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**[fi] **Antonios Anastasopoulos**[ə]
Ryan Cotterell[ð,ʕ] **Naoaki Okazaki**[ŋ]
[c]University of Copenhagen [fi]Johns Hopkins University [ə]Carnegie Mellon University
[ð]University of Cambridge [ʕ]ETH Zürich [ŋ]Tokyo Institute of Technology
emanuele@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 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: *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?

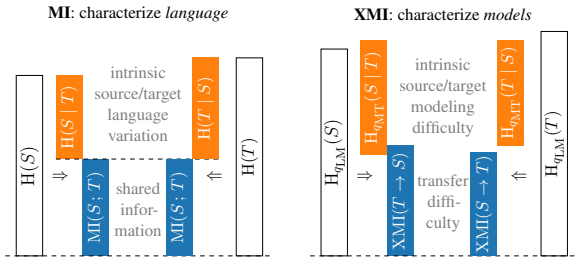

Figure 1: **Left:** Decomposing the uncertainty of a sentence as mutual information plus language-inherent uncertainty: mutual information (MI) corresponds to just how much easier it becomes to predict T when you are given S. MI is symmetric but the relation between $\mathrm{H}(S)$ and $\mathrm{H}(T)$ can be arbitrary. **Right:** estimating cross-entropies using models q_{MT} and q_{LM} invalidates relations between bars, except that $\mathrm{H}_{q_\cdot}(\cdot) \geq \mathrm{H}(\cdot)$. XMI, our proposed metric, is no longer purely a symmetric measure of language, but now an asymmetric measure that mostly highlights models' shortcomings.

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.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1640–1649
July 5 - 10, 2020.

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 rich 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 $1 \leq n \leq 4$.[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 $\mathcal{V}_{\mathsf{S}}$ and $\mathcal{V}_{\mathsf{T}}$ 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 $\mathcal{V}_{\mathsf{S}}^*$ and $\mathcal{V}_{\mathsf{T}}^*$. 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 $q_{\mathrm{MT}}(\mathbf{t} \mid \mathbf{s})$ is defined as:

$$\mathrm{H}_{q_{\mathrm{MT}}}(T \mid S) = - \sum_{\mathbf{t} \in \mathcal{V}_{\mathsf{T}}^*} \sum_{\mathbf{s} \in \mathcal{V}_{\mathsf{S}}^*} p(\mathbf{t}, \mathbf{s}) \log_2 q_{\mathrm{MT}}(\mathbf{t} \mid \mathbf{s}) \tag{1}$$

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

$$\mathrm{H}_{q_{\mathrm{MT}}}(T \mid S) \approx - \frac{1}{N} \sum_{i=1}^{N} \log_2 q_{\mathrm{MT}}(\mathbf{t}^{(i)} \mid \mathbf{s}^{(i)}) \tag{2}$$

In the limit as $N \to \infty$, 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

[1] BLEU also corrects for reference coverage and includes a length penalty, but we focus on the high-level picture.

is much more suitable for measuring the more abstract notion of which language pairs are hardest to translate to and from, which is our purpose here.

3 Disentangling Translation Difficulty and Monolingual Complexity

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** $\mathrm{MI}(S;T)$ 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 **cross-mutual information**, defined as:

$$\mathrm{XMI}(S \to T) = \mathrm{H}_{q_{\mathrm{LM}}}(T) - \mathrm{H}_{q_{\mathrm{MT}}}(T \mid S) \quad (3)$$

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

$$\mathrm{XMI}(S \to T) \approx \quad (4)$$
$$-\frac{1}{N}\sum_{i=1}^{N} \log_2 \frac{q_{\mathrm{LM}}(\mathbf{t}^{(i)})}{q_{\mathrm{MT}}(\mathbf{t}^{(i)} \mid \mathbf{s}^{(i)})}$$

which, again, becomes exact as $N \to \infty$. In practice, we note that we mix different distributions $q_{\mathrm{LM}}(\mathbf{t})$ and $q_{\mathrm{MT}}(\mathbf{t} \mid \mathbf{s})$ and, thus, $q_{\mathrm{LM}}(\mathbf{t})$ is *not* necessarily representable as a marginal: there need not be any distribution $\tilde{q}(\mathbf{s})$ such that $q_{\mathrm{LM}}(\mathbf{t}) = \sum_{\mathbf{s}\in\mathcal{V}_{\mathrm{S}}^*} q_{\mathrm{MT}}(\mathbf{t} \mid \mathbf{s}) \cdot \tilde{q}(\mathbf{s})$. While q_{MT} and q_{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_{MT} as our q_{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.

◌ → **en**	bg	cs	da	de	el	**es**	et	fi	fr	hu	it	**lt**	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(◌ → en)	102.3	97.0	99.7	96.5	105.3	103.8	92.8	92.1	97.0	92.5	92.1	89.2	94.2	86.5	91.9	102.5	106.1	99.8	100.1	96.9	96.9
$H_{q_{LM}}$(en)										154.2											154.2
$H_{q_{MT}}$(en | ◌)	51.8	57.2	54.5	57.7	48.9	50.4	61.4	62.0	57.2	61.6	62.1	65.0	60.0	67.7	62.3	51.7	48.1	54.4	54.1	57.3	57.3
en → ◌	bg	cs	da	de	el	es	et	fi	fr	hu	it	lt	lv	nl	pl	pt	ro	sk	sl	sv	**avg**
BLEU	46.3	34.7	45.0	36.3	45.5	50.2	27.7	30.5	45.7	30.3	37.9	31.0	34.6	34.9	30.5	46.7	44.2	39.8	41.5	41.3	38.73
XMI(en *to* ◌)	106.2	102.8	103.3	104.0	111.0	108.1	100.2	98.0	99.7	99.1	95.3	96.0	99.3	90.4	98.3	105.2	112.4	105.8	107.9	100.1	102.1
$H_{q_{LM}}$(◌)	156.5	164.0	152.7	167.6	163.7	159.3	162.5	158.6	154.9	166.6	158.6	159.2	156.4	159.7	163.4	159.3	160.5	157.7	158.2	153.1	159.6
$H_{q_{MT}}$(◌ | en)	50.3	61.2	49.4	63.6	52.7	51.3	62.4	60.6	55.1	67.5	63.3	63.1	57.0	69.3	65.1	54.1	48.1	51.9	50.3	53.0	57.5

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

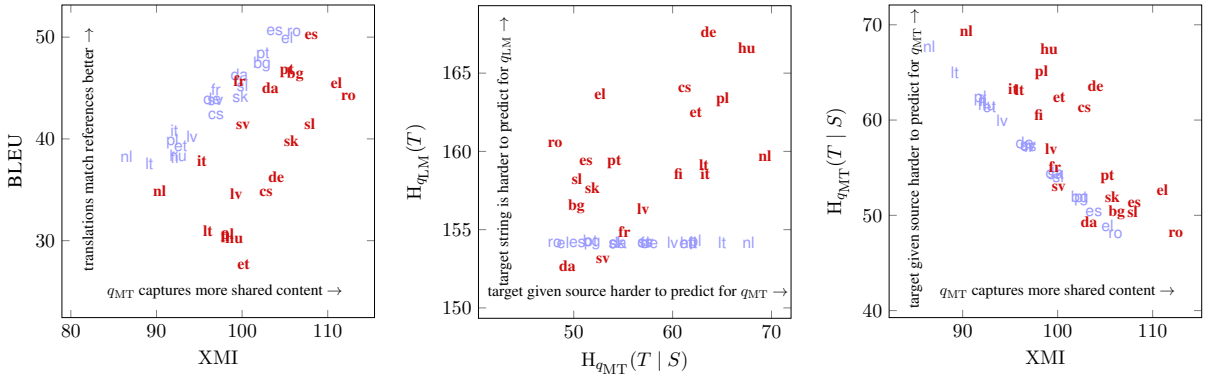

Figure 2: Some correlations between metrics in Table 1, into and **from** English. More correlations in Figure 4.

2020) as it has been shown to prevent models from over-confident predictions, which helps to regularize the models. We report cross-entropies ($H_{q_{MT}}$, $H_{q_{LM}}$), XMI, and BLEU scores obtained using SacreBLEU (Post, 2018).[3] 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.[4] 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 (lt) is the hardest language to translate from, while Spanish (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

[3] Signature: BLEU+c.mixed+#.1+s.exp+tok.13a+v.1.2.12.

[4] Due to resource limitations, we chose these tasks because most of the information available in the web is in English (https://w3techs.com/technologies/overview/content_language) and effectively translating it into any other language would reduce the digital language divide (http://labs.theguardian.com/digital-language-divide/). Besides, translating into English gives most people access to any local information.

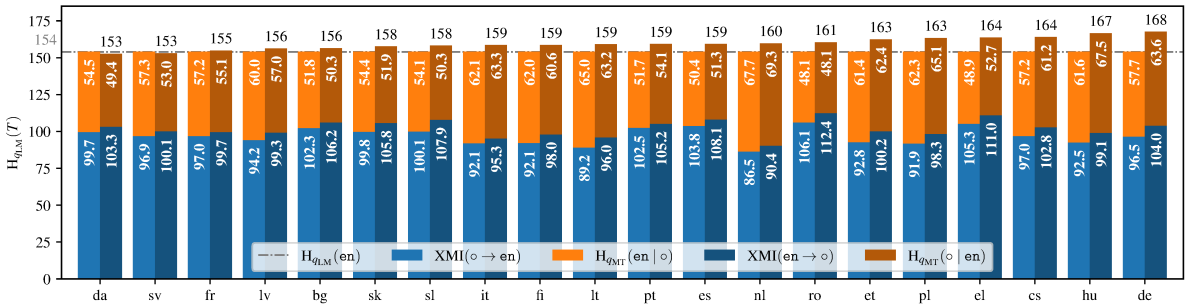

Figure 3: $\mathrm{H}_{q_{\mathrm{LM}}}(T)$, decomposed into $\mathrm{XMI}(S \rightarrow T)$, the information that the system successfully transfers, and $\mathrm{H}_{q_{\mathrm{MT}}}(T \mid S)$, the uncertainty that remains in the target language, all measured in bits. Note that in $\mathrm{XMI}(S \rightarrow T)$ the translation is from the left to the right argument.

Metric	Pearson	Spearman
word number ratio	0.2988 (0.0611)	0.3570 (0.0237)
$\mathrm{TTR}_{\mathrm{src}}$	**-0.5196 (0.0006)**	**-0.5136 (0.0007)**
$\mathrm{TTR}_{\mathrm{tgt}}$	0.1651 (0.3086)	0.3355 (0.0343)
d_{TTR}	-0.4427 (0.0042)	**-0.4660 (0.0024)**
word overlap ratio	0.1383 (0.3949)	0.1731 (0.2853)

Table 2: Correlation coefficients (and p-values) between XMI and data-related features.

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 (`el`) 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-lingually 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

Antonios Anastasopoulos. 2019. An analysis of source-side grammatical errors in NMT. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 213–223, Florence, Italy. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, Vancouver, Canada. Association for Computational Linguistics.

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus phrase-based machine translation quality: a case study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 257–267, Austin, Texas. Association for Computational Linguistics.

Alexandra Birch, Miles Osborne, and Philipp Koehn. 2008. Predicting success in machine translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 745–754, Honolulu, Hawaii. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Sheila Castilho, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley, and Andy Way. 2017. Is neural machine translation the new state of the art? *The Prague Bulletin of Mathematical Linguistics*, 108(1):109–120.

Ryan Cotterell, Sabrina J. Mielke, Jason Eisner, and Brian Roark. 2018. Are all languages equally hard to language-model? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 536–541, New Orleans, Louisiana. Association for Computational Linguistics.

Mathieu Dehouck and Pascal Denis. 2018. A framework for understanding the role of morphology in universal dependency parsing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2864–2870, Brussels, Belgium. Association for Computational Linguistics.

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 944–952, Cambridge, MA. Association for Computational Linguistics.

Diederick P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit*, volume 5, pages 79–96.

Philipp Koehn. 2009. *Statistical Machine Translation*. Cambridge University Press.

Philipp Koehn, Alexandra Birch, and Ralf Steinberger. 2009. 462 machine translation systems for Europe. In *Proceedings of the Twelfth Machine Translation Summit*, pages 65–72.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open

source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

Chi-kiu Lo. 2019. YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy. Association for Computational Linguistics.

Clara Meister, Elizabeth Salesky, and Ryan Cotterell. 2020. Generalized entropy regularization or: There's nothing special about label smoothing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Seattle, USA. Association for Computational Linguistics.

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

Sabrina J. Mielke, Ryan Cotterell, Kyle Gorman, Brian Roark, and Jason Eisner. 2019. What kind of language is hard to language-model? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4975–4989, Florence, Italy. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Benoît Sagot. 2013. Comparing complexity measures. In *Computational Approaches to Morphological Complexity*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*.

Sara Stymne. 2017. The effect of translationese on tuning for statistical machine translation. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 241–246, Gothenburg, Sweden. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826.

Antonio Toral and Víctor M. Sánchez-Cartagena. 2017. A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational*

Linguistics: Volume 1, Long Papers, pages 1063–1073, Valencia, Spain. Association for Computational Linguistics.

Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan N. Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. Tensor2Tensor for neural machine translation. *CoRR*, abs/1803.07416.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Mike Zhang and Antonio Toral. 2019. The effect of translationese in machine translation test sets. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 73–81, Florence, Italy. Association for Computational Linguistics.

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.

Model	Train bootstrap	Test bootstrap
en-es	47.6 (0.233)	50.2 (0.026)
en-et	25.6 (0.167)	27.7 (0.026)
lt-en	34.5 (0.150)	37.6 (0.027)
ro-en	47.5 (0.232)	50.5 (0.027)

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.[5] 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.

[5]The same results were observed in all of the 40 models.

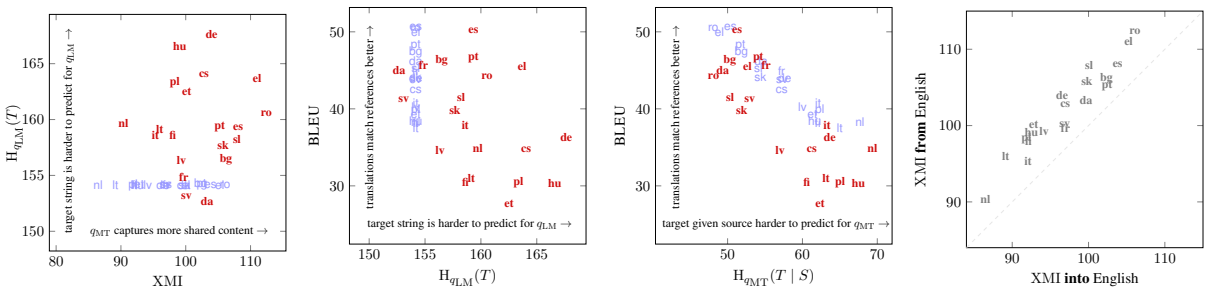

Figure 4: More correlations between metrics in Table 1, into and **from** English.

Metric	Pearson			Spearman		
	◌→ en	en → ◌	both	◌→ en	en → ◌	both
MCC_{src}	-0.2579 (0.2723)	–	-0.4302 (0.0056)	-0.2135 (0.3660)	–	-0.4444 (0.0041)
MCC_{tgt}	–	-0.1260 (0.5965)	0.2619 (0.1025)	–	-0.1263 (0.5957)	0.3778 (0.0162)
ADL_{src}	-0.2972 (0.2032)	–	-0.1166 (0.4737)	-0.2887 (0.2170)	–	0.0166 (0.9188)
ADL_{tgt}	–	-0.2254 (0.3393)	-0.2110 (0.1912)	–	-0.1820 (0.4426)	-0.3798 (0.0156)
HPE-mean$_{src}$	0.2012 (0.3950)	–	0.4567 (0.0031)	0.2000 (0.3979)	–	0.4508 (0.0035)
HPE-mean$_{tgt}$	–	0.0142 (0.9525)	-0.4115 (0.0083)	–	0.0120 (0.9599)	-0.4103 (0.0085)
genetic	0.0433 (0.8563)	0.0777 (0.7446)	0.0544 (0.7387)	-0.1526 (0.5207)	-0.1741 (0.4630)	-0.1360 (0.4028)
syntactic	-0.3643 (0.1143)	-0.2056 (0.3845)	-0.2556 (0.1114)	-0.3560 (0.1234)	-0.2695 (0.2506)	-0.2688 (0.0935)
featural	-0.0561 (0.8142)	-0.0577 (0.8090)	-0.0511 (0.7540)	0.0121 (0.9597)	-0.0093 (0.9690)	-0.0109 (0.9467)
phonological	-0.1442 (0.5441)	-0.2222 (0.3465)	-0.1647 (0.3097)	-0.0435 (0.8556)	-0.0948 (0.6909)	-0.0906 (0.5782)
inventory	0.1125 (0.6369)	0.1048 (0.6601)	0.0976 (0.5492)	0.1231 (0.6052)	0.1472 (0.5356)	0.1128 (0.4884)
geographic	0.1983 (0.4019)	0.3388 (0.1440)	0.2416 (0.1332)	0.1336 (0.5745)	0.2550 (0.2779)	0.2062 (0.2017)
word number ratio	0.4559 (0.0434)	-0.2953 (0.2063)	0.2988 (0.0611)	0.4602 (0.0412)	-0.3278 (0.1582)	0.3570 (0.0237)
TTR_{src}	-0.4746 (0.0345)	–	**-0.5196 (0.0006)**	-0.4857 (0.0299)	–	**-0.5136 (0.0007)**
TTR_{tgt}	–	-0.2931 (0.2099)	0.1651 (0.3086)	–	-0.3128 (0.1794)	0.3355 (0.0343)
d_{TTR}	-0.4434 (0.0502)	-0.2404 (0.3072)	-0.4427 (0.0042)	-0.4857 (0.0299)	-0.3128 (0.1794)	**-0.4660 (0.0024)**
word overlap ratio	0.2563 (0.2754)	0.0526 (0.8258)	0.1383 (0.3949)	0.1474 (0.5352)	0.1474 (0.5352)	0.1731 (0.2853)

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_{src} and TTR_{tgt}: 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:

$$d_{\mathrm{TTR}} = \left(1 - \frac{\mathrm{TTR}_{src}}{\mathrm{TTR}_{tgt}}\right)^2.$$

- 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.