# 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

UNIVERSITY OF COPENHAGEN

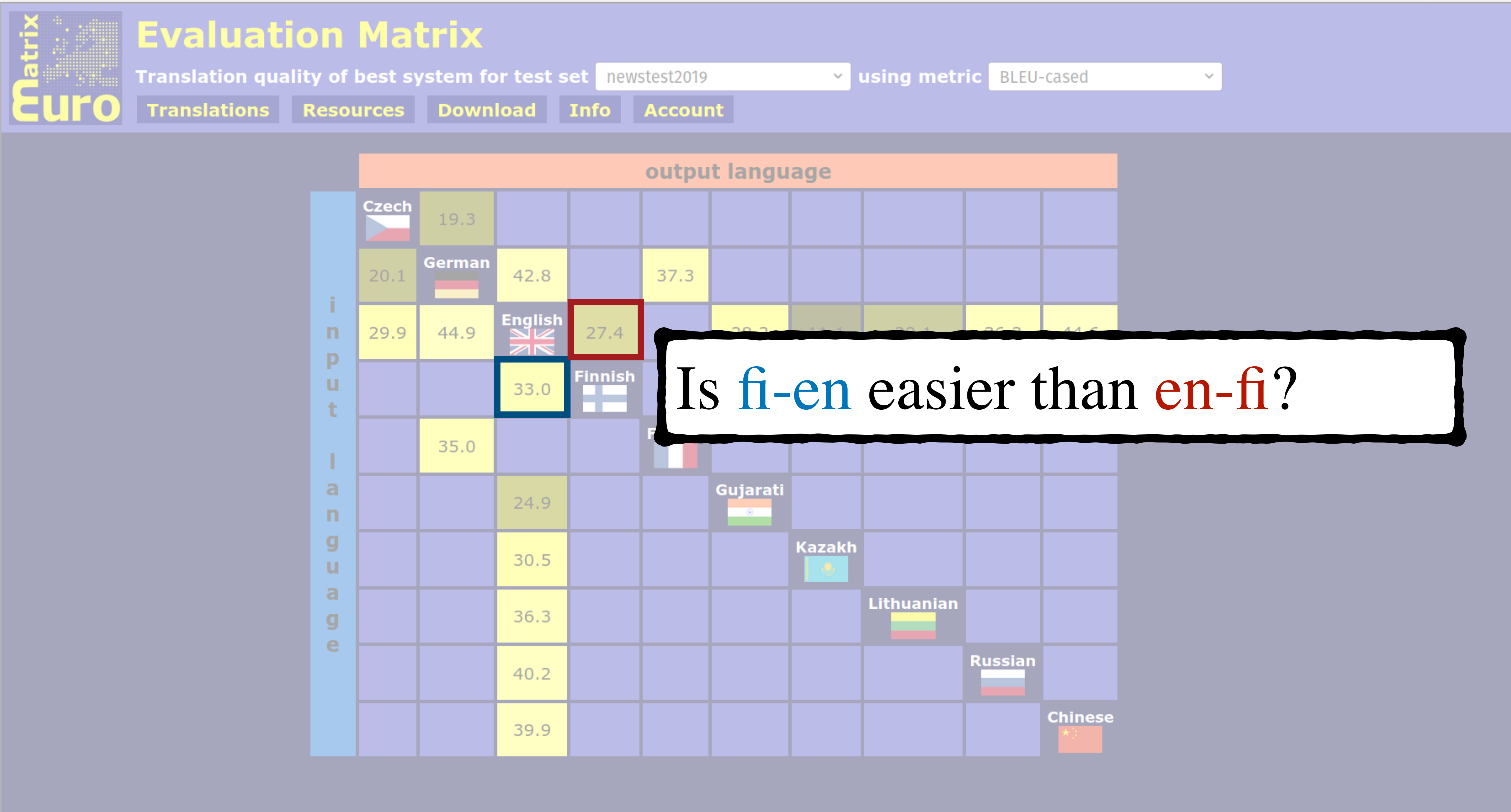
# Evaluation Matrix

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

Translations | Resources | Download | Info | Account

output language

| input language \ output language | Czech | German | English | Finnish | French | Gujarati | Kazakh | Lithuanian | Russian | Chinese |
|---|---|---|---|---|---|---|---|---|---|---|
| Czech | | 19.3 | | | | | | | | |
| German | 20.1 | | 42.8 | | 37.3 | | | | | |
| English | 29.9 | 44.9 | | 27.4 | | 28.2 | 11.1 | 20.1 | 36.3 | 44.6 |
| Finnish | | | 33.0 | | | | | | | |
| French | | 35.0 | | | | | | | | |
| Gujarati | | | 24.9 | | | | | | | |
| Kazakh | | | 30.5 | | | | | | | |
| Lithuanian | | | 36.3 | | | | | | | |
| Russian | | | 40.2 | | | | | | | |
| Chinese | | | 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:** $\mathrm{H}(T) = \mathbb{E}_{t \sim p(T)}[-\log_2(p(t))]$ uncertainty

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

# *Mutual Information* expresses the *act of translation*

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

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

$$\underbrace{\mathrm{H}(T)}_{\text{uncertainty about } T}$$

*a priori*

# *Mutual Information* expresses the *act of translation*

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

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

$$\underbrace{\mathrm{H}(T)}_{\text{uncertainty about } T \text{ a priori}} \qquad \underbrace{\mathrm{H}(T \mid S)}_{\text{uncertainty about } T \text{ after knowing } S}$$

# *Mutual Information* expresses the *act of translation*

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

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

$$\underbrace{\underbrace{\mathrm{H}(T)}_{\substack{\text{uncertainty about } T \\ \textit{a priori}}} \quad - \quad \underbrace{\mathrm{H}(T \mid S)}_{\substack{\text{uncertainty about } T \\ \textit{after knowing } S}}}_{\text{how much knowing } S \text{ reduced uncertainty about } T}$$

# *Mutual Information* expresses the *act of translation*

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

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

$$\underbrace{\mathrm{MI}(S;T)}_{\substack{\textbf{mutual information}\\ \text{between } S \text{ and } T}} = \underbrace{\underbrace{\mathrm{H}(T)}_{\substack{\text{uncertainty about } T\\ \textit{a priori}}} - \underbrace{\mathrm{H}(T \mid S)}_{\substack{\text{uncertainty about } T\\ \textit{after knowing } S}}}_{\text{how much knowing } S \text{ reduced uncertainty about } T}$$

# *Mutual Information* expresses the *act of translation*

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

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

$$\underbrace{\mathrm{MI}(S;T)}_{\substack{\textbf{mutual information} \\ \text{between } S \text{ and } T}} = \underbrace{\underbrace{\mathrm{H}(T)}_{\substack{\text{uncertainty about } T \\ \textit{a priori}}} - \underbrace{\mathrm{H}(T \mid S)}_{\substack{\text{uncertainty about } T \\ \textit{after knowing } S}}}_{\text{how much knowing } S \text{ reduced uncertainty about } T}$$

symmetric!
assuming all entropies w.r.t. same joint $p(S,T)$

# *Mutual Information* expresses the *act of translation*

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

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

$$\underbrace{\underbrace{\mathrm{MI}(S;T)}_{\textbf{mutual information}\text{ between } S \text{ and } T} = \underbrace{\mathrm{H}(T)}_{\text{uncertainty about } T \text{ a priori}} - \underbrace{\mathrm{H}(T \mid S)}_{\text{uncertainty about } T \text{ after knowing } S}}$$

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

symmetric!
assuming all entropies w.r.t. same joint $p(S,T)$

**Example:** en-zh

# *Mutual Information* expresses the *act of translation*

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

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

$$\underbrace{\mathrm{MI}(S;T)}_{\substack{\textbf{mutual information}\\ \text{between } S \text{ and } T}} = \underbrace{\underbrace{\mathrm{H}(T)}_{\substack{\text{uncertainty about } T\\ \textit{a priori}}} - \underbrace{\mathrm{H}(T \mid S)}_{\substack{\text{uncertainty about } T\\ \textit{after knowing } S}}}_{\text{how much knowing } S \text{ reduced uncertainty about } T}$$

**symmetric!**
assuming all entropies w.r.t. same joint $p(S,T)$

**Example:** en-zh

| H(谢谢) |
|---|

uncertainty about “谢谢”

# *Mutual Information* expresses the *act of translation*

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

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

$$\underbrace{\underbrace{\mathrm{MI}(S;T)}_{\substack{\textbf{mutual information}\\ \text{between } S \text{ and } T}} = \underbrace{\mathrm{H}(T)}_{\substack{\text{uncertainty about } T\\ \textit{a priori}}} - \underbrace{\mathrm{H}(T \mid S)}_{\substack{\text{uncertainty about } T\\ \textit{after knowing } S}}}_{}$$

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

symmetric!
assuming all entropies w.r.t. same joint $p(S,T)$

**Example:** en-zh

| | |
|---|---|
| H(谢谢) | uncertainty about “谢谢” |
| H(谢谢 I Thanks) | uncertainty about “谢谢” after knowing its translation |

# *Mutual Information* expresses the *act of translation*

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

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

$$\underbrace{\mathrm{MI}(S;T)}_{\substack{\textbf{mutual information}\\ \text{between } S \text{ and } T}} = \underbrace{\underbrace{\mathrm{H}(T)}_{\substack{\text{uncertainty about } T\\ \textit{a priori}}} - \underbrace{\mathrm{H}(T \mid S)}_{\substack{\text{uncertainty about } T\\ \textit{after knowing } S}}}_{\text{how much knowing } S \text{ reduced uncertainty about } T}$$

symmetric!
assuming all entropies w.r.t. same joint $p(S,T)$

**Example:** en-zh

| | |
|---|---|
| H(谢谢) | uncertainty about “谢谢” |
| H(谢谢 I Thanks) | uncertainty about “谢谢” after knowing its translation |
| MI(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:** $\mathrm{H}(T) = \mathbb{E}_{t \sim p(T)}[-\log_2(p(t))]$ uncertainty

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

$$\underbrace{\mathrm{MI}(S;T)}_{\substack{\textbf{mutual information} \\ \text{between } S \text{ and } T}} = \underbrace{\underbrace{\mathrm{H}(T)}_{\substack{\text{uncertainty about } T \\ \textit{a priori}}} - \underbrace{\mathrm{H}(T \mid S)}_{\substack{\text{uncertainty about } T \\ \textit{after knowing } S}}}_{\text{how much knowing } S \text{ reduced uncertainty about } T}$$

# *Cross-Mutual Information* measures models' performance on the *act of translation*

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

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

$$\underbrace{\mathrm{MI}(S;T)}_{\substack{\textbf{mutual information}\\ \text{between } S \text{ and } T}} = \underbrace{\underbrace{\mathrm{H}(T)}_{\substack{\text{uncertainty about } T\\ \textit{a priori}}} - \underbrace{\mathrm{H}(T \mid S)}_{\substack{\text{uncertainty about } T\\ \textit{after knowing } S}}}_{\text{how much knowing } S \text{ reduced uncertainty about } T}$$

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

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

# *Cross-Mutual Information* measures models' performance on the *act of translation*

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

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

$$\underbrace{\mathrm{MI}(S; T)}_{\substack{\textbf{mutual information} \\ \text{between } S \text{ and } T}} = \underbrace{\underbrace{\mathrm{H}(T)}_{\substack{\text{uncertainty about } T \\ \textit{a priori}}} - \underbrace{\mathrm{H}(T \mid S)}_{\substack{\text{uncertainty about } T \\ \textit{after knowing } S}}}_{\text{how much knowing } S \text{ reduced uncertainty about } T}$$

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

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

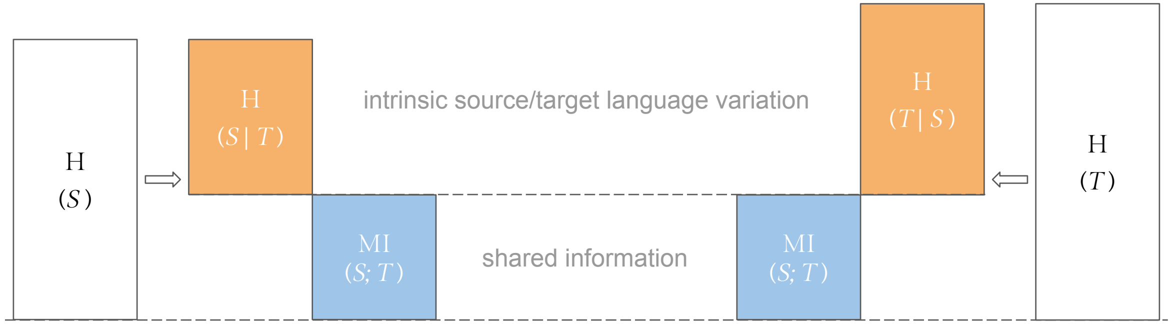
$$\mathrm{XMI}(S \to T) := H_{q_{LM}}(T) - H_{q_{MT}}(T \mid S)$$

# *Cross-Mutual Information* measures models' performance on the *act of translation*

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

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

$$\underbrace{\underbrace{\mathrm{MI}(S;T)}_{\substack{\textbf{mutual information}\\ \text{between } S \text{ and } T}} = \underbrace{\mathrm{H}(T)}_{\substack{\text{uncertainty about } T\\ \textit{a priori}}} - \underbrace{\mathrm{H}(T \mid S)}_{\substack{\text{uncertainty about } T\\ \textit{after knowing } S}}}_{\text{how much knowing } S \text{ reduced uncertainty about } T}$$




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

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

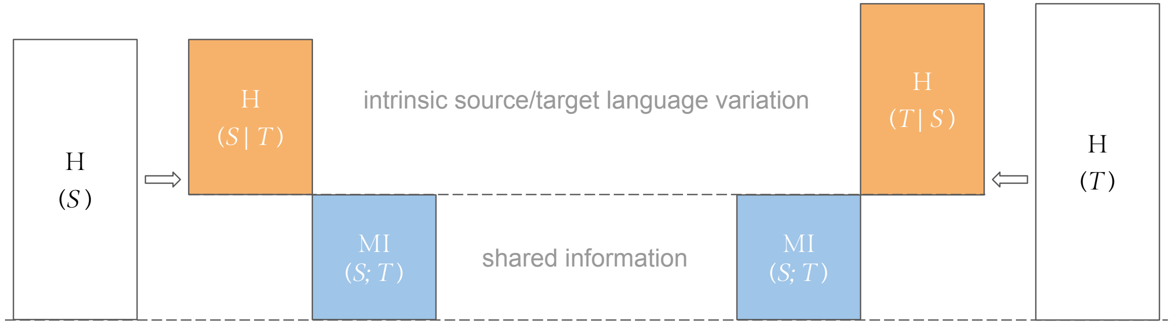
$$\mathrm{XMI}(S \to T) := H_{q_{LM}}(T) - H_{q_{MT}}(T \mid S)$$

# *Cross-Mutual Information* measures models' performance on the *act of translation*

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

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

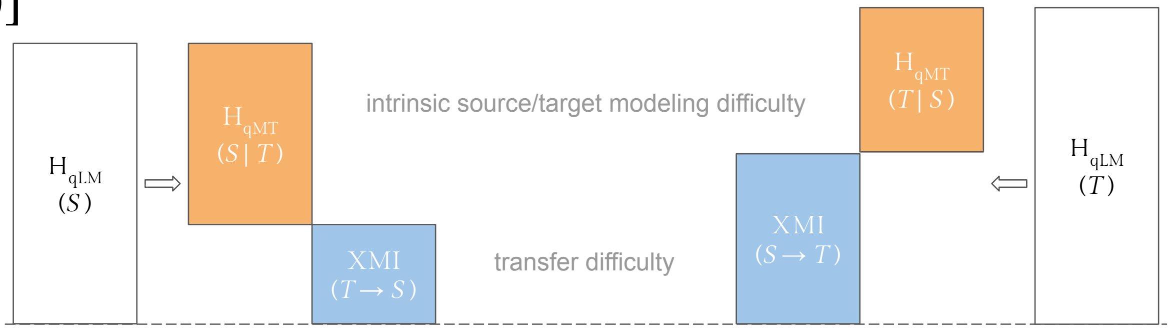
$$\underbrace{\mathrm{MI}(S;T)}_{\substack{\textbf{mutual information}\\ \text{between } S \text{ and } T}} = \underbrace{\underbrace{\mathrm{H}(T)}_{\substack{\text{uncertainty about } T\\ \textit{a priori}}} - \underbrace{\mathrm{H}(T \mid S)}_{\substack{\text{uncertainty about } T\\ \textit{after knowing } S}}}_{\text{how much knowing } S \text{ reduced uncertainty about } T}$$




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

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

$$\mathrm{XMI}(S \rightarrow T) := H_{q_{LM}}(T) - H_{q_{MT}}(T \mid S)$$



# Experiments

# Experiments

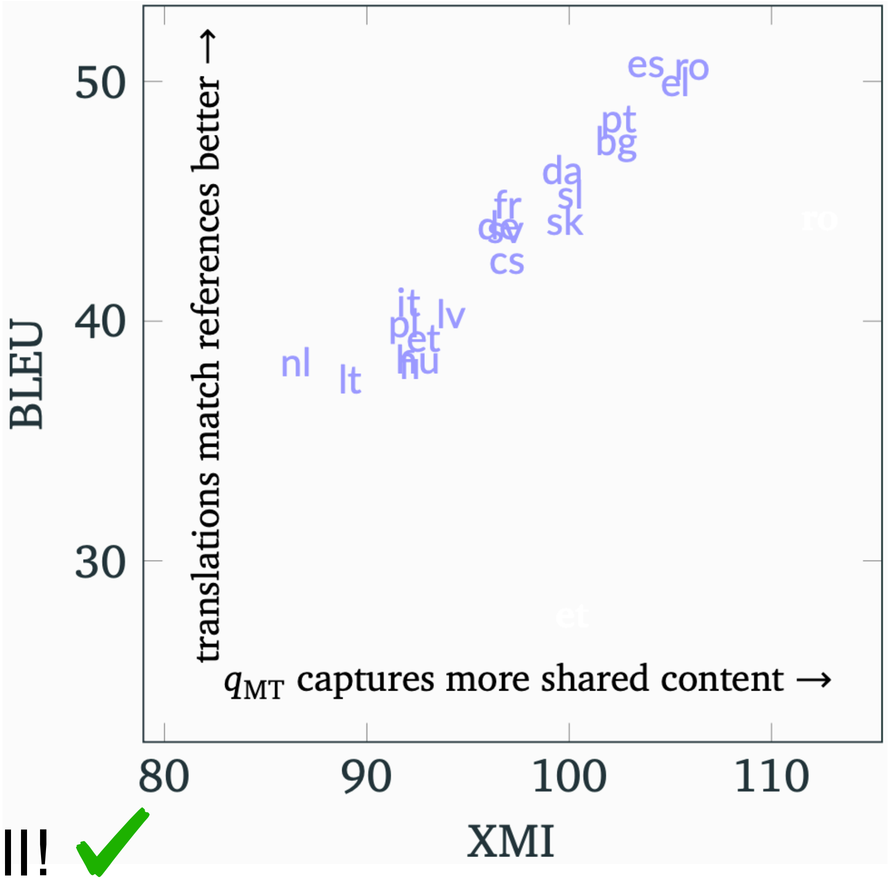
## Setup

- *Data*: Fully 21-parallel subset of Europarl
- *Models*:
  - 20 [ $\circ \rightarrow$ en] Transformers
  - 20 [en $\rightarrow \circ$ ] Transformers

# Experiments

## Setup

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



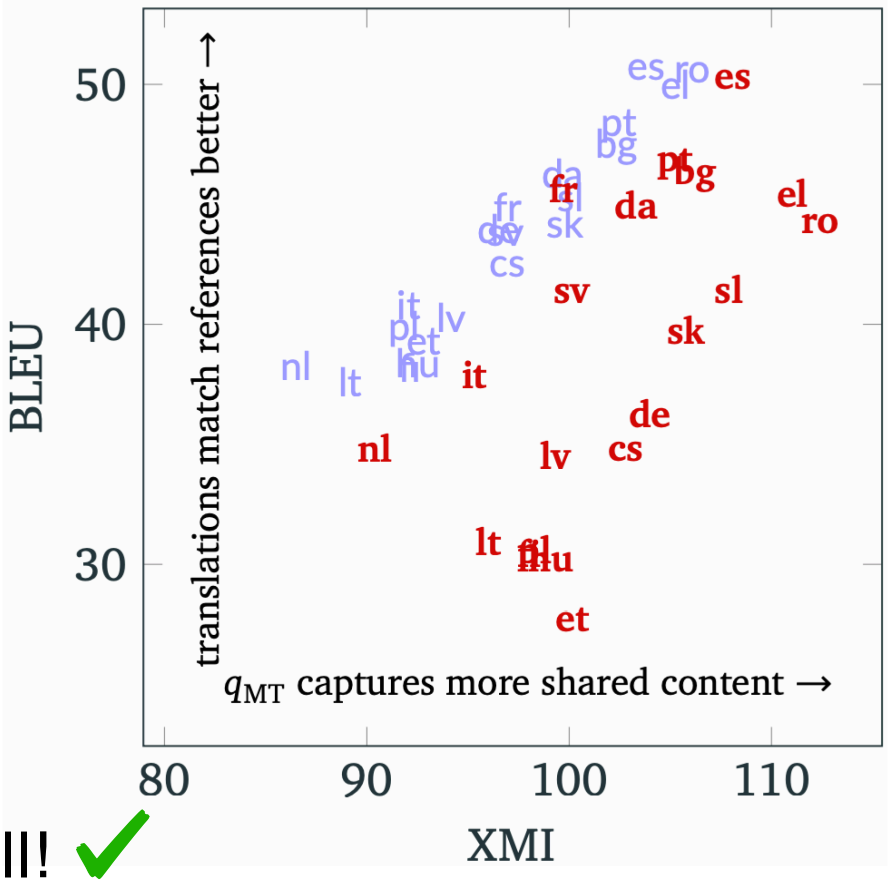

## Results

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

# Experiments

## Setup

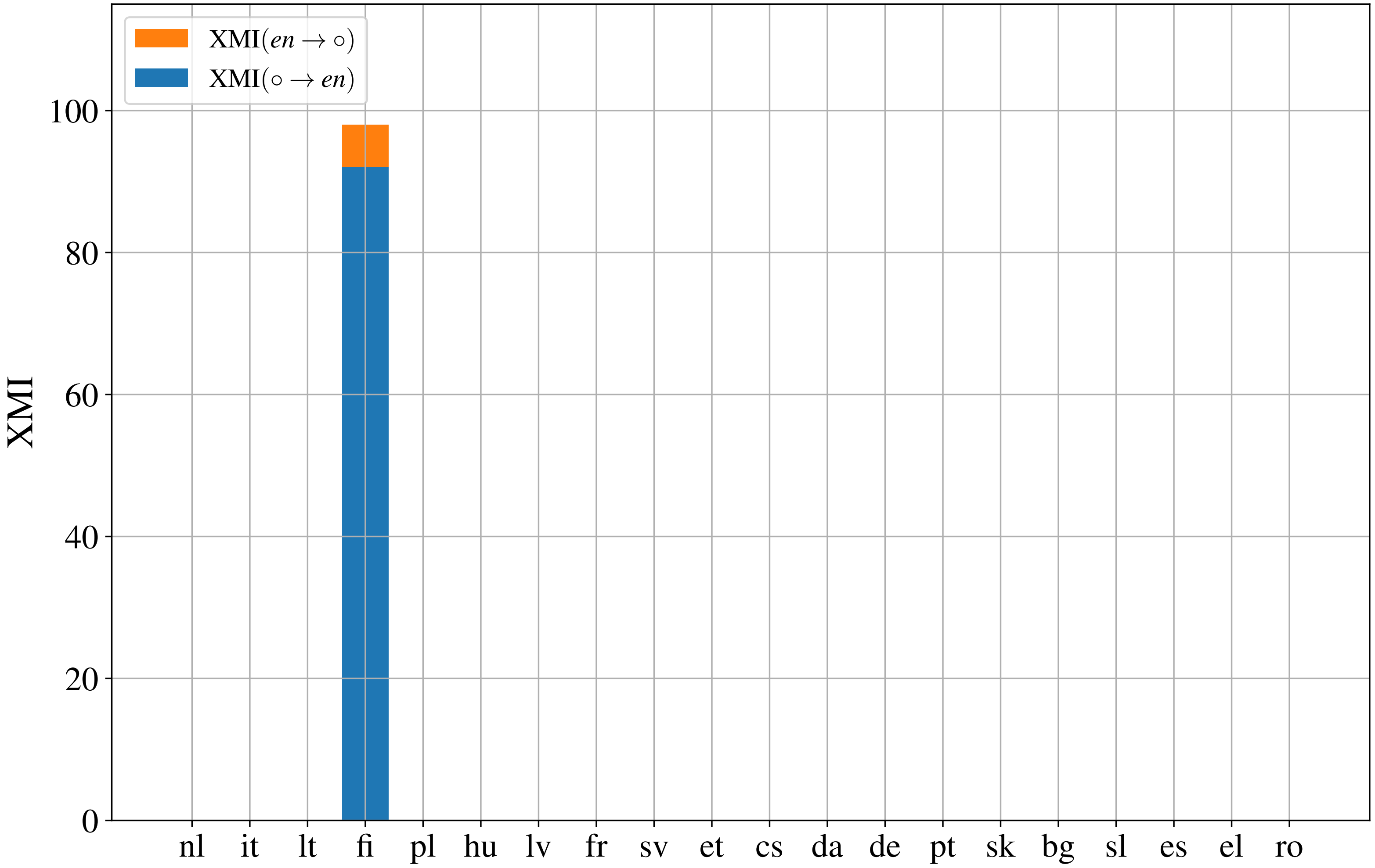
- *Data*: Fully 21-parallel subset of Europarl
- *Models*:
  - 20 [ $\circ \rightarrow$ en] Transformers
  - 20 [en $\rightarrow \circ$ ] Transformers




## Results

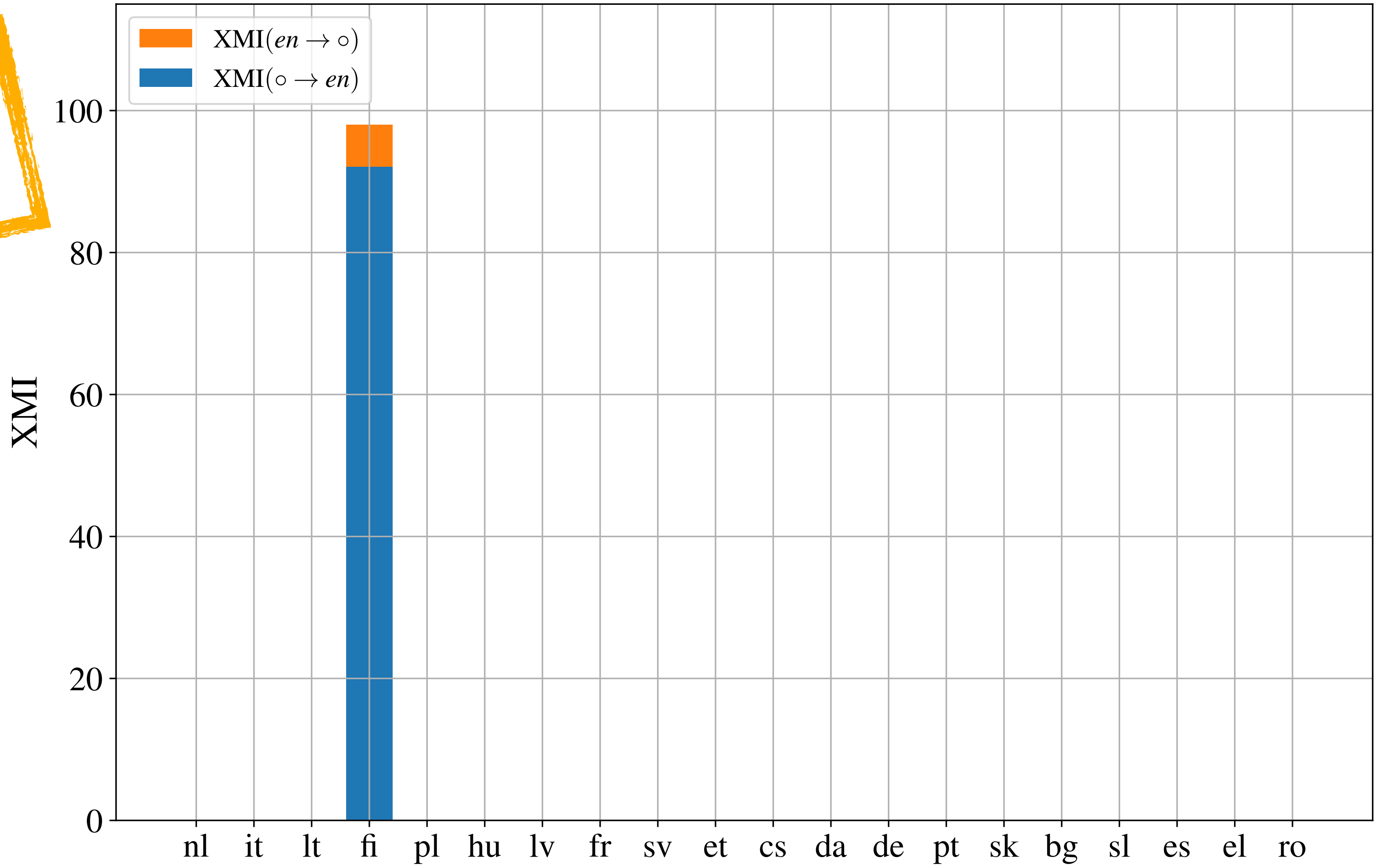
- For fixed target, BLEU and XMI correlate well! ✓
- Check our paper for more correlations

# It’s Easier to Translate *out of* English than *into* it!



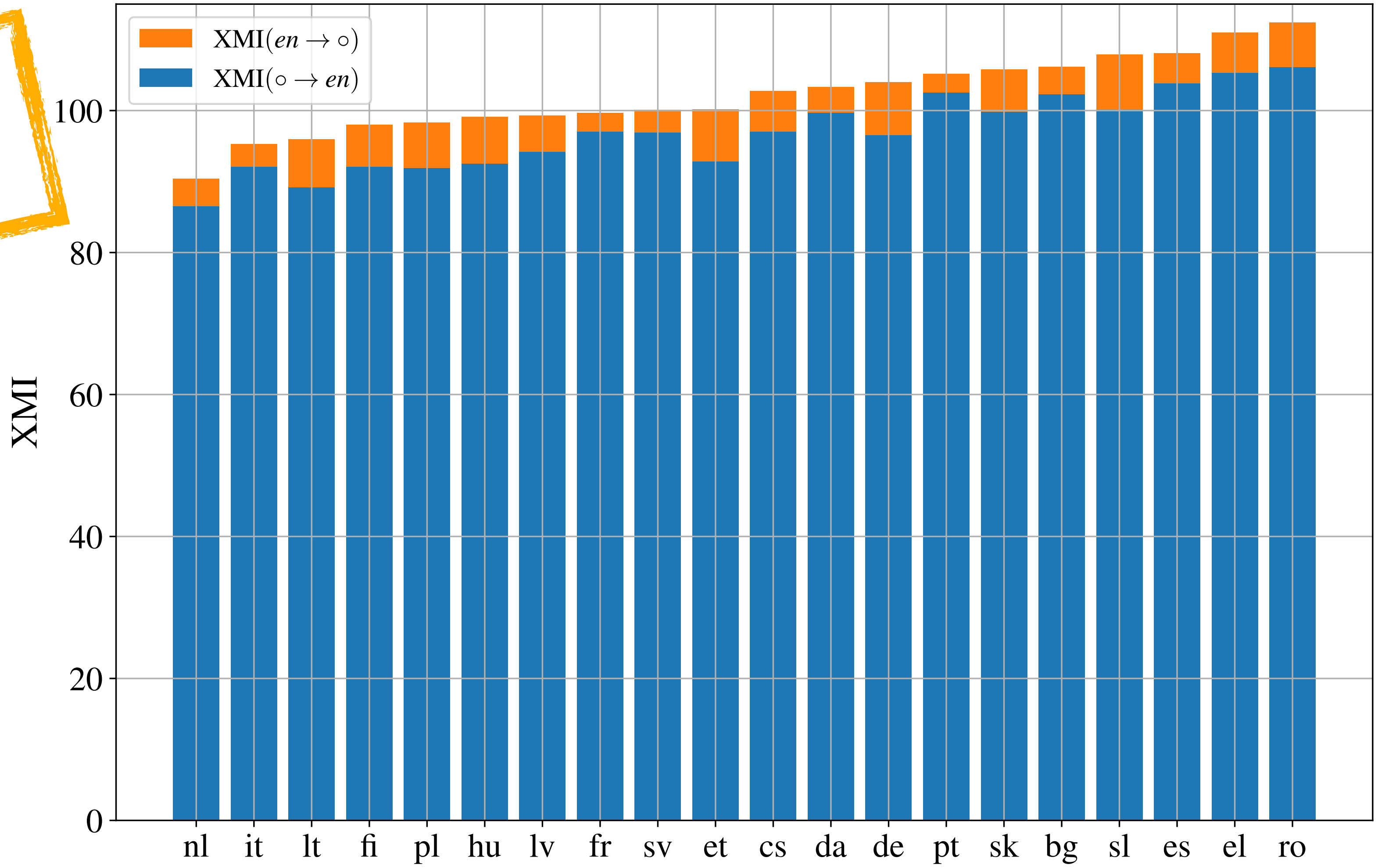
# It’s Easier to Translate *out of* English than *into* it!

en-fi is easier than fi-en!



# It’s Easier to Translate *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 | ◌ → en | en → ◌ | both |
|---|---|---|---|---|
| Mielke et al. (2019) | $\mathrm{MCC}_{\mathrm{src}}$ | nope | nope | maybe? |
| | $\mathrm{MCC}_{\mathrm{tgt}}$ | nope | nope | maybe? |
| | $\mathrm{ADL}_{\mathrm{src}}$ | nope | nope | nope |
| | $\mathrm{ADL}_{\mathrm{tgt}}$ | nope | nope | maybe? |
| | $\text{HPE-mean}_{\mathrm{src}}$ | nope | nope | maybe? |
| | $\text{HPE-mean}_{\mathrm{tgt}}$ | 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? |
| | $\mathrm{TTR}_{\mathrm{src}}$ | maybe? | – | **-0.51** |
| | $\mathrm{TTR}_{\mathrm{tgt}}$ | – | nope | maybe? |
| | $d_{\mathrm{TTR}}$ | 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?

Thanks!

- 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