Beyond Noise: Mitigating the Impact of Fine-grained Semantic Divergences on Neural Machine Translation

Eleftheria Briakou and **Marine Carpuat**
Department of Computer Science
University of Maryland
College Park, MD 20742, USA
ebriakou@cs.umd.edu, marine@cs.umd.edu

Abstract

While it has been shown that Neural Machine Translation (NMT) is highly sensitive to noisy parallel training samples, prior work treats all types of mismatches between source and target as noise. As a result, it remains unclear how samples that are mostly equivalent but contain a small number of semantically divergent tokens impact NMT training. To close this gap, we analyze the impact of different types of fine-grained semantic divergences on Transformer models. We show that models trained on synthetic divergences output degenerated text more frequently and are less confident in their predictions. Based on these findings, we introduce a divergent-aware NMT framework that uses factors to help NMT recover from the degradation caused by naturally occurring divergences, improving both translation quality and model calibration on EN↔FR tasks.

1 Introduction

While parallel texts are essential to Neural Machine Translation (NMT), the degree of parallelism varies widely across samples in practice, for reasons ranging from noise in the extraction process (Roziewski and Stokowiec, 2016) to non-literal translations (Zhai et al., 2019b, 2020a). For instance (Figure 1), a French SOURCE could be paired with an exact translation into English (EQ), with a mostly equivalent translation where only a few tokens convey divergent meaning (fine-DIV), or with a semantically unrelated, noisy reference (coarse-DIV). Yet, prior work treats parallel samples in a binary fashion: coarse-grained divergences are viewed as noise to be excluded from training (Koehn et al., 2018), whilst others are typically regarded as gold-standard equivalent translations. As a result, the impact of fine-grained divergences on NMT remains unclear.

This paper aims to understand and mitigate the impact of fine-grained semantic divergences in

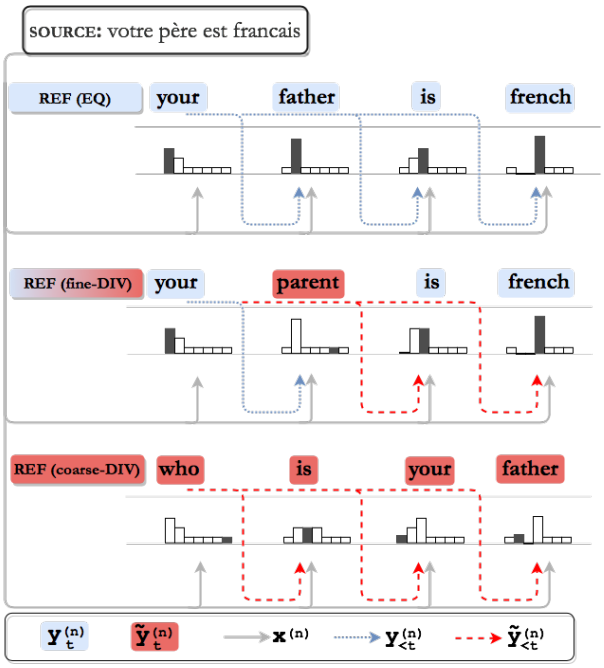


Figure 1: Equivalent vs. Divergent references on NMT training. Fine-grained divergences (i.e., REF (fine-DIV)) provide an imperfect yet potentially useful signal depending on the time step t.

NMT. We first contribute an **analysis** of how fine-grained divergences in training data affect NMT quality and confidence. Starting from a set of equivalent English-French WikiMatrix sentence pairs, we simulate divergences by gradually "corrupting" them with synthetic *fine-grained divergences*. Following Khayrallah and Koehn (2018)—who, in contrast, study the impact of *noise* on MT—we control for different *types* of fine-grained semantic divergences and different *ratios* of equivalent vs. divergent data. Our findings indicate that these imperfect training references: hurt translation quality (as measured by BLEU and METEOR) once they overwhelm equivalents; output degenerated text more frequently; and increase the uncertainty of models' predictions.

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 7236–7249
August 1–6, 2021.

Based on these findings, we introduce a **divergent-aware NMT framework** that incorporates information about which tokens are indicative of semantic divergences between the source and target side of a training sample. Source-side divergence tags are integrated as feature factors (Haddow and Koehn, 2012; Sennrich and Haddow, 2016; Hoang et al., 2016), while target-side divergence tags form an additional output sequence generated in a multi-task fashion (García-Martínez et al., 2016, 2017). Results on EN↔FR translation show that our approach is a successful **mitigation strategy**: it helps NMT recover from the negative impact of fine-grained divergences on translation quality, with fewer degenerated hypotheses, and more confident and better calibrated predictions. We make our code publicly available: https://github.com/Elbria/xling-SemDiv-NMT.

2 Background & Motivation

Cross-lingual Semantic Divergences We use this term to refer to meaning differences in aligned bilingual text (Vyas et al., 2018; Carpuat et al., 2017). Divergences in manual translation might arise due to the translation process (Zhai et al., 2018) and result in non-literal translations (Zhai et al., 2020a). Divergences might also arise in parallel text extracted from multilingual comparable resources. For instance, in Wikipedia, documents aligned across languages might contain parallel segments that share important content, yet they are not perfect translations of each other, yielding fine-grained semantic divergences (Smith et al., 2010). Finally coarse-grained divergences might result from the process of automatically mining and aligning corpora from monolingual data (Fung and Cheung, 2004; Munteanu and Marcu, 2005), or web-scale parallel text (Smith et al., 2013; El-Kishky et al., 2020; Esplà et al., 2019).

Noise vs. Semantic Divergences In the context of MT, noise often refers to mismatches in **web-crawled** parallel corpora that are collected without guarantees about their quality. Khayrallah and Koehn (2018) define five frequent types of noise found in the German-English Paracrawl corpus: *misaligned sentences*, *disfluent text*, *wrong language*, *short segments*, and *untranslated sentences*. They examine the impact of noise on translation quality and find that untranslated training instances cause NMT models to copy the input sentence at inference time. Their findings motivated a shared task dedicated to filtering noisy samples from web-crawled data at WMT, since 2018 (Koehn et al., 2018, 2019, 2020). This work moves beyond such coarse divergences and focuses instead on fine-grained divergences that affect a small number of tokens within mostly equivalent pairs and that can be found even in high-quality parallel corpora.

Training Assumptions NMT models are typically trained to maximize the log-likelihood of the training data, $\mathcal{D} \equiv \{(\boldsymbol{x}^{(n)}, \boldsymbol{y}^{(n)})\}_{n=1}^{N}$, where $(\boldsymbol{x}^{(n)}, \boldsymbol{y}^{(n)})$ is the n-th sentence pair consisting of sentences that are **assumed to be translations of each other**. Under this assumption, model parameters are updated to maximize the **token-level** cross-entropy loss:

$$\mathcal{J}(\theta) = \sum_{n=1}^{N} \sum_{t=1}^{T} \log p(y_t^{(n)} \mid \boldsymbol{y}_{<t}^{(n)}, \boldsymbol{x}^{(n)}; \theta) \quad (1)$$

In Figure 1, we illustrate how semantic divergences interact with NMT training. In the case of coarse divergences, both the prefixes $\widetilde{\boldsymbol{y}}_{t<1}^{(n)}$ and targets $\widetilde{y}_t^{(n)}$, yield a noisy training signal at each time step t, which motivates excluding them from the training pool entirely. In the case of fine-grained divergences, the assumption of *semantic equivalence* is only partially broken. Depending on the time step t, we might thus condition the prediction of the next token on partially corrupted prefixes, encourage the model to make a wrong prediction, or do a combination of the above. This suggests that fine-grained divergent samples provide a noisy yet potentially useful training signal depending on the time step. Meanwhile, fine-grained divergences increase uncertainty in the training data, and as a result might impact models' confidence in their predictions, as noisy untranslated samples do (Ott et al., 2018). This work seeks to clarify and mitigate their impact on NMT, accounting for both translation quality and model confidence.

3 Analyzing the Impact of Divergences

3.1 Method

We evaluate the impact of semantic divergences on NMT by injecting increasing amounts of synthetic divergent samples during training, following the methodology of Khayrallah and Koehn (2018) for noise. We focus on three types of divergences, which were found to be frequent in parallel corpora. They are fine-grained as they represent discrepancies between the source and target segments

at a word or phrase level: **LEXICAL SUBSTITUTION** aims at mimicking *particularization* and *generalization* operations resulting from non-literal translations (Zhai et al., 2019a, 2020b); **PHRASE REPLACEMENT** mimics phrasal mistranslations; **SUBTREE DELETION** simulates missing phrasal content from the source or target side.

Synthetic divergent samples are automatically generated by corrupting semantically equivalent sentence pairs, following the methodology introduced by Briakou and Carpuat (2020). Equivalents are identified by their Divergent mBERT classifier that yields an F1 score of 84, on manually annotated WikiMatrix data, despite being trained on synthetic data. For LEXICAL SUBSTITUTION we corrupt equivalents by substituting words with their hypernyms or hyponyms from WordNet, for PHRASE REPLACEMENT we replace sequences of words with phrases of matching POS tags, and for SUBTREE DELETION we randomly delete subtrees in the dependency parse tree of either the source or the target. Having access to those 4 versions of the same corpus (one initial equivalent and three synthetic divergences), we mix equivalents and divergent pairs introducing one type of divergence at a time (corpora statistics are included in D). Finally, we evaluate the translation quality and uncertainty of the resulting translation models.

3.2 Experimental Set-Up

Training Data We train our models on the parallel WikiMatrix French-English corpus (Schwenk et al., 2019), which consists of sentence pairs mined from Wikipedia pages using language-agnostic sentence embeddings (LASER) (Artetxe and Schwenk, 2019). Previous annotations show that 40% of sentence pairs in a random sample contain fine-grained divergences (Briakou and Carpuat, 2020).

After cleaning noisy samples using simple rules (i.e., exclude pairs that are a) too short or too long, b) mostly numbers, c) almost copies based on edit distance), we extract *equivalent* samples using the Divergent mBERT model. Table 1 presents statistics on the extracted pairs, along with the corpus created if we threshold the LASER score at 1.04, as suggested by Schwenk et al. (2019).

Development and Test data We use the official development and test splits of the TED corpus (Qi et al., 2018), consisting of 4,320 and 4,866 gold-standard translation pairs, respectively. All models share the same BPE vocabulary. We average results across runs with 3 different random seeds.

Corpus	#Sentences
WIKIMATRIX	6,562,360
+ HEURISTIC FILTERING	2,437,108
+ LASER FILTERING	1,250,683
+ divergentmBERT FILTERING	751,792

Table 1: WikiMatrix EN-FR corpus statistics.

Preprocessing We use the standard Moses scripts (Koehn et al., 2007) for punctuation normalization, true-casing, and tokenization. We learn 32K BPEs (Sennrich et al., 2016c) using SentencePiece (Kudo and Richardson, 2018).

Models We use the base Transformer architecture (Vaswani et al., 2017), with embedding size of 512, transformer hidden size of 2,048, 8 attention heads, 6 transformer layers, and dropout of 0.1. Target embeddings are tied with the output layer weights. We train with label smoothing (0.1). We optimize with Adam (Kingma and Ba, 2015) with a batch size of 4,096 tokens and checkpoint models every 1,000 updates. The initial learning rate is 0.0002, and it is reduced by 30% after 4 checkpoints without validation perplexity improvement. We stop training after 20 checkpoints without improvement. We select the best checkpoint based on validation BLEU (Papineni et al., 2002). All models are trained on a single GeForce GTX 1080 GPU.

3.3 Findings

Translation Quality Table 2 presents the impact of semantic divergences on BLEU and METEOR. Corrupting equivalent bitext with fine-grained divergences hurts translation quality across the board. In most cases, the degradation is proportional to the percentage of corrupted training samples. LEXICAL SUBSTITUTION causes the largest degradation for both metrics. The degradation is relatively smaller for METEOR than BLEU, which we attribute to the fact that METEOR allows matches between synonyms when comparing references to hypotheses. SUBTREE DELETION and LEXICAL SUBSTITUTION corruptions lead to significant degradation at $\geq 50\%$ (BLEU; standard deviations across reruns are < 0.4). By contrast, Transformers are more robust to PHRASE REPLACEMENT corruptions, as degradations are only significant after corrupting $\geq 70\%$ (BLEU) of equivalents.

	BLEU						METEOR					
	0%	10%	20%	50%	70%	100%	0%	10%	20%	50%	70%	100%
REPLACEMENT	30.89	31.00	30.82	30.40	29.74	27.01	33.74	33.63	33.66	33.54	33.12	31.02
	+0.00	+0.11	-0.07	-0.49	-1.15	-3.88	+0.00	-0.11	-0.08	-0.20	-0.62	-2.72
DELETION	30.89	30.80	30.62	28.95	29.00	27.50	33.74	33.61	33.38	32.17	32.09	31.44
	+0.00	-0.09	-0.27	-1.94	-1.89	-3.39	+0.00	-0.13	-0.36	-1.57	-1.65	-2.30
SUBSTITUTION	30.89	30.72	30.49	25.04	26.57	25.18	33.74	33.56	33.50	29.59	31.58	30.75
	+0.00	-0.17	-0.40	-5.85	-4.32	-5.71	+0.00	-0.18	-0.24	-4.15	-2.16	-2.99

Table 2: Results for FR→EN translation on the TED test set (means of 3 runs). Bars denote degradation over EQUIVALENTS (i.e., 0%) across different % of corruption. Divergences hurt BLEU and METEOR when they overwhelm the training data. Transformers are particularly sensitive to fine nuances introduced by LEXICAL SUBSTITUTION.

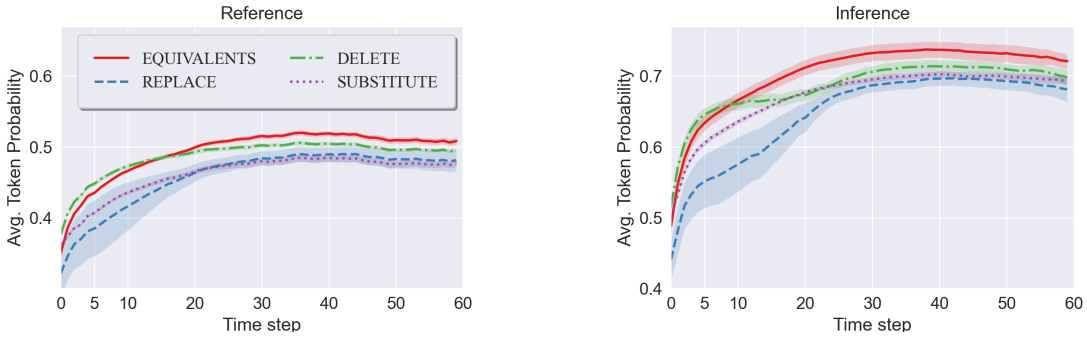


Figure 2: Average token probabilities of predictions conditioned on gold references (left) and beam search (5) prefixes (right). Training on fine-grained divergences (100% corruption) increase NMT model's uncertainty.

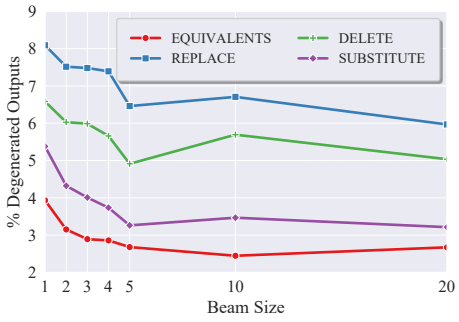


Figure 3: % of degenerated outputs as a function of beam size. NMT training on fine-grained divergences (100% corruptions) increase the frequency of degenerated hypotheses across beams.

Token Uncertainty We measure the impact of divergences on model uncertainty at training time and at test time. For the first, we extract the probability of a reference token conditioned on reference prefixes at each time step. For the latter, we compute the probability of the token predicted by the model given its own history of predictions. Figure 2 shows that models trained on EQUIVALENTS are more confident in their token level predictions both at inference and training time. SUBTREE DELETION mismatches affect models' confidence less than other types, while PHRASE REPLACEMENT hurts confidence the most both at inference and at training time. Finally, we observe that differences across divergence types are larger in early decoding steps, while at later steps, they all converge below the EQUIVALENTS.

Degenerated Hypotheses When models are trained on 50% or more divergent samples, the total length of their hypotheses is longer than the references. Manual analysis on models trained with 100% of divergent samples suggests that this length effect is partially caused by *degenerated* text. Following Holtzman et al. (2019)—who study this phenomenon for unconditional text generation—we define *degenerations* as "output text that is bland, incoherent, or gets stuck in repetitive loops".[1]

[1]For instance, "I've never studied sculpture, engineering and architecture, **and the engineering and architecture".**

We automatically detect degenerated text in model outputs by checking whether they contain repetitive loops of n-grams that do not appear in the reference (details on the algorithm are in C). Figure 3 shows that exposing NMT to divergences increases the percentage of degenerated outputs. Even with large beams, the models trained on divergent data yield more repetitions than the EQUIVALENTS. Moreover, divergences due to phrasal mismatches (PHRASE REPLACEMENT and SUBTREE DELETION) yield more frequent repetitions than token-level mismatches (LEXICAL SUBSTITUTION). Interestingly, the latter almost matches the frequency of repetitions in EQUIVALENTS with larger beams (≥ 5).

Summary Synthetic divergences hurt translation quality, as expected. More surprisingly, our study also reveals that this degradation is partially due to more frequent degenerated outputs, and that divergences impact models' confidence in their predictions. Different types of divergences have different effects: LEXICAL SUBSTITUTION causes the largest degradation in translation quality, SUBTREE DELETION and PHRASE REPLACEMENT increase the number of degenerated beam hypotheses, while PHRASE REPLACEMENT also hurts the models' confidence the most. Nevertheless, the impact of divergences on BLEU appears to be smaller than that of noise (Khayrallah and Koehn, 2018).[2] This suggests that noise filtering techniques are suboptimal to deal with fine-grained divergences.

4 Mitigating the Impact of Fine-grained Divergences

We now turn to naturally occurring divergences in WikiMatrix. We will see that their impact on model quality and uncertainty is consistent with that of synthetic divergences (§ 4.3). We propose a divergent-aware framework for NMT (§ 4.1) that successfully mitigates their impact (§ 4.3).

4.1 Factorizing Divergences for NMT

We use **semantic factors** to inform NMT of tokens that are indicative of meaning differences in each sentence pair. We tag divergent source and target tokens in parallel segments as equivalent (EQ) or divergent (DIV) using an mBERT-based classifier trained on synthetic data. The classifier has a 45 F1 score on a fine-grained divergence test set (Briakou and Carpuat, 2020). The predicted tags are thus noisy, as expected on this challenging task, yet we will see that they are useful. An example is illustrated below:

SRC	TOKENS	*votre*	*père*	*est*	*francais*
	FACTORS	EQ	DIV	EQ	EQ
TGT	TOKENS	*your*	*parent*	*is*	*french*
	FACTORS	EQ	DIV	EQ	EQ

Source Factors We follow Sennrich and Haddow (2016) who represent the encoder input as a combination of token embeddings and linguistic features. Concretely, we look up separate embeddings vectors for tokens and source-side divergent predictions, which are then concatenated. The length of the concatenated vector matches the total embedding size.

Target Factors Target-side divergence tags are an additional output sequence, as in García-Martínez et al. (2016). At each time step the model produces two distributions: one over the token target vocabulary and one over the target factors. The model is trained to minimize a divergent-aware loss (Equation 2). Terms in red (also, underlined) correspond to modifications to the traditional NMT loss. At time step t, the model is rewarded to match the reference target $y_t^{(n)}$, conditioned on the source sequence of tokens ($\boldsymbol{x}^{(n)}$), the source factors ($\boldsymbol{\omega}^{(n)}$), the token target prefix ($\boldsymbol{y}_{<t}^{(n)}$), and the target factors prefix ($\boldsymbol{z}_{<t}^{(n)}$). At the same time ($t$), the model is rewarded to match the factored predictions for the previous time step $\tau = t-1$. The time shift between the two target sequences is introduced so that the model learns to firstly predict the reference token at τ and then its corresponding EQ vs. DIV label, at the same time step. The factored predictions are conditioned again on $\boldsymbol{x}^{(n)}$, $\boldsymbol{\omega}^{(n)}$, the target factor prefix $\boldsymbol{z}_{<\tau}^{(n)}$ and the token prefix ($\boldsymbol{y}_{\leq\tau}^{(n)}$).

$$\mathcal{L} = -\sum_{n=1}^{N} \Bigg(\underbrace{\sum_{t=1}^{T} \log p(y_t^{(n)} \mid \boldsymbol{y}_{<t}^{(n)}, \underline{\boldsymbol{z}_{<t}^{(n)}}, \boldsymbol{x}^{(n)}, \underline{\boldsymbol{\omega}^{(n)}}; \theta)}_{\mathcal{L}_{\text{MT}}^{(n)}} + \underbrace{\sum_{\tau=t-1}^{T} \log p(z_\tau^{(n)} \mid \boldsymbol{z}_{<\tau}^{(n)}, \boldsymbol{y}_{\leq\tau}^{(n)}, \boldsymbol{x}^{(n)}, \boldsymbol{\omega}^{(n)}; \theta)}_{\underline{\mathcal{L}_{factor}^{(n)}}} \Bigg) \tag{2}$$

[2] While the absolute scores are not directly comparable across settings, Khayrallah and Koehn (2018) report that noise has a more striking impact of -8 to -25 BLEU.

Inference At test time, input tokens are tagged with EQ to encourage the model to predict an equivalent translation. We decode using beam search for predicting the translation sequence. The token predictions are conditioned on both the token and the factors prefixes. The factor prefixes are greedily decoded and thus do not participate in beam search.

4.2 Experimental Set-Up

Divergences We conduct an extensive comparison of models exposed to different amounts of equivalent and divergent WikiMatrix samples. Starting from the pool of examples identified as divergent at §3.2, we rank and select the most fine-grained divergences by thresholding the `bicleaner` score (Ramírez-Sánchez et al., 2020) at 0.5, 0.7 and 0.8. For details, see A.

Models We compare the factored models (**DIV-FACTORIZED**) for incorporating divergent tokens (§4.1) against: 1. **LASER** models are trained on WikiMatrix pairs with a LASER score greater than 1.04 – the noise filtering strategy recommended by Schwenk et al. (2019). Our prior work shows that thresholding LASER might introduce a number of divergent data in the training pool varying from fine to coarse mismatches (Briakou and Carpuat, 2020). 2. **EQUIVALENTS** models are trained on WikiMatrix pairs detected as exact translations (§3.2); 3. **DIV-AGNOSTIC** models are trained on equivalent and fine-grained divergent data without incorporating information that distinguishes between them; 4. **DIV-TAGGED** models distinguish equivalences from divergences by appending <EQ> vs. <DIV> tags as source-side constraints (Sennrich et al., 2016a).

Models' details Our models are implemented in the `Sockeye2` toolkit (Domhan et al., 2020).[3] We set the size of factor embeddings to 8, the source token embeddings to 504 and target embeddings to 514, yielding equal model sizes across experiments. All other parameters are kept the same across models, as discussed in §3.2, except that target embeddings are not tied with output layer weights for factored models. More details are included in B.

Other Data & Preprocessing We use the same preprocessing as well as development and test sets as in §3.2, except we learn 5K BPEs as in Schwenk et al. (2019). DIV-FACTORIZED, DIV-AGNOSTIC, and DIV-TAGGED models are compared in controlled setups that use the same training data. We also evaluate out-of-domain on the `khresmoi-summary` test set for the WMT2014 medical translation task (Bojar et al., 2014).

Evaluation We evaluate translation quality with BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005).[4,5] We compute Inference Expected Calibration Error (InfECE) as Wang et al. (2020), which measures the difference in expectation between confidence and accuracy.[6] We measure token-level translation accuracy based on Translation Error Rate (TER) alignments between hypotheses and references.[7] Unless mentioned otherwise, we decode with a beam size of 5.

4.3 Results

We discuss the impact of real divergences along the dimensions surfaced by the synthetic data analysis.

Translation Quality Table 3 presents BLEU and METEOR scores across model configurations and data settings on the TED test sets. First, the model trained on EQUIVALENTS represents a very competitive baseline as it performs better or statistically comparable to all models. This result is in line with prior evidence of Vyas et al. (2018) who show that filtering out the most divergent pairs in noisy corpora (e.g., OpenSubtitles and CommonCrawl) does not hurt translation quality. Interestingly, the EQUIVALENTS model outperforms LASER across metrics and translation directions, despite the fact that it is exposed to only about half of the training data. Gradually adding divergent data (DIV-AGNOSTIC) hurts translation quality across the board compared to the EQUIVALENTS model. The drops are significantly larger when divergences overwhelm the equivalent translations, which is consistent with our findings on synthetic data.

Second, DIV-FACTORIZED is the most effective mitigation strategy. With segment-level constraints (DIV-TAGGED), models can recover from the degradation caused by divergences (DIV-AGNOSTIC), but not consistently. By contrast, token-level factors (DIV-FACTORIZED) help NMT recover from the impact of divergences across data setups and reach

[3] https://github.com/awslabs/sockeye

[4] https://github.com/mjpost/sacrebleu

[5] https://www.cs.cmu.edu/~alavie/METEOR/

[6] https://github.com/shuo-git/InfECE

[7] http://www.cs.umd.edu/~snover/tercom/

METHOD		Training size	FR→EN BLEU	FR→EN METEOR	EN→FR BLEU	EN→FR METEOR
LASER		1.25M	31.80 ±0.36	34.00 ±0.17	32.16 ±0.29	56.49 ±0.24
EQUIVALENTS		0.75M	32.88 ±0.07	34.75 ±0.10	33.53 ±0.35	57.38 ±0.28
+DIV	AGNOSTIC	0.93M	32.47 ±0.40	34.56 ±0.20	33.19 ±0.30	57.10 ±0.30
+DIV	TAGGED	0.93M	31.76 ±1.61	34.17 ±0.91	**33.43** ±0.39	**57.55** ±0.27 ↑
+DIV	FACTORIZED	0.93M	32.73 ±0.38	**34.84** ±0.21 ↑	**33.92** ±0.38 ↑	**57.63** ±0.28 ↑
+DIV	AGNOSTIC	1.12M	32.53 ±0.46	34.40 ±0.21	31.47 ±0.61	56.25 ±0.46
+DIV	TAGGED	1.12M	32.38 ±0.40	34.52 ±0.13	**33.35** ±0.17 ↑	**57.33** ±0.14 ↑
+DIV	FACTORIZED	1.12M	32.79 ±0.24	**34.89** ±0.12 ↑	**33.22** ±0.35 ↑	**57.31** ±0.30 ↑
+DIV	AGNOSTIC	1.68M	31.40 ±0.21	33.79 ±0.11	29.53 ±0.39	54.29 ±0.44
+DIV	TAGGED	1.68M	31.97 ±0.26 ↑	34.30 ±0.10 ↑	31.37 ±0.12 ↑	55.87 ±0.18 ↑
+DIV	FACTORIZED	1.68M	32.57 ±0.19 ↑	**34.70** ±0.11 ↑	31.60 ±0.42 ↑	56.10 ±0.22 ↑

Table 3: Results for EN↔FR translation on the TED test set (averages and stdev of 3 runs). We underline the top scores among all models and boldface the scores lying within one stdev from EQUIVALENTS. ↑ denotes (one stdev) improvements of DIV-TAGGED and DIV-FACTORIZED over DIV-AGNOSTIC. Factorizing divergences helps NMT recover from the degradation caused by divergences, while it achieves comparable scores to EQUIVALENTS.

METHOD	Train. size (M)	FR→EN	EN→FR
LASER	1.25	38.27 ±0.49	39.27 ±0.45
EQUIVALENTS	0.75	39.47 ±0.24	39.63 ±0.52
DIV-AGNOSTIC	0.93	39.45 ±0.50	39.78 ±0.37
DIV-AGNOSTIC	1.12	**40.00** ±0.14	39.20 ±0.50
DIV-AGNOSTIC	1.68	**39.90** ±0.14	38.00 ±0.50
DIV-FACTORIZED	0.93	**40.27** ±0.49	40.13 ±0.46
DIV-FACTORIZED	1.12	**40.03** ±0.42	**40.30** ±0.29
DIV-FACTORIZED	1.68	**39.97** ±0.26	39.30 ±0.16

Table 4: BLEU scores on the medical domain. We underline top scores and boldface (one stdev) improvements over EQUIVALENTS. Divergences improve translation quality when modeled by DIV-FACTORIZED.

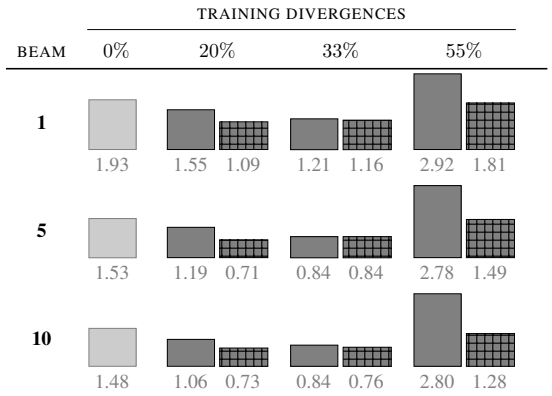


Table 5: Percentage of degenerated outputs for FR→EN models exposed to difference percentage of divergent training data (0% corresponds to EQUIVALENTS; dark gray columns correspond to DIV-AGNOSTIC). DIV-FACTORIZED (grid-columns) help recover from degenerations, yielding fewer repetitions across beams.

translation quality comparable to that of the EQUIVALENTS model, successfully mitigating the impact of the noisy training signals from divergent samples.

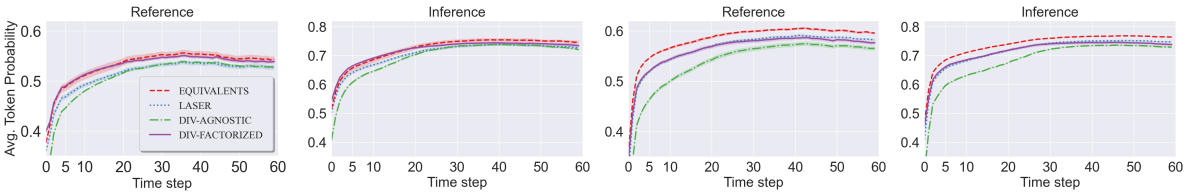
Third, when translating the out-of-domain test set, DIV-FACTORIZED improves over the EQUIVALENTS model, as presented in Table 4. DIV-AGNOSTIC models perform comparably to EQUIVALENTS, while factorizing divergences improves on the latter by $\approx +1$ BLEU, for both directions.[8] Mitigating the impact of divergences is thus important for NMT to benefit from the increased coverage of out-of-domain data provided by the divergent samples.

Degenerated Hypotheses We check for degenerated outputs across models, data setups (we account for different percentages of divergences in the training data), and different beam sizes (Table 5). As with synthetic divergences, we observe that when real divergences overwhelm the training data (55%), degenerated loops are almost twice as frequent for all beam sizes. This phenomenon is consistently mitigated by DIV-FACTORIZED models across the board.[9] Furthermore, in some settings (20%, 33%), DIV-FACTORIZED models decrease the amount of degenerated text by half compared to the EQUIVALENTS models.[10]

[8] We include METEOR results in Appendix E.

[9] We observe similar trends for EN→FR in Appendix F

[10] LASER models degenerate more frequently than EQUIVALENTS and DIV-FACTORIZED.



(a) FR→EN (b) FR→EN (c) EN→FR (d) EN→FR

Figure 4: Average token probability across time steps on TED test set. DIV-AGNOSTIC yield the least confident predictions (for reference and inference prefixes); DIV-FACTORIZED help recover from this drop (55% divergences).

		FR→EN			EN→FR		
METHOD	Train. size	CONF. (↑)	ACC. (↑)	InfECE (↓)	CONF.% (↑)	ACC. (↑)	InfECE (↓)
LASER	1.25M	69.09 ±0.67	62.55 ±0.29	12.34 ±0.38	71.88 ±0.30	60.20 ±0.18	15.10 ±0.12
EQUIVALENTS	0.75M	70.96 ±0.94	63.49 ±0.10	12.37 ±0.24	74.35 ±0.23	61.81 ±0.30	15.09 ±0.18
DIV-AGNOSTIC	0.93M	71.19 ±0.33	63.54 ±0.54	**12.00** ±0.06	73.67 ±0.11	61.44 ±0.22	15.19 ±0.17
DIV-FACTORIZED		**72.16** ±0.10*	**64.29** ±0.44*	**11.81** ±0.04*	74.50 ±0.02*	**62.26** ±0.27*	**14.70** ±0.25*
DIV-AGNOSTIC	1.12M	71.65 ±0.18	61.34 ±0.33	**11.98** ±0.22	71.72 ±0.38	59.29 ±0.48	15.62 ±0.19
DIV-FACTORIZED		71.83 ±0.03	**64.38** ±0.08*	**11.86** ±0.01	74.09 ±0.14*	61.65 ±0.19*	**14.84** ±0.18*
DIV-AGNOSTIC	1.68M	68.01 ±0.37	61.34 ±0.23	12.63 ±0.23	68.38 ±0.25	56.89 ±0.34	16.24 ±0.27
DIV-FACTORIZED		71.01 ±0.39*	63.65 ±0.07*	**11.75** ±0.35*	71.81 ±0.49*	59.78 ±0.39*	14.95 ±0.02*

Table 6: Average token confidence, accuracy, and inference calibration results for EN↔FR translation on the TED test set (average and stdev of 3 runs). We underline top scores and boldface (one stdev) improvements over EQUIVALENTS. * denotes (one stdev) improvements of DIV-FACTORIZED over DIV-AGNOSTIC. DIV-FACTORIZED yield more confident and accurate predictions compared to DIV-AGNOSTIC, yielding the smallest calibration errors.

Uncertainty Figures 4a and 4c show that the gold-standard references are assigned lower probabilities by the DIV-AGNOSTIC models than all other models, especially in early time steps ($t < 30$). We observe similar drops in confidence based on the probabilities of predicted tokens at inference time (4b and 4d). This confirms that exposing models to fine-grained semantic divergences hurts their confidence, whether the divergences are synthetic or not. Furthermore, factorizing divergences helps mitigate the impact of naturally occurring divergences on uncertainty in addition to translation quality.

We conduct a calibration analysis to measure the differences between the confidence (i.e., *probability*) and the correctness (i.e., *accuracy*) of the generated tokens in expectation. Given that deep neural networks are often mis-calibrated in the direction of over-estimation (confidence>accuracy) (Guo et al., 2017), we check whether the increased confidence of DIV-FACTORIZED hurts calibration (Table 6). DIV-FACTORIZED models are on average more confident *and* more accurate than their DIV-AGNOSTIC counterparts. Interestingly, DIV-AGNOSTIC has smaller calibration errors than EQUIVALENTS and LASER models across the board.

5 Related Work

We discuss work related to cross-lingual semantic divergences and noise effects in Section 2 and now turn to the literature that connects with the methods used in this paper.

Factored Models Factored models are introduced to inject word-level linguistic annotations (e.g., Part-of-Speech tags, lemmas) in translation. Source-side factors have been used in statistical MT (Haddow and Koehn, 2012) and in NMT (Sennrich et al., 2016b; Hoang et al., 2016). Target-side factors are used by García-Martínez et al. (2017) as an extension to the traditional NMT framework that outputs multiple sequences. Although their main motivation is to enable models to handle larger vocabularies, Wilken and Matusov (2019) propose a list of novel applications of target-side factors beyond their initial purpose, such as word-case prediction and subword segmentation. Our approach draws inspiration from all the aforementioned works, yet it is unique in its use of *both* source and target factors to incorporate *semantics* in NMT.

Calibration Kumar and Sarawagi (2019) find that NMT models are miscalibrated, even when conditioned on gold-standard prefixes. They attribute this behavior to the poor calibration of the EOS token and the uncertainty of attention and design a recalibration model to improve calibration. Ott et al. (2018) argue that miscalibration can be attributed to the "extrinsic" uncertainty of the noisy, untranslated references found in the training data. Müller et al. (2019) investigate the effect of label smoothing on calibration. On a similar spirit, Wang et al. (2020) propose graduated label smoothing to improve calibration at inference time. They also link miscalibration to linguistic properties of the data (e.g., frequency, position, syntactic roles). Our work, in contrast, focuses on the semantic properties of the training data that affect calibration.

6 Conclusion

This work investigates the impact of semantic mismatches beyond noise in parallel text on NMT quality and confidence. Our experiments on EN↔FR tasks show that fine-grained semantic divergences hurt translation quality when they overwhelm the training data. Models exposed to fine-grained divergences at training time are less confident in their predictions, which hurts beam search and produces degenerated text (repetitive loops) more frequently.

Furthermore, we also show that, unlike noisy samples, fine-grained divergences can still provide a useful training signal for NMT when they are modeled via factors. Evaluated on EN↔FR translation tasks, our divergent-aware NMT framework mitigates the negative impact of divergent references on translation quality, improves the confidence and calibration of predictions, and produces degenerated text less frequently.

More broadly, this work illustrates how understanding the properties of training data can help build better NMT models. In future work, we will extend our analysis to other properties of parallel text and to other language pairs, focusing on low-resource conditions where divergences are expected to be even more prevalent.

Acknowledgements

We thank Sweta Agrawal, Doug Oard, Suraj Rajappan Nair, the anonymous reviewers and the CLIP lab at UMD for helpful comments. This material is based upon work supported by the National Science Foundation under Award No. 1750695. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

M. Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

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.

Rachel Bawden, Kevin Bretonnel Cohen, Cristian Grozea, Antonio Jimeno Yepes, Madeleine Kittner, Martin Krallinger, Nancy Mah, Aurelie Neveol, Mariana Neves, Felipe Soares, Amy Siu, Karin Verspoor, and Maika Vicente Navarro. 2019. Findings of the WMT 2019 biomedical translation shared task: Evaluation for MEDLINE abstracts and biomedical terminologies. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 29–53, Florence, Italy. Association for Computational Linguistics.

Rachel Bawden, Giorgio Maria Di Nunzio, Cristian Grozea, Inigo Jauregi Unanue, Antonio Jimeno Yepes, Nancy Mah, David Martinez, Aurélie Névéol, Mariana Neves, Maite Oronoz, Olatz Perez-de Viñaspre, Massimo Piccardi, Roland Roller, Amy Siu, Philippe Thomas, Federica Vezzani, Maika Vicente Navarro, Dina Wiemann, and Lana Yeganova. 2020. Findings of the WMT 2020 biomedical translation shared task: Basque, Italian and Russian as new additional languages. In *Proceedings of the Fifth Conference on Machine Translation*, pages 660–687, Online. Association for Computational Linguistics.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Eleftheria Briakou and Marine Carpuat. 2020. Detecting Fine-Grained Cross-Lingual Semantic Divergences without Supervision by Learning to Rank. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1563–1580, Online. Association for Computational Linguistics.

Marine Carpuat, Yogarshi Vyas, and Xing Niu. 2017. Detecting cross-lingual semantic divergence for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 69–79, Vancouver. Association for Computational Linguistics.

Tobias Domhan, Michael Denkowski, David Vilar, Xing Niu, Felix Hieber, and Kenneth Heafield. 2020. The sockeye 2 neural machine translation toolkit at AMTA 2020. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 110–115, Virtual. Association for Machine Translation in the Americas.

Ahmed El-Kishky, Philipp Koehn, and Holger Schwenk. 2020. *Searching the Web for Cross-Lingual Parallel Data*, page 2417–2420. Association for Computing Machinery, New York, NY, USA.

Miquel Esplà, Mikel Forcada, Gema Ramírez-Sánchez, and Hieu Hoang. 2019. ParaCrawl: Web-scale parallel corpora for the languages of the EU. In *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*, pages 118–119, Dublin, Ireland. European Association for Machine Translation.

Miquel Esplà-Gomis, Víctor M. Sánchez-Cartagena, Jaume Zaragoza-Bernabeu, and Felipe Sánchez-Martínez. 2020. Bicleaner at WMT 2020: Universitat d'alacant-prompsit's submission to the parallel corpus filtering shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 952–958, Online. Association for Computational Linguistics.

Pascale Fung and Percy Cheung. 2004. Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpus. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 1051–1057, Geneva, Switzerland. COLING.

Mercedes García-Martínez, Loïc Barrault, and Fethi Bougares. 2016. Factored neural machine translation. *CoRR*, abs/1609.04621.

Mercedes García-Martínez, Loïc Barrault, and Fethi Bougares. 2017. Neural machine translation by generating multiple linguistic factors. *CoRR*, abs/1712.01821.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330, International Convention Centre, Sydney, Australia. PMLR.

Barry Haddow and Philipp Koehn. 2012. Interpolated backoff for factored translation models. In *Association for Machine Translation in the Americas, AMTA*.

Cong Duy Vu Hoang, Gholamreza Haffari, and Trevor Cohn. 2016. Improving neural translation models with linguistic factors. In *Proceedings of the Australasian Language Technology Association Workshop 2016*, pages 7–14, Melbourne, Australia.

Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *CoRR*, abs/1904.09751.

Antonio Jimeno Yepes, Aurélie Névéol, Mariana Neves, Karin Verspoor, Ondřej Bojar, Arthur Boyer, Cristian Grozea, Barry Haddow, Madeleine Kittner, Yvonne Lichtblau, Pavel Pecina, Roland Roller, Rudolf Rosa, Amy Siu, Philippe Thomas, and Saskia Trescher. 2017. Findings of the WMT 2017 biomedical translation shared task. In *Proceedings of the Second Conference on Machine Translation*, pages 234–247, Copenhagen, Denmark. Association for Computational Linguistics.

Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 74–83, Melbourne, Australia. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Philipp Koehn, Vishrav Chaudhary, Ahmed El-Kishky, Naman Goyal, Peng-Jen Chen, and Francisco Guzmán. 2020. Findings of the WMT 2020 shared task on parallel corpus filtering and alignment. In *Proceedings of the Fifth Conference on Machine Translation*, pages 726–742, Online. Association for Computational Linguistics.

Philipp Koehn, Francisco Guzmán, Vishrav Chaudhary, and Juan Pino. 2019. Findings of the WMT 2019 shared task on parallel corpus filtering for low-resource conditions. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 54–72, Florence, Italy. Association for Computational Linguistics.

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.

Philipp Koehn, Huda Khayrallah, Kenneth Heafield, and Mikel L. Forcada. 2018. Findings of the WMT

2018 shared task on parallel corpus filtering. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 726–739, Belgium, Brussels. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Aviral Kumar and Sunita Sarawagi. 2019. Calibration of encoder decoder models for neural machine translation. *CoRR*, abs/1903.00802.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. 2019. When does label smoothing help? In *Advances in Neural Information Processing Systems*, volume 32, pages 4694–4703. Curran Associates, Inc.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Comput. Linguist.*, 31(4):477–504.

Mariana Neves, Antonio Jimeno Yepes, Aurélie Névéol, Cristian Grozea, Amy Siu, Madeleine Kittner, and Karin Verspoor. 2018. Findings of the WMT 2018 biomedical translation shared task: Evaluation on Medline test sets. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 324–339, Belgium, Brussels. Association for Computational Linguistics.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. *arXiv preprint arXiv:1803.00047*.

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.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? 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 529–535, New Orleans, Louisiana. Association for Computational Linguistics.

Gema Ramírez-Sánchez, Jaume Zaragoza-Bernabeu, Marta Bañón, and Sergio Ortiz Rojas. 2020. Bifixer and bicleaner: two open-source tools to clean your parallel data. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 291–298, Lisboa, Portugal. European Association for Machine Translation.

Szymon Roziewski and Wojciech Stokowiec. 2016. LanguageCrawl: A generic tool for building language models upon Common-Crawl. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2789–2793, Portorož, Slovenia. European Language Resources Association (ELRA).

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *CoRR*, abs/1907.05791.

Rico Sennrich and Barry Haddow. 2016. Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, pages 83–91, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40, San Diego, California. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016c. 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.

Jason R. Smith, Chris Quirk, and Kristina Toutanova. 2010. Extracting parallel sentences from comparable corpora using document level alignment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 403–411, Los Angeles, California. Association for Computational Linguistics.

Jason R. Smith, Herve Saint-Amand, Magdalena Plamada, Philipp Koehn, Chris Callison-Burch, and Adam Lopez. 2013. Dirt cheap web-scale parallel text from the Common Crawl. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1374–1383, Sofia, Bulgaria. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc.

Yogarshi Vyas, Xing Niu, and Marine Carpuat. 2018. Identifying semantic divergences in parallel text without annotations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1503–1515, New Orleans, Louisiana. Association for Computational Linguistics.

Shuo Wang, Zhaopeng Tu, Shuming Shi, and Yang Liu. 2020. On the inference calibration of neural machine translation. In *ACL*.

Patrick Wilken and Evgeny Matusov. 2019. Novel applications of factored neural machine translation. *CoRR*, abs/1910.03912.

Fangzhou Zhai, Vera Demberg, Pavel Shkadzko, Wei Shi, and Asad Sayeed. 2019a. A hybrid model for globally coherent story generation. In *Proceedings of the Second Workshop on Storytelling*, pages 34–45, Florence, Italy. Association for Computational Linguistics.

Yuming Zhai, Gabriel Illouz, and Anne Vilnat. 2020a. Detecting non-literal translations by fine-tuning cross-lingual pre-trained language models. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5944–5956, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yuming Zhai, Lufei Liu, Xinyi Zhong, Gbariel Illouz, and Anne Vilnat. 2020b. Building an English-Chinese parallel corpus annotated with sub-sentential translation techniques. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4024–4033, Marseille, France. European Language Resources Association.

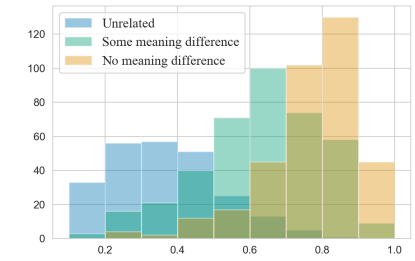
Yuming Zhai, Aurélien Max, and Anne Vilnat. 2018. Construction of a multilingual corpus annotated with translation relations. In *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing*, pages 102–111, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yuming Zhai, Pooyan Safari, Gabriel Illouz, Alexandre Allauzen, and Anne Vilnat. 2019b. Towards recognizing phrase translation processes: Experiments on english-french. *CoRR*, abs/1904.12213.

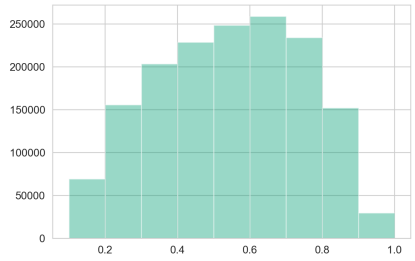
A WikiMatrix Fine-grained Divergences

Starting from the pool of examples identified as divergent under the divergentmBERT classifier, we want to focus on the subset of samples that contain fine meaning differences. Therefore, we use `bicleaner` to filter out training data that are likely to contain coarse meaning differences. Esplà-Gomis et al. (2020) report better NMT results on English↔Portuguese translation after cleaning WikiMatrix data with thresholds of 0.5 and 0.7.

We conduct a preliminary experiment to understand how the `bicleaner` scores of English-French WikiMatrix sentences are distributed. Figure 5(a) shows the distribution of scores among the three classes of the REFRESD dataset, a dataset that distinguishes fine meaning differences ("some meaning difference"), coarse divergences ("unrelated"), and equivalent translation pairs ("no meaning difference").[11] We observe that thresholding the `bicleaner` score at > 0.5 filters out most of the unrelated pairs. We conduct three experiments with thresholds at 0.8, 0.7, and 0.5 to gradually add more fine-grained divergences. Figure 5(b) presents the number of English-French WikiMatrix divergences, binned by `bicleaner` scores.



(a) REFRESD



(b) WIKIMATRIX DIVERGENCES

Figure 5: Distribution of bicleaner score on REFRESD, and English-French WikiMatrix divergences.

[11] https://github.com/Elbria/xling-SemDiv/tree/master/REFreSD

B Sockeye2 configuration details

Tables 7 and 8 present details of NMT training with Sockeye2.

```
--weight-tying-type="trg_softmax"
--num-words 5000:5000
--label-smoothing 0.1
--encoder transformer
--decoder transformer
--num-layers 6
--transformer-attention-heads 84
--transformer-model-size 512
--num-embed 512
--transformer-feed-forward-num-hidden 2048
--transformer-preprocess n
--transformer-postprocess dr
--gradient-clipping-type none
--transformer-dropout-attention 0.1
--transformer-dropout-act 0.1
--transformer-dropout-prepost 0.1
--max-seq-len 80:80
--batch-type word
--batch-size 2048
--min-num-epochs 3
--initial-learning-rate 0.0002
--learning-rate-reduce-factor 0.7
--learning-rate-reduce-num-not-improved 4
--checkpoint-interval 1000
--keep-last-params 30
--max-num-checkpoint-not-improved 20
--decode-and-evaluate 1000
```

Table 7: NMT configurations on Sockeye2 for EQUIVALENTS, LASER, DIV-AGNOSTIC, and DIV-TAGGED.

```
--weight-tying-type none
--source-factors-num-embed 8
--source-factors-combine concat
--target-factors-num-embed 8
--target-factors-combine concat
--transformer-model-size 504:512
--num-embed 504:504
```

Table 8: NMT configurations on Sockeye2 for DIV-FACTORIZED; for missing settings refer to Table 7.

C Measuring Degenerated Hypotheses

We include the pseudo-algorithm that checks if a hypothesis falls under odd repetitions not supported by the reference in Algorithm 1. When measuring repeated n-grams we exclude punctuation and conjunctions. The REPEATED function checks whether an n-gram is repeated (number of occurrences > 1) in the hypothesis h, or reference r.

WikiMatrix version	#Sents.	#Tokens	#Types	Length	%Corr
EQUIVALENTS	751,792	22,723,543	515,154	30.2	0%
SUBTREE DELETION	749,973	20,783,056	483,336	27.7	9.32%
PHRASE REPLACEMENT	750,527	22,735,143	475,567	30.3	16.11%
LEXICAL SUBSTITUTION (HYPERNYMS)	724,326	22,014,609	497,658	30.4	12.33%
LEXICAL SUBSTITUTION (HYPONYMS)	617,913	18,970,039	442,299	30.7	7.42%

Table 9: WikiMatrix statistics corresponding to extracted EQUIVALENTS and the fine-grained corruptions introduced in the synthetic setting (EN-side). %Corr denotes the average % of corrupted tokens in a sentence.

WikiMatrix Version	#Sents.	#Tokens	#Types	Length	%Corr
EQUIVALENTS	751,792	25,554,549	515,194	34.0	0%
SUBTREE DELETION	749,973	23,822,958	486,908	31.8	7.21%
PHRASE REPLACEMENT	750,527	25,554,549	515,194	34.0	12.74%
LEXICAL SUBSTITUTION (HYPERNYMS)	724,326	24,737,604	499,423	34.2	9.82%
LEXICAL SUBSTITUTION (HYPONYMS)	617,913	21,387,650	445,871	34.6	5.78%

Table 10: WikiMatrix statistics corresponding to extracted EQUIVALENTS and the fine-grained corruptions introduced in the synthetic setting (FR-side).

Algorithm 1 Degenerated hypothesis check

Input: h, r (hypothesis, reference)
Output: Deg (True for degenerated hypothesis)

```
function DegenerationCheck(h[ ],r[ ])
    for n-gram ∈ h do
        if Repeated(n-gram, h) then
            if not Repeated(n-gram, r) then
                return True
            end if
        end if
    end for
    return False
end function
```

METHOD	Training data (M)	FR→EN	EN→FR
LASER	1.25	40.67 ±0.28	65.17 ±0.40
EQUIVALENTS	0.75	41.23 ±0.16	65.60 ±0.45
	0.93	41.25 ±0.17	65.67 ±0.22
DIV-AGNOSTIC	1.12	41.19 ±0.31	65.23 ±0.33
	1.68	41.07 ±0.13	63.97 ±0.56
	0.93	41.34 ±0.23	66.03 ±0.38
DIV-FACTORIZED	1.12	41.33 ±0.31	65.88 ±0.35
	1.68	41.25 ±0.08	65.08 ±0.11

Table 11: METEOR scores on medical translation task.

D Synthetic Divergences Statistics

Tables 9 and 10 contain corpus statistics for the 3 versions of synthetic divergences we create, starting from EQUIVALENTS. LEXICAL SUBSTITUTION are sampled at random from the pools of substitutions based on hypernyms and hyponyms.

E METEOR Results (addition)

For completeness, we present METEOR scores to complement the BLEU evaluation of §4.3, which consists the official evaluation metric of WMT biomedical translation tasks (Jimeno Yepes et al., 2017; Neves et al., 2018; Bawden et al., 2019, 2020). The average improvements of DIV-FACTORIZED over EQUIVALENTS and DIV-AGNOSTIC are smaller compared to the differences highlighted by BLEU. However, we note that METEOR results might be misleading when evaluating medical translations, as in this domain we might not want to account for synonyms when comparing references to hypotheses.

F Degenerated Hypotheses (addition)

DIV-FACTORIZED decreases the % of degenerated outputs caused by divergent data (Table 12).

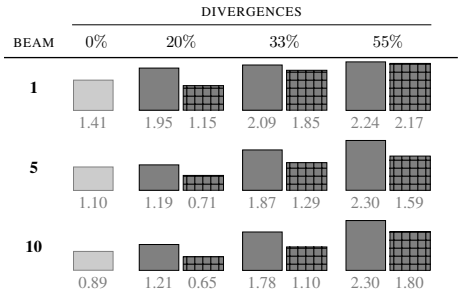


Table 12: % of degenerated outputs across beams (EN→FR). DIV-FACTORIZED (grid-columns) help recover from degenerations, yielding fewer repetitions.