de} & \sim n_2 \cdot \exp d_{de} \\
  d_{bg} & \\
\end{align*}
\]

\[
\begin{align*}
  y_2,_{de} & \sim n_2 \cdot \exp d_{de} \\
\end{align*}
\]

\[
\begin{align*}
  y_3,_{de} & \sim n_3 \cdot \exp d_{de} \\
\end{align*}
\]

\[
\begin{align*}
  y_4,_{de} & \sim n_4 \cdot \exp d_{de} \\
\end{align*}
\]

\[
\begin{align*}
  y_4,_{bg} & \sim n_4 \cdot \exp d_{de} \\
\end{align*}
\]

\[
\begin{align*}
  y_4,_{bg} & \sim n_4 \cdot \exp d_{de} \\
\end{align*}
\]

\[
\begin{align*}
  d_{en} & \\
  d_{de} & \sim n_2 \cdot \exp d_{de} \\
  d_{bg} & \sim n_4 \cdot \exp d_{de} \\
\end{align*}
\]

\[
\begin{align*}
  d_{en} & \sim n_2 \cdot \exp d_{de} \\
  d_{de} & \sim n_2 \cdot \exp d_{de} \\
  d_{bg} & \sim n_4 \cdot \exp d_{de} \\
\end{align*}
\]

\[
\begin{align*}
  d_{en} & \sim n_2 \cdot \exp d_{de} \\
  d_{de} & \sim n_2 \cdot \exp d_{de} \\
  d_{bg} & \sim n_4 \cdot \exp d_{de} \\
\end{align*}
\]

\[
\begin{align*}
  d_{en} & \sim n_2 \cdot \exp d_{de} \\
  d_{de} & \sim n_2 \cdot \exp d_{de} \\
  d_{bg} & \sim n_4 \cdot \exp d_{de} \\
\end{align*}
\]

\[
\begin{align*}
  d_{en} & \sim n_2 \cdot \exp d_{de} \\
  d_{de} & \sim n_2 \cdot \exp d_{de} \\
  d_{bg} & \sim n_4 \cdot \exp d_{de} \\
\end{align*}
\]

\[
\begin{align*}
  d_{en} & \sim n_2 \cdot \exp d_{de} \\
  d_{de} & \sim n_2 \cdot \exp d_{de} \\
  d_{bg} & \sim n_4 \cdot \exp d_{de} \\
\end{align*}
\]

\[
\begin{align*}
  d_{en} & \sim n_2 \cdot \exp d_{de} \\
  d_{de} & \sim n_2 \cdot \exp d_{de} \\
  d_{bg} & \sim n_4 \cdot \exp d_{de} \\
\end{align*}
\]

\[
\begin{align*}
  d_{en} & \sim n_2 \cdot \exp d_{de} \\
  d_{de} & \sim n_2 \cdot \exp d_{de} \\
  d_{bg} & \sim n_4 \cdot \exp d_{de} \\
\end{align*}
\]

\[
\begin{align*}
  d_{en} & \sim n_2 \cdot \exp d_{de} \\
  d_{de} & \sim n_2 \cdot \exp d_{de} \\
  d_{bg} & \sim n_4 \cdot \exp d_{de} \\
\end{align*}
\]

\[
\begin{align*}
  d_{en} & \sim n_2 \cdot \exp d_{de} \\
  d_{de} & \sim n_2 \cdot \exp d_{de} \\
  d_{bg} & \sim n_4 \cdot \exp d_{de} \\
\end{align*}
\]

\[
\begin{align*}
  d_{en} & \sim n_2 \cdot \exp d_{de} \\
  d_{de} & \sim n_2 \cdot \exp d_{de} \\
  d_{bg} & \sim n_4 \cdot \exp d_{de} \\
\end{align*}
\]

\[
\begin{align*}
  d_{en} & \sim n_2 \cdot \exp d_{de} \\
  d_{de} & \sim n_2 \cdot \exp d_{de} \\
  d_{bg} & \sim n_4 \cdot \exp d_{de} \\
\end{align*}
\]

\[
\begin{align*}
  d_{en} & \sim n_2 \cdot \exp d_{de} \\
  d_{de} & \sim n_2 \cdot \exp d_{de} \\
  d_{bg} & \sim n_4 \cdot \exp d_{de} \\
\end{align*}
\]

\[
\begin{align*}
  d_{en} & \sim n_2 \cdot \exp d_{de} \\
  d_{de} & \sim n_2 \cdot \exp d_{de} \\
  d_{bg} & \sim n_4 \cdot \exp d_{de} \\
\end{align*}
\]

\[
\begin{align*}
  d_{en} & \sim n_2 \cdot \exp d_{de} \\
  d_{de} & \sim n_2 \cdot \exp d_{de} \\
  d_{bg} & \sim n_4 \cdot \exp d_{de} \\
\end{align*}
\]

\[
\begin{align*}
  d_{en} & \sim n_2 \cdot \exp d_{de} \\
  d_{de} & \sim n_2 \cdot \exp d_{de} \\
  d_{bg} & \sim n_4 \cdot \exp d_{de} \\
\end{align*}
\]

\[
\begin{align*}
  d_{en} & \sim n_2 \cdot \exp d_{de} \\
  d_{de} & \sim n_2 \cdot \exp d_{de} \\
  d_{bg} & \sim n_4 \cdot \exp d_{de} \\
\end{align*}
\]

\[
\begin{align*}
  d_{en} & \sim n_2 \cdot \exp d_{de} \\
  d_{de} & \sim n_2 \cdot \exp d_{de} \\
  d_{bg} & \sim n_4 \cdot \exp d_{de} \\
\end{align*}
\]

\[
\begin{align*}
  d_{en} & \sim n_2 \cdot \exp d_{de} \\
  d_{de} & \sim n_2 \cdot \exp d_{de} \\
  d_{bg} & \sim n_4 \cdot \exp d_{de} \\
\end{align*}
\]

\[
\begin{align*}
  d_{en} & \sim n_2 \cdot \exp d_{de} \\
  d_{de} & \sim n_2 \cdot \exp d_{de} \\
  d_{bg} & \sim n_4 \cdot \exp d_{de} \\
\end{align*}
\]
How to aggregate multiple intents’ surprisals into “difficulties”?

But what if there’s missing data? Or we want robustness?

LM surprisals/NLLs

\[
\begin{align*}
\sum_{en} y_1 \quad \sum_{de} y_2 \quad \sum_{bg} y_3
\end{align*}
\]

\[
\Rightarrow
\]

\[
\begin{align*}
d_{en} \quad d_{de} \quad d_{bg}
\end{align*}
\]

log-normal noise

not quite, our actual model is

\[
H E T E R O S C E D A S T I C
\]

\[
y_{ij} = n_i \cdot \exp(d_j) \cdot \exp(\epsilon_{ij})
\]

\[
\sigma_i^2 = \ln \left(1 + \frac{\exp(\sigma^2)}{\sigma^2 - \sigma_i^2 n_i}\right)
\]

\[
\epsilon_{ij} \sim \mathcal{N} \left(\frac{\sigma^2 - \sigma_i^2}{2}, \sigma_i^2\right)
\]

Image CC-BY Mike Grauer Jr / flickr
“Difficulty”

Models and languages

What correlates with difficulty?

And... is Translationese really easier?
Good open-vocabulary language models

Formerly state-of-the-art-ish AWD-LSTM (Merity et al., 2018) language models:

char-RNNLM:
Formerly state-of-the-art-ish AWD-LSTM (Merity et al., 2018) language models:

char-RNNLM:

BPE-RNNLM, few merges:
Formerly state-of-the-art-ish AWD-LSTM (Merity et al., 2018) language models:

char-RNNLM:

BPE-RNNLM, few merges:

BPE-RNNLM, many merges:
Choosing the number of BPE merges: how many is best?

It depends on the language (total surprisal, given merges as a ratio of the vocabulary):
Choosing the number of BPE merges: how many is best?

It depends on the language (total surprisal, given merges as a ratio of the vocabulary):

![Graph showing the relationship between the number of BPE merges and surprisal for various languages.](image-url)
Choosing the number of BPE merges: how many is best?

It depends on the language (total surprisal, given merges as a ratio of the vocabulary):

![Graph showing the ratio of BPE merges versus language]

is this one going to be fine?
Choosing the number of BPE merges: how many is best?

It depends on the language (total surprisal, given merges as a ratio of the vocabulary):

Yeah:

it doesn’t matter that much.

average

is this one going to be fine?
Difficulties for char-/BPE-RNNLM: 21 Europarl languages

Difficulties on Europarl

-8 -6 -4 -2 0 2 4 6 8 10

-4 -3 -2 -1 0 1 2 3 4 5

difficulty (×100) using char-RNNLM
difficulty (×100) using BPE-RNNLM with 0.4|V| merges
Difficulties for char-/BPE-RNNLM: 21 Europarl languages

easier with BPE

easier with chars
Difficulties for char-/BPE-RNNLM: 21 Europarl languages

difficulties on Europarl

easier with BPE

easier with chars

Difficulties on Bibles

difficulty (×100) using char-RNNLM

difficulty (×100) using BPE-RNNLM with 0.4|V| merges
Difficulties for char-/BPE-RNNLM: 21 Europarl languages

Difficulties on Europarl

easier with BPE

easier with chars
Di/uniFB03iculties for char-/BPE-RNNLM: 21 Europarl languages and Bibles

difficulty (x100) using BPE-RNNLM with 0.4|V| merges

easier with BPE

easier with chars

difficulty (x100) using BPE-RNNLM with 0.4|V| merges

easier with BPE

easier with chars
Difficulties for char-/BPE-RNNLM: 21 Europarl languages and 106 Bibles

Difficulties on Europarl

easier with BPE

easier with chars

difficulty (×100) using char-RNNLM

difficulty (×100) using BPE-RNNLM with 0.4|V| merges

difficulties for char-/BPE-RNNLM: 21 Europarl languages and 106 Bibles

Difficulties on Bibles

easier with BPE

easier with chars

difficulty (×100) using BPE-RNNLM with 0.4|V| merges
Difficulties for char-/BPE-RNNLM: 21 Europarl languages and 106 Bibles

Difficulties on Europarl

easier with BPE

easier with chars

difficulty (×100) using BPE-RNNLM with 0.4|V| merges

difficulty (×100) using char-RNNLM

Easier with BPE

Difficulties on Bibles

easier with BPE

easier with chars

cmn

difficulty (×100) using BPE-RNNLM with 0.4|V| merges

difficulty (×100) using char-RNNLM
Difficulties for char-/BPE-RNNLM: 21 Europarl languages and 106 Bibles

Difficulties on Europarl vs. Bibles

- **Easier with BPE**
- **Easier with chars**

**Difficulties on Europarl**

- **Harder**
- **Easier**

**Difficulties on Bibles**

- **Easier with BPE**
- **Easier with chars**

- Difficulty (×100) using BPE-RNNLM with 0.4|V| merges

- Difficulty (×100) using char-RNNLM
“Difficulty”

Models and languages

What correlates with difficulty?

And... is Translationese really easier?
How about: morphological counting complexity (Sagot, 2013)

...not particularly striking. Perhaps Finnish was an outlier in Cotterell et al. (2018)?
How about: morphological counting complexity (Sagot, 2013)

...not particularly striking. Perhaps Finnish was an outlier in Cotterell et al. (2018)?
Other linguistically motivated regressors

WALS: “Prefixing vs. Suffixing [...] Morphology” (for languages where present)?
WALS: “Prefixing vs. Suffixing [...] Morphology” (for languages where present)?
...no visible differences.
Other linguistically motivated regressors

WALS: “Prefixing vs. Suffixing [...] Morphology” (for languages where present)? ...no visible differences.

WALS: “Order of Subject, Object and Verb” (for languages where present)?
Other linguistically motivated regressors

WALS: “Prefixing vs. Suffixing [...] Morphology” (for languages where present)?
...no visible differences.

WALS: “Order of Subject, Object and Verb” (for languages where present)?
...no visible differences.
Other linguistically motivated regressors

**WALS: “Prefixing vs. Suffixing [...] Morphology”** (for languages where present)?
...no visible differences.

**WALS: “Order of Subject, Object and Verb”** (for languages where present)?
...no visible differences.

**Head-POS Entropy** (Dehouck and Denis, 2018)?

...neither mean and skew show correlation.

Average dependency length (computed using UDPipe (Straka et al., 2016))?
...correlation! But not significant after correcting for multiple hypotheses.

This is disappointing.
Other linguistically motivated regressors

WALS: “Prefixing vs. Suffixing [...] Morphology” (for languages where present)?
...no visible differences.

WALS: “Order of Subject, Object and Verb” (for languages where present)?
...no visible differences.

Head-POS Entropy (Dehouck and Denis, 2018)?
...neither mean and skew show correlation.
Other linguistically motivated regressors

WALS: “Prefixing vs. Suffixing [...] Morphology” (for languages where present)?
...no visible differences.

WALS: “Order of Subject, Object and Verb” (for languages where present)?
...no visible differences.

Head-POS Entropy (Dehouck and Denis, 2018)?
...neither mean and skew show correlation.

Average dependency length (computed using UDPipe (Straka et al., 2016))?
Other linguistically motivated regressors

**WALS: “Prefixing vs. Suffixing [...] Morphology”** (for languages where present)?
...no visible differences.

**WALS: “Order of Subject, Object and Verb”** (for languages where present)?
...no visible differences.

**Head-POS Entropy** (Dehouck and Denis, 2018)?
...neither mean and skew show correlation.

**Average dependency length** (computed using UDPipe (Straka et al., 2016))?  
...correlation! But not significant after correcting for multiple hypotheses.
Other linguistically motivated regressors

WALS: “Prefixing vs. Suffixing [...] Morphology” (for languages where present)?
...no visible differences.

WALS: “Order of Subject, Object and Verb” (for languages where present)?
...no visible differences.

Head-POS Entropy (Dehouck and Denis, 2018)?
...neither mean and skew show correlation.

Average dependency length (computed using UDPipe (Straka et al., 2016))? 
...correlation! But not significant after correcting for multiple hypotheses.

This is disappointing.
Very simple heuristics are very predictive

<table>
<thead>
<tr>
<th>Raw sequence length / # predictions</th>
<th>char-RNNLM difficulty</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Significant on:</strong></td>
<td></td>
</tr>
<tr>
<td>• Europarl at $p &lt; .01$</td>
<td></td>
</tr>
<tr>
<td>• Bibles at $p &lt; .001$</td>
<td></td>
</tr>
</tbody>
</table>

i.e., for the char-RNNLM

puč\textsubscript{cz} is easier than Putsch\textsubscript{de}!
Very simple heuristics are very predictive

Raw sequence **length** / # predictions
→ **char-RNNLM difficulty**

**Significant** on:
- Europarl at $p < .01$
- Bibles at $p < .001$

i.e., for the char-RNNLM
pučcz is easier than Putschde!

Raw **vocabulary size**
→ **BPE-RNNLM difficulty**

**Significant** on:
- not Europarl
- but Bibles at $p < .00000000001$

i.e., the BPE-RNNLM still suffers if a language has high type-token-ratio!
Very simple heuristics are very predictive

### Raw sequence length / # predictions → char-RNNLM difficulty

**Significant on:**
- Europarl at $p < .01$
- Bibles at $p < .001$

i.e., for the char-RNNLM
pučcz is easier than Putsch_de!

### Raw vocabulary size → BPE-RNNLM difficulty

**Significant on:**
- not Europarl
- but Bibles at $p < .00000000001$

i.e., the BPE-RNNLM still suffers if a language has high type-token-ratio!

Wow! What is happening here? We have many conjectures...
“Difficulty”

What correlates with difficulty?

And... is Translationese really easier?
Common assumption: Translationese is somehow simpler than “native” text.
Common assumption: *Translationese is somehow simpler than “native” text.*

We have partial parallel data that we can use to evaluate our models:

<table>
<thead>
<tr>
<th>en&lt;sub&gt;original&lt;/sub&gt;</th>
<th>en&lt;sub&gt;translated&lt;/sub&gt;</th>
<th>de&lt;sub&gt;original&lt;/sub&gt;</th>
<th>de&lt;sub&gt;translated&lt;/sub&gt;</th>
<th>nl&lt;sub&gt;original&lt;/sub&gt;</th>
<th>nl&lt;sub&gt;translated&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resumption...</td>
<td></td>
<td>Wiederauf...</td>
<td></td>
<td>Hervatten...</td>
<td></td>
</tr>
<tr>
<td>The German...</td>
<td></td>
<td>Der deutsche...</td>
<td></td>
<td>De Duitse...</td>
<td></td>
</tr>
<tr>
<td>Thank you...</td>
<td></td>
<td>Vielen Dank...</td>
<td></td>
<td>Hartelijk...</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td>...</td>
<td></td>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>
Translationese: translations as a separate language?

Common assumption: Translationese is somehow simpler than “native” text.

We have partial parallel data that we can use to evaluate our models:

| en_{original} | en_{translated} | de_{original} | de_{translated} | nl_{original} | nl_{translated} | ...
|---------------|-----------------|---------------|-----------------|---------------|-----------------|-----
| Resumption... |                 | Wiederauf...  |                 | Hervatten...  |                 |     
| The German... |                 | Der deutsche... |                | De Duitse...  |                 |     
| Thank you...  |                 | Vielen Dank... |                | Hartelijk...  |                 |     
| ...           |                 | ...           |                | ...           |                 |     

...and indeed the original languages **seem** harder.
Translationese: translations as a separate language?

Common assumption: Translationese is somehow simpler than “native” text.

We have partial parallel data that we can use to evaluate our models:

<table>
<thead>
<tr>
<th>en\textsubscript{original}</th>
<th>en\textsubscript{translated}</th>
<th>de\textsubscript{original}</th>
<th>de\textsubscript{translated}</th>
<th>nl\textsubscript{original}</th>
<th>nl\textsubscript{translated}</th>
<th>…</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resumption...</td>
<td></td>
<td>Wiederauf...</td>
<td></td>
<td>Hervatten...</td>
<td></td>
<td>…</td>
</tr>
<tr>
<td>The German...</td>
<td></td>
<td>Der deutsche...</td>
<td></td>
<td>De Duitse...</td>
<td></td>
<td>…</td>
</tr>
<tr>
<td>Thank you...</td>
<td></td>
<td>Vielen Dank...</td>
<td></td>
<td>Hartelijk...</td>
<td></td>
<td>…</td>
</tr>
<tr>
<td>…</td>
<td></td>
<td>…</td>
<td></td>
<td>…</td>
<td></td>
<td>…</td>
</tr>
</tbody>
</table>

...and indeed the original languages seem harder. But we missed something!
We trained on mostly translationese!

Of course we will then find it easier...
Repeat the experiment with fairly balancing training data.

Change the training sets!

We can **rebalance a single language**, leaving the others merged, i.e.:

<table>
<thead>
<tr>
<th>en\textsubscript{original}</th>
<th>en\textsubscript{translated}</th>
<th>de</th>
<th>nl</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resumption...</td>
<td></td>
<td>Wiederauf-...</td>
<td>Hervatten...</td>
<td>...</td>
</tr>
<tr>
<td>The German...</td>
<td></td>
<td>Der-deutsche...</td>
<td>De-Duitse...</td>
<td>...</td>
</tr>
<tr>
<td>Thank you...</td>
<td></td>
<td>Vielen Dank...</td>
<td>Hartelijk...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Repeat the experiment with fairly balancing training data

Change the training sets!

We can **rebalance a single language**, leaving the others merged, i.e.:

<table>
<thead>
<tr>
<th>en\textsubscript{original}</th>
<th>en\textsubscript{translated}</th>
<th>de</th>
<th>nl</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resumption...</td>
<td></td>
<td>Wiederauf-...</td>
<td>Hervatten...</td>
<td>...</td>
</tr>
<tr>
<td>The-German...</td>
<td></td>
<td>Der-deutsche...</td>
<td>De-Duitse...</td>
<td>...</td>
</tr>
<tr>
<td>Thank you...</td>
<td></td>
<td>Vielen Dank...</td>
<td>Hartelijk...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

And the result: the **difficulties are now the same**!

(more precisely, “native” is $0.004 \pm 0.02$ easier)
Conclusion: cross-linguistic comparisons are tricky  (hope we didn’t mess up!)
1. Make sure your training data is comparable and fair.
Conclusion: cross-linguistic comparisons are tricky (hope we didn’t mess up!)

1. Make sure your training data is comparable and fair.
2. Make sure your metrics are comparable and fair.
Conclusion: cross-linguistic comparisons are tricky (hope we didn’t mess up!)

1. Make sure your training data is comparable and fair.
2. Make sure your metrics are comparable and fair.
3. Make sure your stats are fair (no p-hacking!).
Conclusion: cross-linguistic comparisons are tricky  (hope we didn’t mess up!)

1. Make sure your training data is comparable and fair.
2. Make sure your metrics are comparable and fair.
3. Make sure your stats are fair (no p-hacking!).
4. Work on more NLP resources for more languages!
What Kind of Language Is Hard to Language-Model?
ACL 2019

Sebastian J. Mielke and Ryan Cotterell, Kyle Gorman, Brian Roark, Jason Eisner

Johns Hopkins University // City University of New York Graduate Center // Google
sjmielke@jhu.edu

Twitter: @sjmielke – paper and thread pinned!