It’s Easier to Translate out of English than into it: Measuring Neural Translation Difficulty by Cross-Mutual Information

ACL 2020

Emanuele Bugliarello, Sabrina J. Mielke, Antonios Anastasopoulos, Ryan Cotterell, Naoaki Okazaki
Is **fi-en** easier than **en-fi**?
Is **fi-en** easier than **en-fi**?

We can’t tell based on BLEU!
BLEU’s shortcomings for cross-linguistic comparisons
BLEU’s shortcomings for cross-linguistic comparisons

BLEU is a precision-based metric
BLEU’s shortcomings for cross-linguistic comparisons

BLEU is a precision-based metric

1. BLEU depends on *tokenization* and the *notion of “word”*!
BLEU’s shortcomings for cross-linguistic comparisons

BLEU is a precision-based metric

1. BLEU depends on tokenization and the notion of “word”!

   **Example:**
   “I will have been programming”  English
   “Programlayacağım”  Turkish
BLEU’s shortcomings for cross-linguistic comparisons

BLEU is a precision-based metric

1. BLEU depends on *tokenization* and the *notion of “word”*

   **Example:**
   
   “I will have been programming”  English
   “Programlayacağım”  Turkish

   ➞ More partial credit for English!
BLEU’s shortcomings for cross-linguistic comparisons

BLEU is a precision-based metric

1. BLEU depends on *tokenization* and the *notion of “word”*

   Example:
   
   “I will have been programming”  
   “Programlayacağım”

   → More partial credit for English!

   Remedy: Look at the likelihood
BLEU’s shortcomings for cross-linguistic comparisons

BLEU is a precision-based metric

1. BLEU depends on tokenization and the notion of “word”!

   **Example:**
   
   “I will have been programming”  English
   “Programlayacağım”  Turkish

   → More partial credit for English!

   **Remedy:** Look at the likelihood

2. We are still measuring: difficulty of translation and generation
Mutual Information expresses the act of translation

**Entropy:**

\[
H(T) = \mathbb{E}_{t \sim p(T)}[-\log_2(p(t))] \quad \text{uncertainty}
\]

\[
H(T|S) = \mathbb{E}_{(s,t) \sim p(S,T)}[-\log_2(p(t|s))]
\]
Mutual Information expresses the act of translation

Entropy: $H(T) = \mathbb{E}_{t \sim p(T)}[-\log_2(p(t))]$ uncertainty

$H(T | S) = \mathbb{E}_{(s,t) \sim p(S,T)}[-\log_2(p(t | s))]$

$H(T)$

uncertainty about $T$

*a priori*
**Mutual Information** expresses the act of translation

**Entropy:**

\[ H(T) = \mathbb{E}_{t \sim p(T)}[-\log_2(p(t))] \]

uncertainty about \( T \)

\[ H(T \mid S) = \mathbb{E}_{(s,t) \sim p(S,T)}[-\log_2(p(t \mid s))] \]

uncertainty about \( T \) after knowing \( S \)
Mutual Information expresses the act of translation

\[ H(T) = \mathbb{E}_{t \sim p(T)}[-\log_2(p(t))] \]

\[ H(T | S) = \mathbb{E}_{(s,t) \sim p(S,T)}[-\log_2(p(t | s))] \]

\[ H(T) - H(T | S) \]

- uncertainty about \( T \)
- uncertainty about \( T \)
- uncertainty about \( T \)
- uncertainty about \( T \)

- a priori
- after knowing \( S \)

how much knowing \( S \) reduced uncertainty about \( T \)
Mutual Information expresses the act of translation

**Entropy:**

\[ H(T) = \mathbb{E}_{t \sim p(T)}[-\log_2(p(t))] \quad \text{uncertainty about } T \]

\[ H(T | S) = \mathbb{E}_{(s, t) \sim p(S, T)}[-\log_2(p(t | s))] \]

\[
\text{MI}(S; T) = H(T) - H(T | S)
\]

- **Mutual Information** between \( S \) and \( T \)
- uncertainty about \( T \) *a priori*
- uncertainty about \( T \) *after knowing \( S \)*
- how much knowing \( S \) reduced uncertainty about \( T \)
Mutual Information expresses the act of translation

Mutual Information: $\text{MI}(S; T) = \text{H}(T) - \text{H}(T ∣ S)$

Entropy: $H(T) = \mathbb{E}_{t ∼ p(T)}[−\log_2(p(t))]$  
$H(T ∣ S) = \mathbb{E}_{(s,t) ∼ p(S,T)}[−\log_2(p(t ∣ s))]$

- Mutual Information between $S$ and $T$
- Uncertainty about $T$ 
  - a priori
  - after knowing $S$
- How much knowing $S$ reduced uncertainty about $T$

Symmetric! assuming all entropies w.r.t. same joint $p(S, T)$
Mutual Information expresses the *act of translation*

**Entropy:**
\[
H(T) = \mathbb{E}_{t \sim p(T)}[-\log_2(p(t))] \quad \text{uncertainty about } T
\]

\[
H(T | S) = \mathbb{E}_{(s,t) \sim p(S,T)}[-\log_2(p(t | s))] \quad \text{uncertainty about } T \text{ after knowing } S
\]

**Mutual Information (MI):**
\[
\text{MI}(S; T) = H(T) - H(T | S)
\]

- **Mutual Information** 
  - between $S$ and $T$
- **Entropy of $T$** 
  - *a priori* uncertainty about $T$
- **Conditional Entropy $H(T | S)$** 
  - *after knowing $S$* uncertainty about $T$

**Example:** en-zh
Mutual Information expresses the act of translation

**Entropy:**

\[ H(T) = \mathbb{E}_{t \sim p(T)}[-\log_2(p(t))] \]

\[ H(T | S) = \mathbb{E}_{(s,t) \sim p(S,T)}[-\log_2(p(t | s))] \]

\[
\text{MI}(S; T) = \underbrace{H(T)}_{\text{uncertainty about } T \text{ } a \text{ priori}} - \underbrace{H(T | S)}_{\text{uncertainty about } T \text{ after knowing } S}
\]

**Mutual Information**

between \( S \) and \( T \)

\( \text{how much knowing } S \text{ reduced uncertainty about } T \)

**Example:** en-zh

\[ H(谢谢) \]

uncertainty about “谢谢”

**symmetric!**

assuming all entropies w.r.t. same joint \( p(S, T) \)
**Mutual Information** expresses the act of translation

**Entropy:**
\[
H(T) = \mathbb{E}_{t \sim p(T)}[- \log_2(p(t))] \quad \text{uncertainty}
\]
\[
H(T \mid S) = \mathbb{E}_{(s,t) \sim p(S,T)}[- \log_2(p(t \mid s))] \quad \text{after knowing } S
\]

\[
\text{MI}(S; T) = H(T) - H(T \mid S) \quad \text{symmetric! assuming all entropies w.r.t. same joint } p(S,T)
\]

- **mutual information** between \( S \) and \( T \)
- uncertainty about \( T \) \textit{a priori}
- uncertainty about \( T \) \textit{after knowing } \( S \)
- how much knowing \( S \) reduced uncertainty about \( T \)

**Example:** en-zh

- \( H(谢谢) \)
- \( H(谢谢 \mid \text{Thanks}) \)

uncertainty about “谢谢”

uncertainty about “谢谢” after knowing its translation
Mutual Information expresses the act of translation

Entropy: $H(T) = \mathbb{E}_{t \sim p(T)}[-\log_2(p(t))]$ uncertainty

$H(T \mid S) = \mathbb{E}_{(s,t) \sim p(S,T)}[-\log_2(p(t \mid s))]$

$\text{MI}(S; T) = H(T) - H(T \mid S)$

Mutual information between $S$ and $T$

- uncertainty about $T$ a priori
- uncertainty about $T$ after knowing $S$

how much knowing $S$ reduced uncertainty about $T$

Example: en-zh

$H(谢谢)$

uncertainty about “谢谢”

$H(谢谢 \mid \text{Thanks})$

uncertainty about “谢谢” after knowing its translation

$\text{MI}(\text{Thanks};谢谢)$

how much easier it has become to predict “谢谢”
Cross-Mutual Information measures models’ performance on the act of translation

Entropy: $H(T) = \mathbb{E}_{t \sim p(T)}[-\log_2(p(t))]$ uncertainty

$H(T|S) = \mathbb{E}_{(s,t) \sim p(S,T)}[-\log_2(p(t|s))]$

$\text{MI}(S; T) = H(T) - H(T|S)$

- $H(T)$: uncertainty about $T$ a priori
- $H(T|S)$: uncertainty about $T$ after knowing $S$
- $\text{MI}(S; T)$: mutual information between $S$ and $T$
- how much knowing $S$ reduced uncertainty about $T$
**Cross-Mutual Information** measures models’ performance on the act of translation

**Entropy:**
- $H(T) = \mathbb{E}_{t \sim p(T)}[-\log_2(p(t))]$ uncertainty
- $H(T \mid S) = \mathbb{E}_{(s,t) \sim p(S,T)}[-\log_2(p(t \mid s))]$

**MI(S; T) =** $H(T) - H(T \mid S)$

- **Mutual information** between $S$ and $T$
- **Uncertainty about $T$ a priori**
- **Uncertainty about $T$ after knowing $S$**
- **How much knowing $S$ reduced uncertainty about $T$**

**Cross-Entropy:**
- $H_q(T) = \mathbb{E}_{t \sim p(T)}[-\log_2(q(t))]$ how surprised is model $q$ in reality $p$?
- $H_q(T \mid S) = \mathbb{E}_{(s,t) \sim p(S,T)}[-\log_2(q(t \mid s))]$
**Cross-Mutual Information** measures models’ performance on the act of translation

**Entropy:** $H(T) = \mathbb{E}_{t \sim p(T)}[-\log_2(p(t))]$ uncertainty

$H(T \mid S) = \mathbb{E}_{(s,t) \sim p(S,T)}[-\log_2(p(t \mid s))]$

**MI($S; T$) = $H(T) - H(T \mid S)$**

- mutual information between $S$ and $T$
- uncertainty about $T$ a priori
- uncertainty about $T$ after knowing $S$
- how much knowing $S$ reduced uncertainty about $T$

**Cross-Entropy:** $H_q(T) = \mathbb{E}_{t \sim p(T)}[-\log_2(q(t))]$ how surprised is model $q$ in reality $p$?

$H_q(T \mid S) = \mathbb{E}_{(s,t) \sim p(S,T)}[-\log_2(q(t \mid s))]$

**XMI($S \rightarrow T$) := $H_{q_{LM}}(T) - H_{q_{MT}}(T \mid S)$**
Cross-Mutual Information measures models’ performance on the act of translation

**Entropy:**

\[
H(T) = \mathbb{E}_{t \sim p(T)}[-\log_2(p(t))] \quad \text{uncertainty}
\]

\[
H(T \mid S) = \mathbb{E}_{(s,t) \sim p(S,T)}[-\log_2(p(t \mid s))] \quad \text{uncertainty about } T \text{ after knowing } S
\]

\[
\text{MI}(S; T) = H(T) - H(T \mid S)
\]

- **mutual information**
  - uncertainty about \( T \) **a priori**
  - uncertainty about \( T \) **after knowing** \( S \)
- **how much knowing** \( S \) **reduced uncertainty** about \( T \)

**Cross-Entropy:**

\[
H_q(T) = \mathbb{E}_{t \sim p(T)}[-\log_2(q(t))] \quad \text{how surprised is model } q \text{ in reality } p?
\]

\[
H_q(T \mid S) = \mathbb{E}_{(s,t) \sim p(S,T)}[-\log_2(q(t \mid s))]
\]

**XMI**

\[
\text{XMI}(S \rightarrow T) := H_{q_{LM}}(T) - H_{q_{MT}}(T \mid S)
\]
**Cross-Mutual Information** measures models’ performance on the act of translation

**Entropy:**

\[
H(T) = \mathbb{E}_{t \sim p(T)}[-\log_2(p(t))] \quad \text{uncertainty}
\]

\[
H(T \mid S) = \mathbb{E}_{(s,t) \sim p(S,T)}[-\log_2(p(t \mid s))] \quad \text{uncertainty about } T \text{ given } S
\]

**Mutual Information:**

\[
\text{MI}(S; T) = H(T) - H(T \mid S)
\]

- mutual information between S and T
- uncertainty about T a priori
- uncertainty about T after knowing S
- how much knowing S reduced uncertainty about T

**Cross-Entropy:**

\[
H_q(T) = \mathbb{E}_{t \sim p(T)}[-\log_2(q(t))] \quad \text{how surprised is model } q \text{ in reality } p?
\]

\[
H_q(T \mid S) = \mathbb{E}_{(s,t) \sim p(S,T)}[-\log_2(q(t \mid s))]
\]

**XMI(S \rightarrow T):**

\[
\text{XMI}(S \rightarrow T) := H_{q_{LM}}(T) - H_{q_{MT}}(T \mid S)
\]
Experiments
Experiments

Setup

• **Data**: Fully 21-parallel subset of Europarl

• **Models**:
  • 20 [○ → en] Transformers
  • 20 [en → ○] Transformers
Experiments

Setup

- **Data**: Fully 21-parallel subset of Europarl
- **Models**:
  - 20 [● → en] Transformers
  - 20 [en → ●] Transformers

Results

- For fixed target, BLEU and XMI correlate well! ✓
Experiments

Setup

- **Data**: Fully 21-parallel subset of Europarl
- **Models**:
  - 20 \([\circ \rightarrow \text{en}]\) Transformers
  - 20 \([\text{en} \rightarrow \circ]\) Transformers

Results

- For fixed target, BLEU and XMI correlate well! ✓
- Check our paper for more correlations
It’s Easier to Translate *out of* English than *into* it!
It’s Easier to Translate *out of* English than *into* it!

en-fi is easier than fi-en!
It’s Easier to Translate \textit{out of} English than \textit{into} it!

\textbf{en- – is easier than -en!}
Correlations with XMI?

The usual: type-token ratio...
but on the source side!

<table>
<thead>
<tr>
<th>Spearman’s $\rho$</th>
<th>Metric</th>
<th>$\emptyset \rightarrow \text{en}$</th>
<th>$\text{en} \rightarrow \emptyset$</th>
<th>both</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MCC$_{\text{src}}$</td>
<td>nope</td>
<td>nope</td>
<td>maybe?</td>
</tr>
<tr>
<td></td>
<td>MCC$_{\text{tgt}}$</td>
<td>nope</td>
<td>nope</td>
<td>maybe?</td>
</tr>
<tr>
<td></td>
<td>ADL$_{\text{src}}$</td>
<td>nope</td>
<td>nope</td>
<td>nope</td>
</tr>
<tr>
<td></td>
<td>ADL$_{\text{tgt}}$</td>
<td>nope</td>
<td>nope</td>
<td>maybe?</td>
</tr>
<tr>
<td></td>
<td>HPE-mean$_{\text{src}}$</td>
<td>nope</td>
<td>nope</td>
<td>maybe?</td>
</tr>
<tr>
<td></td>
<td>HPE-mean$_{\text{tgt}}$</td>
<td>nope</td>
<td>nope</td>
<td>maybe?</td>
</tr>
<tr>
<td>Mielke et al. (2019)</td>
<td>genetic</td>
<td>nope</td>
<td>nope</td>
<td>nope</td>
</tr>
<tr>
<td></td>
<td>syntactic</td>
<td>nope</td>
<td>nope</td>
<td>nope</td>
</tr>
<tr>
<td></td>
<td>featural</td>
<td>nope</td>
<td>nope</td>
<td>nope</td>
</tr>
<tr>
<td></td>
<td>phonological</td>
<td>nope</td>
<td>nope</td>
<td>nope</td>
</tr>
<tr>
<td></td>
<td>inventory</td>
<td>nope</td>
<td>nope</td>
<td>nope</td>
</tr>
<tr>
<td></td>
<td>geographic</td>
<td>nope</td>
<td>nope</td>
<td>nope</td>
</tr>
<tr>
<td>Lin et al. (2019)</td>
<td>word number ratio</td>
<td>maybe?</td>
<td>nope</td>
<td>maybe?</td>
</tr>
<tr>
<td></td>
<td>TTR$_{\text{src}}$</td>
<td>maybe?</td>
<td>–</td>
<td>-0.51</td>
</tr>
<tr>
<td></td>
<td>TTR$_{\text{tgt}}$</td>
<td>–</td>
<td>nope</td>
<td>maybe?</td>
</tr>
<tr>
<td></td>
<td>$d_{\text{TTR}}$</td>
<td>maybe?</td>
<td>nope</td>
<td>-0.47</td>
</tr>
<tr>
<td>Lin et al. (2019)</td>
<td>word overlap ratio</td>
<td>nope</td>
<td>nope</td>
<td>nope</td>
</tr>
</tbody>
</table>
Where to go from here?
Where to go from here?

- Cross-mutual information (XMI)
Where to go from here?

• Cross-mutual information (XMI)
  • A metric for translation difficulties between *any* two directions
Where to go from here?

• Cross-mutual information (XMI)
  • A metric for translation difficulties between *any* two directions

• No linguistic correlations, but TTR... again
Where to go from here?

• Cross-mutual information (XMI)
  • A metric for translation difficulties between any two directions

• No linguistic correlations, but TTR... again
  • Let’s scale this up and evaluate more pairs!
Where to go from here?

• Cross-mutual information (XMI)
  • A metric for translation difficulties between *any* two directions

• No linguistic correlations, but TTR... again
  • Let’s scale this up and evaluate more pairs!
  • Let’s build better models!
Where to go from here?

• Cross-mutual information (XMI)
  • A metric for translation difficulties between any two directions

• No linguistic correlations, but TTR... again
  • Let’s scale this up and evaluate more pairs!
  • Let’s build better models!

Code available online at https://github.com/e-bug/nmt-difficulty