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

Emanuele Bugliarello\(^c\) Sabrina J. Mielke\(^d\) Antonios Anastasopoulos\(^a\)
Ryan Cotterell\(^{b,5}\) Naoaki Okazaki\(^g\)

\(^c\)University of Copenhagen \(^d\)Johns Hopkins University \(^a\)Carnegie Mellon University
\(^b\)University of Cambridge \(^g\)ETH Zürich \(^g\)Tokyo Institute of Technology
emaneule@di.ku.dk, sjmielke@jhu.edu, aanastas@cs.cmu.edu, rcotterell@inf.ethz.ch, okazaki@c.titech.ac.jp

Abstract

The performance of neural machine translation systems is commonly evaluated in terms of BLEU. However, due to its reliance on target language properties and generation, the BLEU metric does not allow an assessment of which translation directions are more difficult to model. In this paper, we propose cross-mutual information (XMI): an asymmetric information-theoretic metric of machine translation difficulty that exploits the probabilistic nature of most neural machine translation models. XMI allows us to better evaluate the difficulty of translating text into the target language while controlling for the difficulty of the target-side generation component independent of the translation task. We then present the first systematic and controlled study of cross-lingual translation difficulties using modern neural translation systems. Code for replicating our experiments is available online at \url{https://github.com/e-bug/nmt-difficulty}.

1 Introduction

Machine translation (MT) is one of the core research areas in natural language processing. Current state-of-the-art MT systems are based on neural networks (Sutskever et al., 2014; Bahdanau et al., 2015), which generally surpass phrase-based systems (Koehn, 2009) in a variety of domains and languages (Bentivogli et al., 2016; Toral and Sánchez-Cartagena, 2017; Castilho et al., 2017; Bojar et al., 2018; Barrault et al., 2019). Using phrase-based MT systems, various controlled studies to understand where the translation difficulties lie for different language pairs were conducted (Birch et al., 2008; Koehn et al., 2009). However, comparable studies have yet to be performed for neural machine translation (NMT). As a result, it is still unclear whether all translation directions are equally easy (or hard) to model for NMT. This paper hence aims at filling this gap: \emph{Ceteris paribus, is it easier to translate from English into Finnish or into Hungarian? And how much easier is it? Conversely, is it equally hard to translate Finnish and Hungarian into another language?}

Based on BLEU (Papineni et al., 2002) scores, previous work (Belinkov et al., 2017) suggests that translating into morphologically rich languages, such as Hungarian or Finnish, is harder than translating into morphologically poor ones, such as English. However, a major obstacle in the cross-lingual comparison of MT systems is that many automatic evaluation metrics, including BLEU and METEOR (Banerjee and Lavie, 2005), are not cross-lingually comparable. In fact, being a function of \(n\)-gram overlap between candidate and reference translations, they only allow for a fair comparison of the performance between models when translating into the same test set in the same target language. Indeed, one cannot and should not draw conclusions about the difficulty of translating a source language into different target languages purely based on BLEU (or METEOR) scores.
In response, we propose cross-mutual information (XMI), a new metric towards cross-linguistic comparability in NMT. In contrast to BLEU, this information-theoretic quantity no longer explicitly depends on language, model, and tokenization choices. It does, however, require that the models under consideration are probabilistic. As an initial starting point, we perform a case study with a controlled experiment on 21 European languages. Our analysis showcases XMI’s potential for shedding light on the difficulties of translation as an effect of the properties of the source or target language. We also perform a correlation analysis in an attempt to further explain our findings. Here, in contrast to the general wisdom, we find no significant evidence that translating into a morphologically impoverished language is harder than translating into a morphologically impoverished one. In fact, the only significant correlate of MT difficulty we find is source-side type–token ratio.

2 Cross-Linguistic Comparability through Likelihoods, not BLEU

Human evaluation will always be the gold standard of MT evaluation. However, it is both time-consuming and expensive to perform. To help researchers and practitioners quickly deploy and evaluate new systems, automatic metrics that correlate fairly well with human evaluations have been proposed over the years (Banerjee and Lavie, 2005; Snover et al., 2006; Isozaki et al., 2010; Lo, 2019). BLEU (Papineni et al., 2002), however, has remained the most common metric to report the performance of MT systems. BLEU is a precision-based metric: a BLEU score is proportional to the geometric average of the number of n-grams in the candidate translation that also appear in the reference translation for 1 ≤ n ≤ 4.1

In the context of our study, we take issue with two shortcomings of BLEU scores that prevent a cross-linguistically comparable study. First, it is not possible to directly compare BLEU scores across languages because different languages might express the same meaning with a very different number of words. For instance, agglutinative languages like Turkish often use a single word to express what other languages have periphrastic constructions for. To be concrete, the expression “I will have been programming” is five words in English, but could easily have been one word in a language with sufficient morphological markings; this unfairly boosts BLEU scores when translating into English. The problem is further exacerbated by tokenization techniques as finer granularities result in more partial credit and higher n for the n-gram matches (Post, 2018). In summary, BLEU only allows us to compare models for a fixed target-language and tokenization scheme, i.e. it only allows us to draw conclusions about the difficulty of translating different source languages into a specific target one (with downstream performance as a proxy for difficulty). Thus, BLEU scores cannot provide an answer to which translation direction is easier between any two source–target pairs.

In this work, we address this particular shortcoming by considering an information-theoretic evaluation. Formally, let \( V_S \) and \( V_T \) be source- and target-language vocabularies, respectively. Let \( S \) and \( T \) be source- and target-sentence-valued random variables for languages \( S \) and \( T \), respectively; then \( S \) and \( T \) respectively range over \( V_S^2 \) and \( V_T^2 \). These random variables \( S \) and \( T \) are distributed according to some true, unknown probability distribution \( p \). The cross-entropy between the true distribution \( p \) and a probabilistic neural translation model \( q_{MT}(t \mid s) \) is defined as:

\[
H_{q_{MT}}(T \mid S) = -\sum_{t \in V_T} \sum_{s \in V_S} p(t, s) \log q_{MT}(t \mid s)
\]

Since we do not know \( p \), we cannot compute eq. (1). However, given a held-out data set of sentence pairs \( \{(s^{(i)}, t^{(i)})\}_{i=1}^N \) assumed to be drawn from \( p \), we can approximate the true cross-entropy as follows:

\[
H_{q_{MT}}(T \mid S) \approx -\frac{1}{N} \sum_{i=1}^N \log q_{MT}(t^{(i)} \mid s^{(i)})
\]

In the limit as \( N \rightarrow \infty \), eq. (2) converges to eq. (1).

We emphasize that this evaluation does not rely on language tokenization provided that the model \( q_{MT} \) does not (Mielke, 2019). While common in the evaluation of language models, cross-entropy evaluation has been eschewed in machine translation research since (i) not all MT models are probabilistic and (ii) we are often interested in measuring the quality of the candidate translation our model actually produces, e.g. under approximate decoding. However, an information-theoretic evaluation
We contend that simply reporting cross-entropies is not enough. A second issue in performing a controlled, cross-lingual MT comparison is that the language generation component (without translation) is not equally difficult across languages (Cotterell et al., 2018). We claim that the difficulty of translation corresponds more closely to the mutual information \( \text{MI}(S; T) \) between the source and target language, which tells us how much easier it becomes to predict \( T \) when \( S \) is given (see Figure 1). But what is the appropriate analogue of mutual information for cross-entropy? One such natural generalization is a novel quantity that we term \text{cross-mutual information}, defined as:

\[
\text{XMI}(S \rightarrow T) = H_{\text{qLM}}(T) - H_{\text{qMT}}(T \mid S)
\]

(3)

where \( H_{\text{qLM}}(T) \) denotes the cross-entropy of the target sentence \( T \) under the model \( \text{qLM} \). As in §2, this can, analogously, be approximated by the cross-entropy of a separate target-side language model \( \text{qMT} \) over our held-out data set:

\[
\text{XMI}(S \rightarrow T) \approx -\frac{1}{N} \sum_{i=1}^{N} \log_2 \frac{q_{\text{LM}}(t^{(i)})}{q_{\text{MT}}(t^{(i)} \mid s^{(i)})}
\]

(4)

which, again, becomes exact as \( N \rightarrow \infty \). In practice, we note that we mix different distributions \( q_{\text{LM}}(t) \) and \( q_{\text{MT}}(t \mid s) \) and, thus, \( q_{\text{LM}}(t) \) is not necessarily representable as a marginal: there need not be any distribution \( \tilde{q}(s) \) such that \( q_{\text{LM}}(t) = \sum_{s \in V_s^t} q_{\text{MT}}(t \mid s) \cdot \tilde{q}(s) \). While \( q_{\text{MT}} \) and \( q_{\text{LM}} \) can, in general, be any two models, we exploit the characteristics of NMT models to provide a more meaningful, model-specific estimate of XMI. NMT architectures typically consist of two components: an encoder that embeds the input text sequence, and a decoder that generates translated output text. The latter acts as a conditional language model, where the source-language sentence embedded by the encoder drives the target-language generation. Hence, we use the decoder of \( q_{\text{MT}} \) as our \( q_{\text{LM}} \) to accurately estimate the difficulty of translation for a given architecture in a controlled way.

In summary, by looking at XMI, we can effectively decouple the language generation component, whose difficulties have been investigated by Cotterell et al. 2018 and Mielke et al. 2019, from the translation component. This gives us a measure of how rich and useful the information extracted from the source language is for the target-language generation component.

4 Experiments

In order to measure which pairs of languages are harder to translate to and from, we make use of the latest release v7 of Europarl (Koehn, 2005): a corpus of the proceedings of the European Parliament containing parallel sentences between English (en) and 20 other European languages: Bulgarian (bg), Czech (cs), Danish (da), German (de), Greek (el), Spanish (es), Estonian (et), Finnish (fi), French (fr), Hungarian (hu), Italian (it), Lithuanian (lt), Latvian (lv), Dutch (nl), Polish (pl), Portuguese (pt), Romanian (ro), Slovak (sk), Slovene (sl) and Swedish (sv).

Pre-processing steps In order to precisely effect a fully controlled experiment, we enforce a fair comparison by selecting the set of parallel sentences available across all 21 languages in Europarl. This fully controls for the semantic content of the sentences; however, we cannot adequately control for translationese (Stymne, 2017; Zhang and Toral, 2019). Our subset of Europarl contains 190,733 sentences for training, 1,000 unique, random sentences for validation and 2,000 unique, random sentences for testing. For each parallel corpus, we jointly learn byte-pair encodings (BPE; Sennrich et al., 2016) for the source and target languages, using 16,000 merge operations. We use the same vocabularies for the language models.\(^2\)

Setup In our experiments, we train Transformer models (Vaswani et al., 2017), which often achieve state-of-the-art performance on MT for various language pairs. In particular, we rely on the PyTorch (Paszke et al., 2019) re-implementation of the Transformer model in the fairseq toolkit (Ott et al., 2019). For language modeling, we use the decoder from the same architecture, training it at the sentence level, as opposed to commonly used fixed-length chunks. We train our systems using label smoothing (LS; Szegedy et al., 2016; Meister et al.,

\(^2\)For English, we arbitrarily chose the English portion of the en-bg vocabulary.
2020) as it has been shown to prevent models from over-confident predictions, which helps to regularize the models. We report cross-entropies ($H_{\text{qMT}}$, $H_{\text{qLM}}$), XMI, and BLEU scores obtained using SACREBLEU (Post, 2018). Finally, in a similar vein to Cotterell et al. (2018), we multiply cross-entropy values by the number of sub-word units generated by each model to make our quantities independent of sentence lengths (and divide them by the total number of sentences to match our approximations of the true distributions). See App. A for experimental details.

5 Results and Analysis

We train 40 systems, translating each language into and from English. The models’ performance in terms of BLEU scores, and the cross-mutual information (XMI) and cross-entropy values over the test sets are reported in Table 1 with significant values marked in App. B.

Translating into English When translating into the same target language (in this case, English), BLEU scores are, in fact, comparable, and can be used as a proxy for difficulty. We can then conclude, for instance, that Lithuanian ($\text{lt}$) is the hardest language to translate from, while Spanish ($\text{es}$) is the easiest. In this scenario, given the good correlation of BLEU scores with human evaluations, it is desirable that XMI correlates well with BLEU. This behavior is indeed apparent in the blue points in the left part of Figure 2, confirming the efficacy of XMI in evaluating the difficulty of translation while still being independent of the target language generation component.

Translating from English Despite the large gaps between BLEU scores in Table 1, one should not be tempted to claim that it is easier to translate into English than from English for these languages as often hinted at in previous work (e.g., Belinkov et al., 2017). As we described above, different target languages are not directly comparable, and we actually find that XMI is slightly higher, on average, when translating from English, indicating that it is actually easier, on average, to transfer information correctly in this direction. For instance, translation from English to Finnish is shown to be easier than from Finnish to English, despite the large gap

| $\text{en} \rightarrow$ | bg | cs | da | de | el | es | et | fi | fr | hu | it | lv | nl | pl | pt | ro | sk | sl | sv | avg |
| BLEU                    | 47.4 | 42.4 | 46.3 | 44.0 | 50.0 | 50.6 | 39.3 | 38.2 | 44.9 | 38.4 | 40.8 | 37.6 | 40.3 | 38.3 | 39.8 | 48.3 | 50.5 | 44.2 | 45.3 | 43.7 | 43.5 |
| XMI ($\text{en} \rightarrow \text{en}$) | 17.2 | 79.0 | 11.9 | 99.7 | 96.5 | 103.8 | 92.8 | 92.1 | 97.0 | 92.5 | 92.1 | 98.9 | 94.2 | 94.5 | 90.1 | 102.5 | 106.1 | 99.8 | 100.1 | 96.9 | 96.9 |

Table 1: Test scores, from and into English, Europarl, visualized in Figure 2 and Figure 3.

Figure 2: Some correlations between metrics in Table 1, into and from English. More correlations in Figure 4.
in BLEU scores. This suggests that the former model is heavily penalized by the target-side language model; this is likely because Finnish has a large number of inflections for nouns and verbs. Another interesting example is given by Greek (en) and Spanish (es) in Table 1, where, again, the two tasks achieve very different BLEU scores but similar XMI. In light of the correlation with BLEU when translating into English, this shows us that Greek is just harder to language-model, corroborating the findings of Mielke et al. (2019). Moreover, Figure 2 clearly shows that, as expected, XMI is not as well correlated with BLEU when translating from English, given that BLEU scores are not cross-linguistically comparable.

**Correlations with linguistic and data features**

Last, we conduct a correlation study between the translation difficulties as measured by XMI and the linguistic and data-dependent properties of each translation task, following the approaches of Lin et al. (2019) and Mielke et al. (2019). Table 2 lists Pearson’s and Spearman’s correlation coefficients for data-dependent metrics, where bold values indicate statistically significant results ($p < 0.05$) after Bonferroni correction ($p < 0.0029$). Interestingly, the only features that significantly correlate with our metric are related to the type-to-token ratio (TTR) for the source language and the distance between source and target TTRs. This implies that a potential explanation for the differences in translation difficulty lies in lexical variation. For full correlation results, refer to App. D.

### 6 Conclusion

In this work, we propose a novel information-theoretic approach, XMI, to measure the translation difficulty of probabilistic MT models. Differently from BLEU and other metrics, ours is language- and tokenization-agnostic, enabling the first systematic and controlled study of cross-lingual translation difficulties. Our results show that XMI correlates well with BLEU scores when translating into the same language (where they are comparable), and that higher BLEU scores in different languages do not necessarily imply easier translations. In future work, we plan to extend this analysis across more translation pairs, more diverse languages and multiple domains, as well as investigating the effect of translationese or source-side grammatical errors (Anastasopoulos, 2019).

### Acknowledgments

The authors are thankful to the anonymous reviewers for their valuable feedback. The second-to-last author acknowledges a Facebook Fellowship and discussions with Tiago Pimentel. This project has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 801199, the National Science Foundation under grant 1761548, and by “Research and Development of Deep Learning Technology for Advanced Multilingual Speech Translation,” the Commissioned Research of National Institute of Information and Communications Technology (NICT), Japan.
References


Sabrina J. Mielke. 2019. Can you compare perplexity across different segmentations?


A Experimental Details

Pre-processing steps To precisely determine the effect of the different properties of each language in translation difficulty, we enforce a fair comparison by selecting the same set of parallel sentences across all the languages evaluated in our data set. The number of parallel sentences available in Europarl varies considerably, ranging from 387K sentences for Polish-English to 2.3M sentences for Dutch-English. Therefore, we proceed by taking the set of English sentences that are shared by all the language pairs. This leaves us with 197,919 sentences for each language pair, from which we then extract 1,000 and 2,000 unique, random sentences for validation and test, respectively.

We follow the same pre-processing steps used by Vaswani et al. (2017) to train the Transformer model on WMT data: Data sets are first tokenized using the Moses toolkit (Koehn et al., 2007) and then filtered by removing sentences longer than 80 tokens in either source or target language. Due to this cleaning step that is specific to each training corpus, different sentences are dropped in each data set. We then only select the set of sentence pairs that are shared across all languages. This results in a final number of 190,733 training sentences. For each parallel corpus, we jointly learn byte-pair encodings (BPE; Sennrich et al., 2016) for source and target languages, using 16,000 merge operations.

Training setup In our experiments, we train a Transformer model (Vaswani et al., 2017), which achieves state-of-the-art performance on a multitude of language pairs. In particular, we rely on the PyTorch re-implementation of the Transformer model in the Fairseq toolkit (Ott et al., 2019). All experiments are based on the Base Transformer architecture, which we trained for 20,000 steps and evaluated using the checkpoint corresponding to the lowest validation loss. We trained our models on a cluster of 4 machines, each equipped with 4 Nvidia P100 GPUs, resulting in training times of almost 70 minutes for each system. Sentence pairs with similar sequence length were batched together, with each batch containing a total of approximately $32K$ source tokens and $32K$ target tokens.

We used the hyper-parameters specified in latest version (3) of Google’s Tensor2Tensor (Vaswani et al., 2018) implementation, with the exception of the dropout rate, as we found 0.3 to be more robust across all the models trained on Europarl.

<table>
<thead>
<tr>
<th>Model</th>
<th>Train bootstrap</th>
<th>Test bootstrap</th>
</tr>
</thead>
<tbody>
<tr>
<td>en-es</td>
<td>47.6 (0.233)</td>
<td>50.2 (0.026)</td>
</tr>
<tr>
<td>en-et</td>
<td>25.6 (0.167)</td>
<td>27.7 (0.026)</td>
</tr>
<tr>
<td>lt-en</td>
<td>34.5 (0.150)</td>
<td>37.6 (0.027)</td>
</tr>
<tr>
<td>ro-en</td>
<td>47.5 (0.232)</td>
<td>50.5 (0.027)</td>
</tr>
</tbody>
</table>

Table 3: Mean test BLEU scores when bootstrapping train and test sets. Numbers in brackets denote standard deviation over 5 runs (train bootstrap) and 95% confidence interval over 1,000 samples (test bootstrap).

Models are optimized using Adam (Kingma and Ba, 2015) and following the learning schedule specified by Vaswani et al. (2017) with 8,000 warm-up steps. We employed label smoothing $\epsilon_{ls} = 0.1$ (Szegedy et al., 2016) during training and we used beam search with a beam size of 4 and length penalty $\alpha = 0.6$ (Wu et al., 2016).

For language models, we use a Transformer decoder with the same hyperparameters used in the translation task to effectively measure the contribution given by a translation. These models were trained, using label smoothing $\epsilon_{ls} = 0.1$, for 10,000 steps on sequences consisting of separate sentences in our corpus. Analogously to translation models, the checkpoints corresponding to the lowest validation losses were used for evaluation.

B Statistical Significance Tests

Table 3 presents the results when applying bootstrap re-sampling (Koehn, 2004) on either training or test sets to the systems achieving the highest and the lowest BLEU scores in the validation set for each direction. In our experiments, we observe a general trend where the performance of different models varies similarly. For instance, when we bootstrap test sets, we see that the average BLEU scores are equal to the ones seen in Table 1, and that all the models have similar confidence intervals. When bootstrapping the training data, we observe a consistent drop in mean performance of 2 – 3 BLEU points across the translation tasks. The drop in performance is not surprising as the resulting training sets are more redundant, having fewer unique sentences than the original sets, but it is interesting to see that all models are similarly affected. The standard deviation over 5 runs is also similar across all models but slightly larger on the high-performing ones.

1 The same results were observed in all of the 40 models.
Figure 4: More correlations between metrics in Table 1, into and from English.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Pearson</th>
<th>Spearman</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCC&lt;sub&gt;src&lt;/sub&gt;</td>
<td>-0.2579 (0.2723)</td>
<td>-0.4302 (0.0056)</td>
</tr>
<tr>
<td>MCC&lt;sub&gt;tgt&lt;/sub&gt;</td>
<td>-0.1260 (0.5965)</td>
<td>-0.2135 (0.3660)</td>
</tr>
<tr>
<td>ADL&lt;sub&gt;src&lt;/sub&gt;</td>
<td>0.2619 (0.1025)</td>
<td>-0.1263 (0.5957)</td>
</tr>
<tr>
<td>ADL&lt;sub&gt;tgt&lt;/sub&gt;</td>
<td>-0.1166 (0.4737)</td>
<td>-0.2887 (0.2170)</td>
</tr>
<tr>
<td>HPE-mean&lt;sub&gt;src&lt;/sub&gt;</td>
<td>0.0112 (0.9188)</td>
<td>0.0166 (0.9188)</td>
</tr>
<tr>
<td>HPE-mean&lt;sub&gt;tgt&lt;/sub&gt;</td>
<td>0.3570 (0.0237)</td>
<td>0.1450 (0.035)</td>
</tr>
</tbody>
</table>

Table 4: All Pearson’s and Spearman’s correlation coefficients and corresponding p-values (in brackets) between XMI and various metrics. Values in black are statistically significant at \( p < 0.05 \), and bold values are also statistically significant after Bonferroni correction (\( p < 0.0029 \)).

C More Correlations between Metrics

Figure 4 shows more correlations between the metrics we reported in our experiments (see Table 1).

D Correlation Analysis

Table 4 shows Pearson’s and Spearman’s correlations between XMI and all investigated predictors, including per-direction results. Following Lin et al. (2019) and Mielke et al. (2019), we evaluated:

- MCC: Morphological counting complexity (Sagot, 2013), using the values for Europarl reported by Cotterell et al. (2018).
- ADL: Average dependency length (Futrell et al., 2015), using the values reported for Europarl by Mielke et al. (2019).
- HPE-mean: mean over all Europarl tokens of Head-POS Entropy (Dehouck and Denis, 2018), as reported by Mielke et al. (2019).
- Six different linguistic distances (genetic, syntactic, featural, phonological, inventory, geographic) from the URIEL Typological Database (Littell et al., 2017). We refer the reader to Lin et al. (2019) for more details.

- Word number ratio: number of source tokens over number of target tokens used for training.
- TTR<sub>src</sub> and TTR<sub>tgt</sub>: type-to-token ratio evaluated on the source and target language training data, respectively, to measure lexical diversity.
- \( d_{TTR} \): distance between the TTRs of the source and target language corpora, as a rough indication of their morphological similarity:
  \[
  d_{TTR} = \left( 1 - \frac{TTR_{src}}{TTR_{tgt}} \right)^2.
  \]
- Word overlap ratio: we measure the similarity between the vocabularies of source and target languages as the ratio between the number of shared types and the size of their union.