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
### Evaluation Matrix

Translation quality of best system for test set `newstest2019` using metric `BLEU-cased`.

| Output Language |
|-----------------|
| Czech 19.3 |
| German 42.8 |
| English 27.4 |
| Finnish 33.0 |
| French 35.0 |
| Gujarati 24.9 |
| Kazakh 30.5 |
| Lithuanian 36.3 |
| Russian 40.2 |
| Chinese 39.9 |

### Scores

- **Czech** to **German**: 20.1
- **English** to **French**: 28.2
- **French** to **Kazakh**: 11.1
- **Kazakh** to **Russian**: 20.1
- **Russian** to **Chinese**: 36.3
- **Italian** to **Greek**: 44.6
| Language  | Output Language | Czech | German | English | Finnish | Gujarati | Kazakh | Lithuanian | Russian | Chinese |
|----------|-----------------|-------|--------|---------|---------|----------|--------|------------|---------|---------|
| Input Language | 19.3       | 20.1  | 29.9   | 33.0    | 35.0    | 24.9    | 30.5   | 36.3       | 40.2    | 39.9    |

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”  English
    “Programlayacağım”   Turkish

    → 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:**

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

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

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

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

\[
\frac{H(T)}{}
\]

\text{uncertainty about } T \text{ a priori}
Mutual Information expresses the act of translation

**Entropy:**

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

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

- **\(H(T)\)**: uncertainty about \(T\) *a priori*
- **\(H(T \mid S)\)**: uncertainty about \(T\) *after knowing S*
Mutual Information expresses the act of translation

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

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

$H(T) - H(T | S)$

- uncertainty about $T$ 
  - a priori

- how much knowing $S$ reduced uncertainty about $T$
**Mutual Information** expresses the *act of translation*

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

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

- **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

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

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

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

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

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

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

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

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

**mutual information**

uncertainty about $T$

**a priori**

uncertainty about $T$

after knowing $S$

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

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

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

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

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

- mutual information between $S$ and $T$
- uncertainty about $T$ before knowing $S$ (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” | symmetric! assuming all entropies w.r.t. same joint $p(S, T)$ | symmetric! assuming all entropies w.r.t. same joint $p(S, T)$ |

| $H(谢谢)$ | uncertainty about “谢谢” | symmetric! assuming all entropies w.r.t. same joint $p(S, T)$ | symmetric! assuming all entropies w.r.t. same joint $p(S, T)$ |
Mutual Information expresses the act of translation

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

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
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
  - 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:**

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

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

**Mutual Information (MI):**

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

This represents the mutual information between \( S \) and \( T \), which measures how much knowing \( S \) reduced the 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:**

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

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

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

- **Mutual Information (MI)** between S and T:
  - **a priori uncertainty** about T
  - **a posteriori uncertainty** about T
  - how much knowing S reduced uncertainty about T

**Cross-Entropy:**

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

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

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

**Entropy:**

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

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

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

\[ \text{mutual information} \]

\[ \text{uncertainty about } T \]  

\[ \text{a priori} \]

\[ \text{uncertainty about } T \]  

\[ \text{after knowing } S \]

\[ \text{how much knowing } S \text{ reduced uncertainty about } T \]

**Cross-Entropy:**

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

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

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

**Entropy:** \( H(T) = \mathbb{E}_{t \sim p(T)}[-\log_2(p(t))] \) 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 \( T \)
- \( H(T) \): uncertainty about \( T \) \textit{a priori}
- \( H(T \mid S) \): uncertainty about \( T \) \textit{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 \( p \)?
\( H_q(T \mid S) = \mathbb{E}_{(s,t) \sim p(S,T)}[-\log_2(q(t \mid s))] \)

\[
\text{XMI}(S \rightarrow T) := H_{q_{LM}}(T) - H_{q_{MT}}(T \mid S)
\]
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 [$\odot \rightarrow \text{en}$] Transformers
  - 20 [\text{en} \rightarrow \odot] Transformers

Results

- For fixed target, BLEU and XMI correlate well! ✔️
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! ✓
• 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 *out of* English than *into* it!

en- ∘ is easier than ∘ -en!
Correlations with XMI?

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

| Spearman’s $\rho$ | Metric     | $\text{∅} \rightarrow \text{en}$ | en $\rightarrow \text{∅}$ | both     |
|------------------|------------|-----------------|-----------------|---------|
| Mielke et al. (2019) | MCC_{src} | nope            | nope            | maybe?  |
|             | MCC_{tgt} | nope            | nope            | maybe?  |
|             | ADL_{src} | nope            | nope            | nope    |
|             | ADL_{tgt} | nope            | nope            | maybe?  |
|             | HPE-mean_{src} | nope         | nope            | maybe?  |
|             | HPE-mean_{tgt} | nope         | nope            | maybe?  |
| Lin et al. (2019) | genetic    | nope            | nope            | nope    |
|             | syntactic  | nope            | nope            | nope    |
|             | featural   | nope            | nope            | nope    |
|             | phonological | nope          | nope            | nope    |
|             | inventory  | nope            | nope            | nope    |
|             | geographic | nope            | nope            | nope    |
| Lin et al. (2019) | word number ratio | maybe?        | nope            | maybe?  |
|             | TTR_{src}  | maybe?          | –               | -0.51   |
|             | TTR_{tgt}  | –               | nope            | maybe?  |
|             | $d_{TTR}$  | maybe?          | nope            | -0.47   |
|             | word overlap ratio | nope        | nope            | nope    |
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](https://github.com/e-bug/nmt-difficulty)