To Diverge or Not to Diverge: A Morphosyntactic Perspective on Machine Translation vs Human Translation

Jiaming Luo and **Colin Cherry** and **George Foster**
Google Translate Research, Canada
{jmluo,colincherry,fosterg}@google.com

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00645/2362224/tacl_a_00645.pdf by guest on 24 April 2024

Abstract

We conduct a large-scale fine-grained comparative analysis of machine translations (MTs) against human translations (HTs) through the lens of morphosyntactic divergence. Across three language pairs and two types of divergence defined as the structural difference between the source and the target, MT is consistently more conservative than HT, with less morphosyntactic diversity, more convergent patterns, and more one-to-one alignments. Through analysis on different decoding algorithms, we attribute this discrepancy to the use of beam search that biases MT towards more convergent patterns. This bias is most amplified when the convergent pattern appears around 50% of the time in training data. Lastly, we show that for a majority of morphosyntactic divergences, their presence in HT is correlated with decreased MT performance, presenting a greater challenge for MT systems.

1 Introduction

Translation divergences occur when the translations differ structurally from the source sentences, typically as a result of either inherent crosslingual differences or idiosyncratic preferences of translators. These divergences happen naturally in the translation process and can be readily found in human translations (HTs), including those used for training machine translation (MT) systems (see the table in Figure 1 for some examples). Their existence in HT has long been regarded as a key challenge for MT (Dorr, 1994) and more recent empirical studies have demonstrated the abundance of translation divergences in HT (Deng and Xue, 2017; Nikolaev et al., 2020).

In contrast to HT, MT outputs tend to be less diverse and more literal (i.e., absence of translation divergence), exhibiting the features of *translationese* (Gellerstam, 1986). This *qualitative* difference between HT and MT has inspired a rich body of work attempting to narrow the gap, such as automatic detection of machine translated texts in the training data (Kurokawa et al., 2009; Lembersky et al., 2012; Aharoni et al., 2014; Riley et al., 2020; Freitag et al., 2022), training MT systems on more diverse translations (Khayrallah et al., 2020; Bao et al., 2023), and carefully reordering the examples to reduce the degree of divergence between the source and the target (Wang et al., 2007; Zhang and Zong, 2016; Zhou et al., 2019). The challenges that translation divergences present do not just concern training MT systems, but also their evaluation (Koppel and Ordan 2011; Freitag et al., 2020).

Nonetheless, even as we gain deepened understanding of how to address these challenges, it remains unclear *how quantitatively different* MT and HT are in terms of divergences.[1] Control verbs,[2] for instance, provide a great case study to showcase this difference. There is much uncertainty when translating them from English to French, and human translators employ a wide variety of constructions including many divergent patterns (Figure 1). In comparison, MT is much more likely to preserve the source structure, with the convergent pattern constituting about 20% more of all translations of control verbs. This difference exemplifies MT's undesirable tendency to produce translationese that is too literal and lacks structural diversity (Freitag et al., 2019; Bizzoni et al., 2020).

In this work, we seek to systematically investigate this difference by conducting a *large-scale fine-grained comparative* analysis on the distribution of translation divergences for HT and MT, all

[1] We use the term MT to mean the version of MT tested in this project's experiments: bilingual encoder-decoder Transformer-base networks with beam search decoding (see Models in Section 3).

[2] https://en.wikipedia.org/wiki/Control_(linguistics). They are coded as xcomp in Universal Dependencies (see https://universaldependencies.org/u/dep/all.html#xcomp-open-clausal-complement).

Transactions of the Association for Computational Linguistics, vol. 12, pp. 355–371, 2024. https://doi.org/10.1162/tacl_a_00645
Action Editor: Yue Zhang. Submission batch: 8/2023; Revision batch: 12/2023; Published 4/2024.
© 2024 Association for Computational Linguistics. Distributed under a CC-BY 4.0 license.

Readers are **cautioned** not to **place** undue reliance on …	Il est recommandé aux lecteurs de ne pas accorder…
PCO has **continued** to **assist** …	Le BCP a constamment épaulé …
These meetings **seek** to **strengthen** …	Ces réunions ont pour but de renforcer …
… weaknesses will **become** more **apparent** …	Les faiblesses apparaîtront…

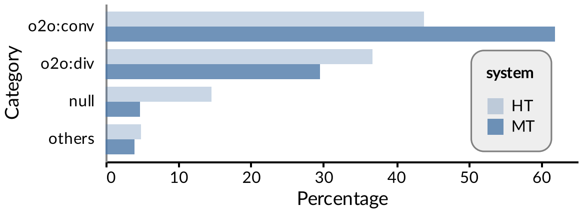

Figure 1: **Top table**: Examples of divergences in HT for En→Fr WMT15 training data (Bojar et al., 2015), with relevant fragments of the source/target shown in the first/second rows. The English control constructions are bolded including both the finite root verb and the controlled word, while the French phrases of interest are underlined. **Bottom figure**: Percentages of target patterns for HT and MT, with obligatory control finite verbs as the source pattern. `o2o:conv`: one-to-one convergent patterns where the target phrase uses a similar control construction to the source; `o2o:div`: one-to-one divergent patterns where the target differs structurally from the source; `null`: no target word is aligned; `others`: other less frequent patterns (e.g., one-to-many alignments). The percentages of all four categories sum up to 100%.

through the lens of morphosyntax. More specifically, we aim to answer the following research questions: 1) How are MT and HT quantitatively different in terms of morphosyntactic divergence? 2) How do we explain or understand this difference? 3) How do translation divergences in HT affect MT quality? In other words, do MT systems have more difficulty translating source sentences that exhibit divergences in HT?

Through extensive analyses based on three language pairs and two types of morphosyntactic divergence using the annotational framework of Universal Dependencies (Nivre et al., 2016), we make the following empirical observations:

1. MT is more *conservative* than HT, with less morphosyntactic diversity, more convergent patterns, and more one-to-one alignments.

2. MT is morphosyntactically less similar to HT for less frequent source patterns.

3. The distributional difference can be largely attributed to the use of beam search, which is biased towards convergent patterns. This bias is most amplified when the convergent target patterns appear around 50% of the time out of all translations of the same source pattern in the training data.

4. A majority of the most frequent divergent patterns are correlated with decreased MT performance. This correlation cannot be fully explained by the lower frequencies of the relevant divergences.

To the best of our knowledge, this is the first work to present the comparative perspective of HT vs MT in such fine granularity covering thousands of morphosyntactic constructions. In the remaining sections, we first briefly describe related work in Section 2. The experimental setup is described in detail in Section 3. We demonstrate the quantitative difference between MT and HT in Section 4, and seek to understand this discrepancy in Section 5. Lastly, we explore the correlation between the presence of divergences in HT with MT performance in Section 6 and make conclusions in Section 7.

2 Related Work

Translation Divergence Systematic and theoretical treatment of translation divergences started in the early 1990s, focusing on European languages (Dorr, 1992, 1993, 1994). Later work has expanded into more languages, and focused on the automatic detection of divergences (Gupta and Chatterjee, 2001, 2003; Sinha et al., 2005; Mishra and Mishra, 2009; Saboor and Khan, 2010) or their empirical distributions in human translations (Wong et al., 2017; Deng and Xue, 2017; Wein and Schneider, 2021). Relatedly, Carpuat et al. (2017), Vyas et al. (2018), and Briakou and Carpuat (2020) focused on identifying semantic divergences that manifest in translations not entirely semantically equivalent to the original sources.

The closest work to ours is from Nikolaev et al. (2020), who proposed to investigate fine-grained crosslingual morphosyntactic divergence based on Universal Dependencies. They augmented a subset of the Parallel Universal Dependencies (PUD) corpus (Zeman et al., 2017) with human-annotated word alignments for five language pairs and focused exclusively on content words. While our

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00645/2362224/tacl_a_00645.pdf by guest on 24 April 2024

work shares a similar conceptional and methodological foundation to theirs, our goal is to conduct a comparative analysis between HT and MT. In addition, we rely on a dependency parser and a word aligner (see Section 3 for more details) to reach a sufficiently large scale to enable the investigation of more fine-grained divergences.

Diverse Machine Translation MT systems tend to produce less diverse outputs in general (Gimpel et al., 2013; Ott et al., 2018), which is particularly harmful for back translation (Edunov et al., 2018; Soto et al., 2020; Burchell et al., 2022). To address this issue, various techniques have been proposed in the literature, including modified decoding algorithms (Li et al., 2016; Sun et al., 2020; Li et al., 2021), mixtures of experts (Shen et al., 2019), Bayesian models (Wu et al., 2020), additional codes (syntax or latent) (Shu et al., 2019; Lachaux et al., 2020), and training with simulated multi-reference corpora (Lin et al., 2022). In all aforementioned works, the emphasis is on the lack of diversity in MT outputs rather than comparing them systematically against HT. Notable exceptions include Roberts et al. (2020), who investigated the distributional differences between MT and HT in terms of n-grams, sentence length, punctuation, and copy rates. Marchisio et al. (2022) compared translations from supervised MT and unsupervised MT and noted their systematic style differences based on similarity and monotonicity in their POS sequences. In contrast, our work goes beyond surface features and focuses on fine-grained morphosyntactic divergences.

Algorithmic Bias Another closely related line of work studies algorithmic biases of current NLP systems, with particular emphasis on gender and racial biases (Bolukbasi et al., 2016; Caliskan et al., 2017; Zhao et al., 2017; Garg et al., 2018). Specifically for MT, researchers have focused on lexical diversity by comparing HT against post-editese (Toral, 2019) or MT outputs directly (Vanmassenhove et al., 2019); Bizzoni et al. (2020) have compared HT, MT, and simultaneous interpreting in terms of translationese using POS perplexity and dependency length. Most related to our work, Vanmassenhove et al. (2021) have conducted an extensive comparison between HT and MT based on a suite of lexical and morphological diversity metrics. While our study reaches a similar conclusion that MT is less diverse than HT,

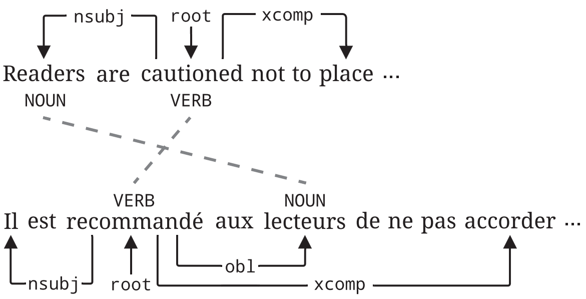
nsubj
root
xcomp
Readers are cautioned not to place ···
NOUN
VERB
VERB
NOUN
Il est recommandé aux lecteurs de ne pas accorder ···
nsubj
root
obl
xcomp

Figure 2: An illustration of the two types of morphosyntactic divergence. See Section 3 for details.

we explore morphosyntactic patterns on a more fine-grained level, and also reveal the bias of MT (and more specifically beam search) towards convergent structures.

3 Experimental Setup

Types of Morphosyntactic Divergence In this study, we experiment with two types of translation patterns based on the annotational scheme of Universal Dependencies:

(A) *Word-based*: POS tags for the aligned word pair. We additionally include their parent and child syntactic dependencies for more granularity. Order of the children dependencies is ignored.

(B) *Arc-based*: The source dependency arc, and the target path between the aligned words of the arc's head and tail. Directionality of the target dependencies is ignored. We additionally include the POS tags of both the head and the tail for more granularity.

These types are largely based on the proposal of Nikolaev et al. (2020), with modifications to accommodate more granularity. With either type, the translation pattern is a *convergence* if the source and the target sides have the same structure (word-based or arc-based), and otherwise a *divergence*. Notationally, we use tildes to connect the various parts of the pattern in a fixed order. For instance, for the control verb *''cautioned''* in Figure 2, its word-based divergence has `root~VERB~nsubj+xcomp` on the source side, where `VERB` corresponds to its POS tag, `root` its parent dependency, and `nsubj` and `xcomp` its two child dependencies. Similarly, we have `root~VERB~nsubj+obl+xcomp` on the

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00645/2362224/tacl_a_00645.pdf by guest on 24 April 2024

Language pair	Source	Target
En→De	17 055	15 040
En→Zh	14 816	19 471
En→Fr	18 321	12 212

Table 1: Number of distinct source or target patterns found in the analysis set (1M sentences from WMT).

target side. With regard to an arc-based divergence, for the source arc between the words *''cautioned''* and *''readers''*, we denote it as `VERB~nsubj~NOUN`, where `nsubj` is the dependency relation of the arc, and `VERB` and `NOUN` the POS tags of the head and the tail, respectively. Similarly, we denote the aligned target pattern as `VERB~obl~NOUN`.

Data We conduct experiments for three language pairs using WMT datasets (Bojar et al., 2015; Barrault et al., 2019): En→Zh (WMT19), En→Fr (WMT15), and En→De (WMT19). All training datasets are lightly filtered based on length, length ratio, and language ID, and deduplicated. For each language pair, one million sentences are held out from the training split to form an analysis subset. All analyses in our study are based on this subset to eliminate potential confounding effects from domain mismatch. Table 1 shows the number of distinct source or target patterns found in the analysis set for each language pair.

Models We train a bilingual Transformer base model (Vaswani et al., 2017) for each language pair using the `T5X` framework (Roberts et al., 2022). All models are trained with `Adafactor` optimizer (Shazeer and Stern, 2018) for 2M steps with 0.1 dropout rate, 1024 batch size, and 0.1 label smoothing. We use an inverse square root learning rate schedule with a base rate of 2.0. As summarized in Table 2 part (i), all models achieve similar BLEU scores[3] on the development set as reported in the literature with a comparable setup.

Annotations We rely on two automatic tools to conduct a large-scale analysis: a dependency parser and a word aligner. More specifically, the dependency parser is an implementation of Dozat and Manning, (2017) based on mBERT (Devlin et al., 2019). The neural word aligner is based on AMBER (Hu et al., 2021) and fine-tuned on human-annotated alignments. We follow Nikolaev et al. (2020) to keep the content words[4] and their dependencies and alignments only, and focus on one-to-one alignments unless otherwise noted.

(i) TRANSLATION

Target	Dev dataset	BLEU	Reported
Fr	newstest2014	39.9	38.1†
De	newstest2018	46.3	46.4‡
Zh	newstest2018	34.4	34.8††

(ii) DEPENDENCY PARSING

Language	Dataset	UPOS	UAS	LAS
En	EWT	95.56	89.55	96.67
Fr	GSD	97.77	93.20	90.90
De	GSD	94.80	87.87	83.25
Zh	GSD	94.58	86.41	80.70

(iii) WORD ALIGNMENT

Target	Precision	Recall	F1
Fr	85.6	81.7	83.6
Zh	85.5	81.9	83.7

† Vaswani et al. (2017)
‡ Ng et al. (2019)
††Bawden et al. (2019)

Table 2: Performance for (i) MT (ii) dependency parsing and (iii) word alignment. No human annotations for En→De are provided by Nikolaev et al. (2020).

As reported in Table 2 parts (ii) and (iii), we validate that both tools have high accuracy on public datasets: UD test sets for parsing and human-annotated PUD datasets (Nikolaev et al., 2020) for word alignment. The automatic annotations are available to the public here: `https://github.com/google-research-datasets/Crosslingual-Morphosyntactic-Divergence-dataset`.

[3] All reported BLEU scores for our models are obtained through `SacreBLEU` (Post, 2018).

[4] Content words are words with semantic content, used in various ''contentful'' positions such as subjects, objects, and adjectival modifiers. We identify content words by matching their parent dependencies against a manually selected set, as defined in footnote 10 of the original paper (Nikolaev et al., 2020). This criterion kept around 40%-50% of all the tokens for all three language pairs in our experiments. Please see Appendix B for a more detailed analysis.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00645/2362224/tacl_a_00645.pdf by guest on 24 April 2024

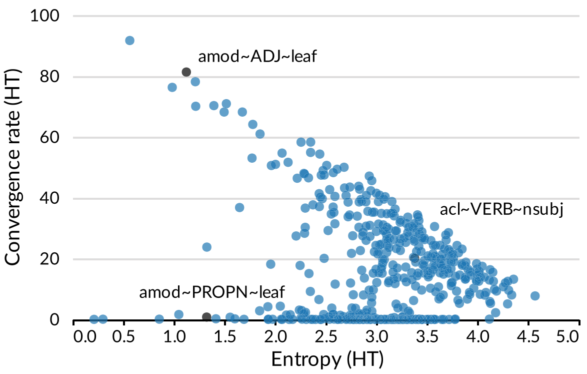

Figure 3: Plot of convergence rate vs entropy for the most frequent word-based source patterns in En→Fr human translations, three of which are highlighted in black: (1) `amod~ADJ~leaf` (high convergence rate, low entropy): the most common cases of adjectival modifiers; (2) `acl~VERB~nsubj` (low convergence rate, high entropy): object relative clauses without a relative pronoun, or subject relative clauses. The high entropy reflects a major difference between English and French, where the relative pronoun *que* is obligatory in French but not in English. (3) `amod~PROPN~leaf` (low convergence rate, low entropy): adjectives as part of a proper nouns. Adjectives in official institutions and titles are typically capitalized and annotated as `PROPN` in English (e.g., *Secretary General*) but lowercased and annotated as `ADJ` in French (e.g., *secrétaire général*).

4 Comparative Analysis of MT vs HT

We proceed to conduct a comparative analysis of MT vs HT based on the fine-grained morphosyntactic patterns defined in the previous section. For any given source pattern p according to the word-based or arc-based definition as detailed in the previous section, we study the distribution of its aligned target patterns, i.e., $\mathrm{Pr}_{\mathrm{HT}}(\cdot \mid p)$ and $\mathrm{Pr}_{\mathrm{MT}}(\cdot \mid p)$, along two major dimensions: diversity/uncertainty as measured by entropy of the target pattern, and convergence/divergence rate. Figure 3 shows that there is considerable variance in how the most frequent source patterns in HT are distributed along these two axes, and that each dimension captures a different property of the distribution.

Through analyses on both the aggregate level and the individual pattern level, we conclude that MT is more *conservative* than HT, with less morphosyntactic diversity, more convergent patterns, and more one-to-one alignments. We also observe that MT tends to be less similar to HT for the less frequent source patterns. The analyses in this section are based on the held-out subset consisting of one million sentence pairs. We refer readers to Appendix A for similar results on a subset that is further filtered using LaBSE crosslingual embeddings (Feng et al., 2022) with a remarkably similar trend, which we include to show that it does not change our conclusions when we test on data that has been filtered to improve its cross-lingual equivalence.

Target	Word-based			Arc-based		
	HT	MT	Δ%	HT	MT	Δ%
			(i) DIVERSITY			
Fr	2.23	1.84	−17.6	2.24	1.75	−22.2
De	2.23	1.90	−15.0	2.38	1.96	−17.8
Zh	2.95	2.77	−5.9	3.79	3.46	−8.6
			(ii) CONVERGENCE RATE			
Fr	37.9	44.7	18.1	46.3	53.2	15.0
De	45.8	51.8	13.7	51.8	57.8	11.7
Zh	21.2	22.6	6.9	23.4	25.2	7.4

Table 3: Aggregate diversity scores and convergence rates. The Δ% columns show the relative change in percentage from HT to MT.

4.1 MT is Less Morphosyntactically Diverse Than HT

Preliminaries We define diversity score as the conditional entropy of target patterns given source patterns, which reflects the *aggregate* level of uncertainty when translating a morphosyntactic pattern. More formally, let P and Q denote the categorical random variables for source patterns and their aligned target patterns, respectively. The aggregate diversity score is defined as

$$H(Q \mid P) = \sum_{p} \mathrm{Pr}(p) \cdot H(Q \mid P = p), \quad (1)$$

where p is any specific source pattern that occurs in the corpus.

In addition, for any given source pattern p, we define a *source pattern-specific* diversity score as the entropy of the target patterns aligned to that source pattern p. This score corresponds to the term $H(Q \mid P = p)$ in Equation (1).

Aggregate Finding As summarized in Table 3 part (i), MT is less morphosyntactically diverse than HT in aggregate, across three language pairs and two types of divergence. The relative reduction in diversity for MT compared to HT ranges from 5.9% for En→Zh (2.77 vs 2.95 with

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00645/2362224/tacl_a_00645.pdf by guest on 24 April 2024

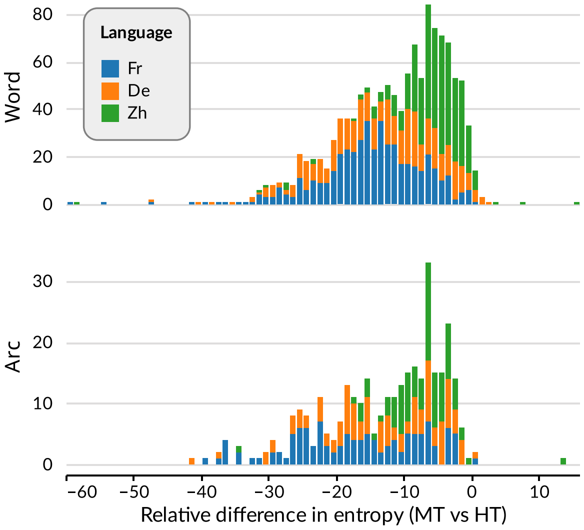

Figure 4: Stacked histogram of the relative differences in source pattern-specific diversity score.

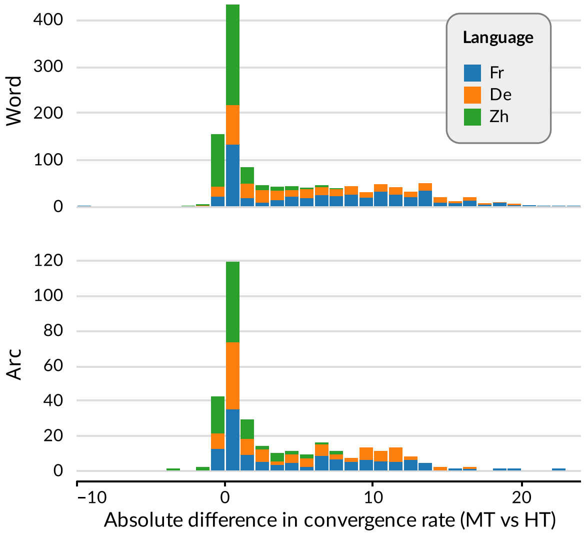

Figure 5: Stacked histogram of the absolute differences in source pattern-specific convergence rate.

word-based patterns) to 22.2% for En→Fr (1.75 vs 2.24 with arc-based patterns). Interestingly, En→Zh has noticeably higher diversity scores than En→Fr and En→De but lower overall reduction. This may be attributed to the larger linguistic difference between Chinese and English.[5]

Finding by Source Pattern On the level of individual source patterns, we observe that the reduction of diversity among their aligned target patterns is *across-the-board* but *unevenly distributed*. Figure 4 plots a stacked histogram of the relative differences in diversity score (MT vs HT) for the most frequent source patterns with at least 1000 occurrences, and it shows that the vast majority of them see a drop of diversity (i.e., negative difference). This reduction varies from pattern to pattern, ranging from 0% to 60%.

4.2 MT is More Convergent Than HT

Preliminaries We tally divergences and convergences according to the two types detailed in Section 3. We then define the convergence or divergence rate as the percentage of convergent or divergent patterns out of all translation patterns. Similar to diversity, we can compute convergence/ divergence rates for both the entire corpus in aggregate and individual source patterns. For the latter case, we tally all the aligned target patterns for a specific source pattern and calculate the rates accordingly.

Aggregate Finding As summarized in Table 3 part (ii), we observe a consistent increase of convergence rate for all three language pairs and two types of divergence. This increase is most pronounced for En→Fr and En→De, whereas En→Zh has a less noticeable although still consistent increase and starts with a much lower convergence rate for HT: the highest rate for En→Zh is 23.4%, whereas En→De can reach 57.8%.

Finding by Source Pattern On a more granular level, we again notice a consistent increase of convergent patterns for MT among the top source patterns (Figure 5). For the vast majority of top source patterns, MT has produced more convergent translations than HT, and this discrepancy ranges from a negligible amount (~0%) for most patterns to more than 20%. This discrepancy is distributed differently for the three languages: En→Fr and En→De have seen more patterns with increased convergence rate while En→Zh has most patterns barely changed and clustered around 0%. As we later show in Figure 9, this trend is unsurprising given the much lower convergence rates for En→Zh in general.

4.3 MT Looks Less Like HT For Less Frequent Patterns

Preliminaries Both diversity score and convergence rate are properties of translations produced by one system, *either* MT *or* HT. To directly measure the distributional difference between

[5] Note that, however, our setup is not entirely comparable across language pairs since the data is not multi-way parallel.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00645/2362224/tacl_a_00645.pdf by guest on 24 April 2024

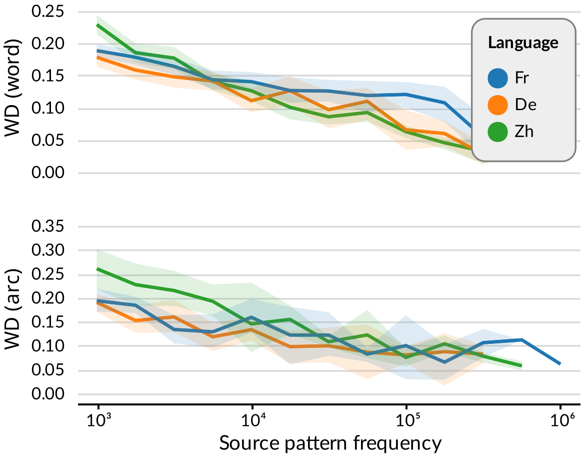

Figure 6: Wasserstein distance with a unit cost matrix between $\mathrm{Pr}_{\mathrm{MT}}(\cdot \mid p)$ and $\mathrm{Pr}_{\mathrm{HT}}(\cdot \mid p)$ for any given source pattern p. Patterns are binned by frequency on a log scale, and both the means (lines) and the 95% confidence intervals (shaded areas) are shown. The plot shows a negative correlation between WD and the source pattern frequency.

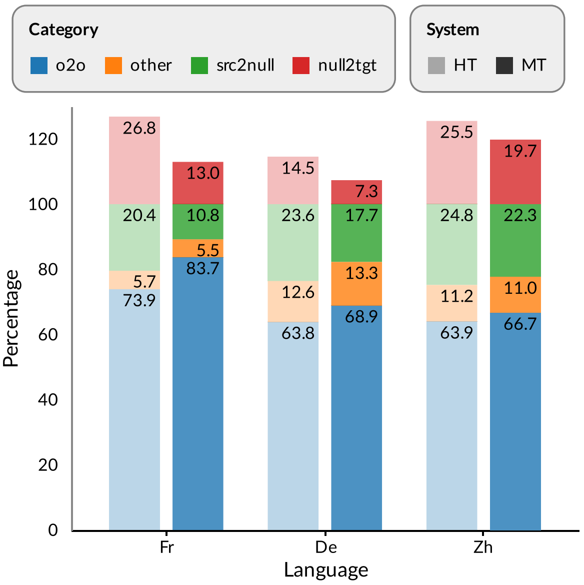

Figure 7: Distribution for all types of alignments. Percentages are defined relative to the total number of source content words. `o2o`: one-to-one; `src2null`: deletions; `null2tgt`: insertions; `other`: other types such as one-to-many.

MT and HT, we resort to Wasserstein distance (WD) between the two conditional distributions[6] $\mathrm{Pr}_{\mathrm{MT}}(\cdot \mid p)$ and $\mathrm{Pr}_{\mathrm{HT}}(\cdot \mid p)$ using a unit cost matrix.[7] This metric can be intuitively interpreted as the minimal amount of probability mass that has to be moved from $\mathrm{Pr}_{\mathrm{MT}}(\cdot \mid p)$ to match $\mathrm{Pr}_{\mathrm{HT}}(\cdot \mid p)$, with an upper bound of 1 (i.e., sum of all probability mass).[8]

Finding As Figure 6 shows, there is a negative correlation between WD and the source pattern frequency: MT matches HT more closely for the more frequent source patterns while having difficulty in reproducing the HT distribution for the less frequent ones. This trend persists for all tested settings, and points to a potential weakness of MT systems when it comes to learning the distributions of the less common structures.

4.4 Beyond One-to-one Alignments

Preliminaries One-to-one alignments constitute a majority of all detected alignments, but they fail to account for translation patterns involving deletions and insertions. To investigate the quantitative differences between HT and MT on those special patterns, we conduct additional analyses on the distribution of all categories of alignments based on the word-based definition. Besides deletions (`src2null`) and insertions (`null2tgt`), the remaining alignments are collapsed into the `other` category (e.g., one-to-many mapping).

Finding Figure 7 summarizes the distribution of all alignment categories,[9] which demonstrates a significant and consistent difference between HT and MT. More specifically, MT produces fewer deletions (green), fewer insertions (red), and more one-to-one translations (blue). En→Fr again exhibits the biggest discrepancy with 9.6% less deletions (10.8% vs 20.4%) and 14.8% less insertions (13.0% vs 26.8%), both around 50% relative reduction. This trend contributes to the overall conservative nature of MT predictions, favoring one-to-one alignments at the expense of the other (more uncertain) categories.

[6] Recall that we treat both source patterns and target patterns as categorical random variables where every unique source or target pattern is treated as a distinct value that the random variables can take.

[7] In which diagonal/off-diagonal entries are 0/1.

[8] We note that other metrics such as KL-divergence can also be used to measure distributional difference, but we eventually chose WD for its interpretability.

[9] The percentages are computed in terms of source words. By definition, `src2null`, `o2o`, and `other` add up to 100%. Since `null2tgt` alignments do not have aligned source words, their percentages indicate how many target content words are inserted for each content word on the source side.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00645/2362224/tacl_a_00645.pdf by guest on 24 April 2024

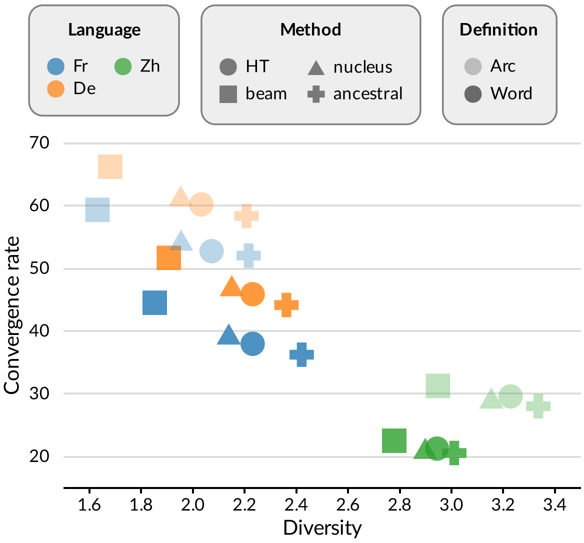

Figure 8: Convergence rates (y-axis) and diversity scores (x-axis) on the aggregate level for translations through different sampling methods and HT. Sampling methods consistently obtain higher diversity score and lower convergence rate than beam search.

5 Understanding the Discrepancy

In this section, we seek to understand the source of discrepancy between HT and MT as demonstrated in the previous section. By investigating different decoding algorithms, we attribute this discrepancy to the use of beam search, echoing the thesis laid out by previous work (Edunov et al., 2018; Eikema and Aziz, 2020). More specifically in our experiments, we show that beam search is biased towards less diverse and more convergent translations, even when the learned model distribution actually resembles HT. This bias is most prominent when the convergent patterns appear around 50% of the time in training data. Moreover, frequencies of convergent patterns in MT are increased even when they are uncommon in HT, suggesting perhaps a more inherent structural bias in current MT architectures.

Decoding Algorithms Besides beam search, we additionally obtain translations through two sampling methods. More specifically, to make fair comparison with single-reference HT, we sample one translation using ancestral sampling or nucleus sampling with $p = 0.95$ (Holtzman et al., 2020) for each source sentence.

Beam Search is Biased Against Diversity and Divergence As Figure 8 illustrates, for all three language pairs and two types of divergence, translations obtained through beam search are significantly less diverse and more convergent compared to either sampling method. Indeed, ancestral sampling consistently produces higher diversity scores and lower convergence rates than even HT.[10] Since ancestral sampling is an unbiased estimator of the model distribution, this suggests that on the aggregate distribution level, the model learns to be as least as morphosyntactically diverse and divergent as HT.

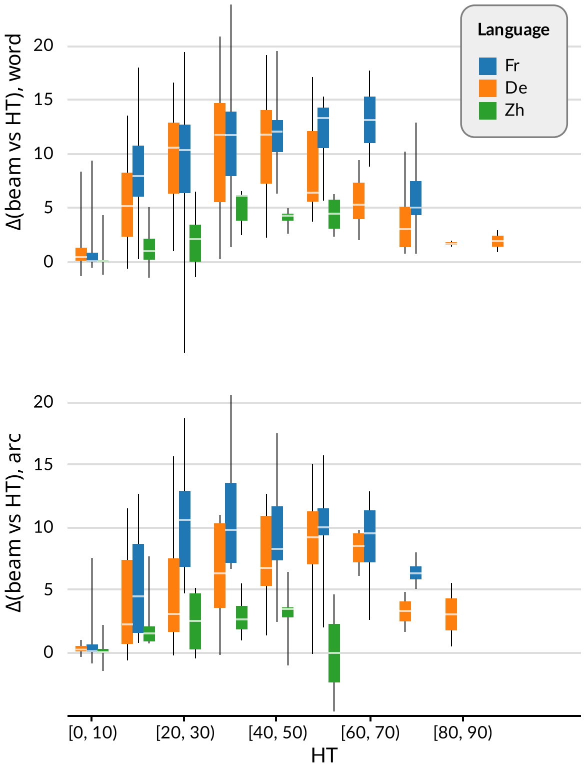

Figure 9: Plot of difference in convergence rate (beam search vs HT) against convergence rate of HT. The plot is similar when comparing beam search against ancestral sampling.

A further breakdown of most frequent[11] individual source patterns reveals that beam search's bias towards convergent translations is a function of the relative frequencies of the convergent patterns. As Figure 9 demonstrates, the increase of convergence rate for beam search compared to ancestral sampling seems to be quadratically correlated with the convergence rate for ancestral sampling: Peak difference is reached at around

[10] We hypothesize that the increased diversity score and the higher divergence rate for ancestral sampling compared to HT are attributable to the use of label smoothing during training. Roberts et al. (2020) have also demonstrated the effect of label smoothing on various diversity diagnostics.

[11] With at least 1000 occurrences.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00645/2362224/tacl_a_00645.pdf by guest on 24 April 2024

40%–50%. This suggests that beam search favors the convergent pattern more when the pattern appears around 50% of the time in training data. This could be because the model has seen the pattern enough to assign it substantial probability mass, but there is still enough uncertainty that humans will frequently choose other patterns.

We additionally note that convergence rate increases for the overwhelming majority of the most frequent source patterns *even when the convergence patterns are uncommon in HT*. This strongly suggests an inherent bias of beam search towards convergent patterns,[12] and that this bias is distinct from the typical bias amplification due to data exposure, e.g., ‘‘cooking’’ is more likely to co-occur with ‘‘women’’ than ‘‘men’’ in the training data (Zhao et al., 2017). We suspect that this bias towards convergence is due to the architectural design of MT systems, but we leave the subject matter for future work.

6 Divergence and MT Quality

In our final analysis, we investigate how the presence of morphosyntactic divergence in HT might affect MT quality. In contrast to the previous sections analyzing conditional distributions given a source pattern, we focus instead on individual divergences/convergences. The potential connection between divergence and MT quality is motivated by second-language acquisition research that describes language inference from their first languages (i.e., negative transfer) as one source of difficulty for learners (Gass et al., 2020), which can happen when the two languages diverge structurally. Do MT systems have similar problems with divergences?

Preliminaries To answer this question, we conduct an analysis on the presence (or absence) of a word-based morphosyntactic divergence in HT and the corresponding MT quality as measured by BLEU (Papineni et al., 2002) and BLEURT (Sellam et al., 2020). The basic idea is to construct two contrastive groups of source sentences (called the experiment group and the control group) and compare the MT performance on each group. The HT references of the experiment group contain a given divergent pattern, corresponding to sentences that are perhaps more challenging to translate, whereas those of the control group do not.

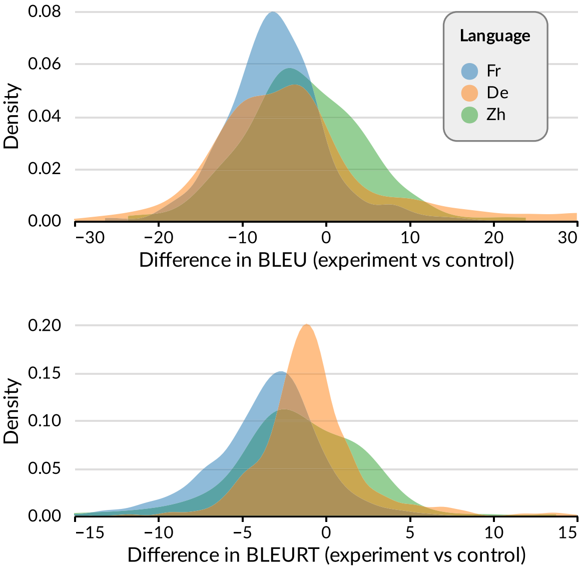

Figure 10: Kernel density estimation for the difference in BLEU or BLEURT scores between the experiment group and the control group. Negative values indicate that the experiment group has lower score than the control group.

More specifically, for a given divergence with source pattern p and target pattern q ($p \neq q$), its control group consists of source sentences for which HT translates every source p into target p (i.e., a convergent pattern), and its experiment group consists of source sentences for which HT translates every source p into p except for one that is translated into q. For an simplified example, if we are interested the divergence that translates nouns into verbs, the corresponding control group contains source sentences for which HT translates every noun into a noun, whereas its experiment group contains source sentences for which exactly one noun is translated into a verb and the rest of nouns into nouns.

We then collect the MT outputs for both groups and compute the differences in BLEU and BLEURT. This procedure is repeated for every divergence pattern for which both groups have at least 100 sentences.

Findings We treat each difference in BLEU or BLEURT as one data point and plot their estimated probability density function. As illustrated in Figure 10, divergences are more often associated with significantly lower BLEU scores (i.e., negative differences), with a fairly large amount of

[12] We do not observe a similar trend when comparing ancestral sampling against HT.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00645/2362224/tacl_a_00645.pdf by guest on 24 April 2024

Target	Metric	Pearson	Kendall τ
Zh	BLEURT	−0.072 0.32	0.030 0.54
	BLEU	−0.080 0.27	−0.026 0.59
Fr	BLEURT	0.319 1.4e-16	0.240 1.0e-19
	BLEU	0.206 1.6e-7	0.161 1.2e-9
De	BLEURT	0.289 1.4e-11	0.253 3.6e-18
	BLEU	0.159 2.5e-4	0.171 4.4e-9

Table 4: Correlation between the difference in BLEURT score and ratio of frequencies (i.e., the number of training examples with divergences over that with convergences). p-values are displayed in gray.

variance. Trends for BLEURT scores are similar, but with En→De showing less drastic differences compared to BLEU.[13] On the other hand, a substantial number of divergent patterns have either virtual no change or an increase of BLEU or BLEURT scores. This suggests that being a divergence pattern in itself is not associated with decreased MT performance.

What could explain this variance? Why are some divergent patterns associated with worse MT performance while others aren't? One obvious hypothesis is that these patterns are seen less frequently during training. However, a closer inspection seems to suggest that frequency of divergent patterns alone is not an adequate predictor. More specifically, we use the absolute or relative frequency[14] of the divergent pattern, with or without taking a log of the number, and correlate it with BLEU or BLEURT scores. Even with the best option (log of relative frequency) presented in Table 4, there is only weak correlation (Pearson or Kendall τ) for En→Fr and En→De, and no correlation for En→Zh. It is unclear what aspects of divergent patterns make them more difficult to translate, or whether they are merely co-occurring with those elements that are the true cause of difficulty. We leave it to future work to investigate the underlying cause.

7 Conclusion

We conduct a large-scale fine-grained comparative investigation between HT and MT outputs, through the lens of morphosyntactic divergence. Based on extensive analyses on three language pairs, we demonstrate that MT is less morphosyntactic diverse and more convergent than HT. We further attribute to this difference to the use of beam search that biases MT outputs towards less diverse and less divergent patterns. Finally, we show that the presence of divergent patterns in HT has overall an adverse effect on MT quality.

In future work, we are interested in applying the same analysis to large language model (LLM)-based MT systems. Recent studies have noted that LLM-based systems tend to produce less literal translations, compared to the traditional encoder-decoder models (Vilar et al., 2023; Raunak et al., 2023). It would be interested to see whether and to what extent the LLM translations might differ from those produced by traditional models when viewed from a morphological lens.

Acknowledgments

We thank Julia Kreutzer, Eleftheria Briakou, Markus Freitag, and Macduff Hughes for providing useful feedback that helped shape this paper. We would also like to thank the anonymous reviewers and the action editor for their constructive comments.

References

Roee Aharoni, Moshe Koppel, and Yoav Goldberg. 2014. Automatic detection of machine translated text and translation quality estimation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 289–295, Baltimore, Maryland. Association for Computational Linguistics. https://doi.org/10.3115/v1/P14-2048

Guangsheng Bao, Zhiyang Teng, and Yue Zhang. 2023. Target-side augmentation for document-level machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10725–10742, Toronto, Canada. Association for Computational Linguistics.

[13] We also note that n-gram overlap-based metrics such as BLEU are more likely to penalize diverse translations (Freitag et al., 2019).

[14] Here, relative frequency is defined as the ratio of the number of training examples with the divergence over that with the convergence. It is a way to counterbalance the fact that some extremely common source patterns will have a lot more frequent divergences.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00645/2362224/tacl_a_00645.pdf by guest on 24 April 2024

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. https://doi.org/10.18653/v1/W19-5301

Rachel Bawden, Nikolay Bogoychev, Ulrich Germann, Roman Grundkiewicz, Faheem Kirefu, Antonio Valerio Miceli Barone, and Alexandra Birch. 2019. The University of Edinburgh's submissions to the WMT19 news translation task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 103–115, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-5304

Yuri Bizzoni, Tom S. Juzek, Cristina España-Bonet, Koel Dutta Chowdhury, Josef van Genabith, and Elke Teich. 2020. How human is machine translationese? Comparing human and machine translations of text and speech. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 280–290, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.iwslt-1.34

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal. Association for Computational Linguistics. https://doi.org/10.18653/v1/W15-3001

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems*, 29.

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. https://doi.org/10.18653/v1/2020.emnlp-main.121

Laurie Burchell, Alexandra Birch, and Kenneth Heafield. 2022. Exploring diversity in back translation for low-resource machine translation. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 67–79, Hybrid. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.deeplo-1.8

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186. https://doi.org/10.1126/science.aal4230, PubMed: 28408601

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. https://doi.org/10.18653/v1/W17-3209

D. Deng and Nianwen Xue. 2017. Translation divergences in Chinese–English machine translation: An empirical investigation. *Computational Linguistics*, 43:521–565. https://doi.org/10.1162/COLI_a_00292

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186,

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00645/2362224/tacl_a_00645.pdf by guest on 24 April 2024

Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

B. Dorr. 1994. Machine translation divergences: A formal description and proposed solution. *Computational Linguistics*, 20:597–633.

Bonnie J. Dorr. 1992. The use of lexical semantics in interlingual machine translation. *Machine Translation*, 7(3):135–193. https://doi.org/10.1007/BF00402510

Bonnie J. Dorr. 1993. Interlingual machine translation a parameterized approach. *Artificial Intelligence*, 63(1–2):429–492. https://doi.org/10.1016/0004-3702(93)90023-5

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. In *International Conference on Learning Representations*.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1045

Bryan Eikema and Wilker Aziz. 2020. Is MAP decoding all you need? The inadequacy of the mode in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4506–4520, Barcelona, Spain (Online). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.398

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.62

Markus Freitag, Isaac Caswell, and Scott Roy. 2019. APE at scale and its implications on MT evaluation biases. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 34–44, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-5204

Markus Freitag, David Grangier, and Isaac Caswell. 2020. BLEU might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.5

Markus Freitag, David Vilar, David Grangier, Colin Cherry, and George Foster. 2022. A natural diet: Towards improving naturalness of machine translation output. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3340–3353, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-acl.263

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Y. Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115:E3635–E3644. https://doi.org/10.1073/pnas.1720347115, PubMed: 29615513

Susan M. Gass, Jennifer Behney, and Luke Plonsky. 2020. *Second language acquisition: An introductory course*, Routledge. https://doi.org/10.4324/9781315181752

Martin Gellerstam. 1986. Translationese in Swedish novels translated from English. *Translation studies in Scandinavia*, 1:88–95.

Kevin Gimpel, Dhruv Batra, Chris Dyer, and Gregory Shakhnarovich. 2013. A systematic exploration of diversity in machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1100–1111.

Deepa Gupta and Niladri Chatterjee. 2001. Study of divergence for example based English-Hindi machine translation. *STRANS-2001, IIT Kanpur*, pages 43–51.

Deepa Gupta and Niladri Chatterjee. 2003. Identification of divergence for English to Hindi ebmt. In *MTSUMMIT*.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00645/2362224/tacl_a_00645.pdf by guest on 24 April 2024

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Junjie Hu, Melvin Johnson, Orhan Firat, Aditya Siddhant, and Graham Neubig. 2021. Explicit alignment objectives for multilingual bidirectional encoders. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3633–3643, Online. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2021.naacl-main.284`

Huda Khayrallah, Brian Thompson, Matt Post, and Philipp Koehn. 2020. Simulated multiple reference training improves low-resource machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 82–89, Online. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2020.emnlp-main.7`

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1318–1326, Portland, Oregon, USA. Association for Computational Linguistics.

David Kurokawa, Cyril Goutte, and Pierre Isabelle. 2009. Automatic detection of translated text and its impact on machine translation. In *Proceedings of Machine Translation Summit XII: Papers*. Ottawa, Canada.

Marie-Anne Lachaux, Armand Joulin, and Guillaume Lample. 2020. Target conditioning for one-to-many generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2853–2862, Online. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2020.findings-emnlp.256`

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. 2012. Adapting translation models to translationese improves SMT. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 255–265, Avignon, France. Association for Computational Linguistics.

Jicheng Li, Pengzhi Gao, Xuanfu Wu, Yang Feng, Zhongjun He, Hua Wu, and Haifeng Wang. 2021. Mixup decoding for diverse machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 312–320.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. A simple, fast diverse decoding algorithm for neural generation. *ArXiv*, abs/1611.08562.

Huan Lin, Baosong Yang, Liang Yao, Dayiheng Liu, Haibo Zhang, Jun Xie, Min Zhang, and Jinsong Su. 2022. Bridging the gap between training and inference: Multi-candidate optimization for diverse neural machine translation. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2622–2632, Seattle, United States. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2022.findings-naacl.200`

Kelly Marchisio, Markus Freitag, and David Grangier. 2022. On systematic style differences between unsupervised and supervised MT and an application for high-resource machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2214–2225, Seattle, United States. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2022.naacl-main.161`

Vimal Mishra and Ravi Bhushan Mishra. 2009. Divergence patterns between English and Sanskrit machine translation. *INFOCOMP Journal of Computer Science*, 8:62–71.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy. Association for Computational Linguistics. `https://doi.org/10.18653/v1/W19-5333`

Dmitry Nikolaev, Ofir Arviv, Taelin Karidi, Neta Kenneth, Veronika Mitnik, Lilja Saeboe, and Omri Abend. 2020. Fine-grained analysis of cross-linguistic syntactic divergences. In

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00645/2362224/tacl_a_00645.pdf by guest on 24 April 2024

ACL. https://doi.org/10.18653/v1/2020.acl-main.109

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. In *International Conference on Machine Learning*, pages 3956–3965. PMLR.

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. https://doi.org/10.3115/1073083.1073135

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, Belgium, Brussels. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6319

Vikas Raunak, Arul Menezes, Matt Post, and Hany Hassan. 2023. Do GPTs produce less literal translations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1041–1050, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-short.90

Parker Riley, Isaac Caswell, Markus Freitag, and David Grangier. 2020. Translationese as a language in ''multilingual'' NMT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7737–7746, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.691

Adam Roberts, Hyung Won Chung, Anselm Levskaya, Gaurav Mishra, James Bradbury, Daniel Andor, Sharan Narang, Brian Lester, Colin Gaffney, Afroz Mohiuddin, Curtis Hawthorne, Aitor Lewkowycz, Alex Salcianu, Marc van Zee, Jacob Austin, Sebastian Goodman, Livio Baldini Soares, Haitang Hu, Sasha Tsvyashchenko, Aakanksha Chowdhery, Jasmijn Bastings, Jannis Bulian, Xavier Garcia, Jianmo Ni, Andrew Chen, Kathleen Kenealy, Jonathan H. Clark, Stephan Lee, Dan Garrette, James Lee-Thorp, Colin Raffel, Noam Shazeer, Marvin Ritter, Maarten Bosma, Alexandre Passos, Jeremy Maitin-Shepard, Noah Fiedel, Mark Omernick, Brennan Saeta, Ryan Sepassi, Alexander Spiridonov, Joshua Newlan, and Andrea Gesmundo. 2022. Scaling up models and data with t5x and seqio. *arXiv preprint arXiv:2203.17189*. https://doi.org/10.48550/arXiv.2203.17189

Nicholas Roberts, Davis Liang, Graham Neubig, and Zachary Lipton. 2020. Decoding and diversity in machine translation. In *NeurIPS 2020 Workshop on Resistance AI*.

Abdus Saboor and Mohammad Abid Khan. 2010. Lexical-semantic divergence in Urdu-to-English example based machine translation. *2010 6th International Conference on Emerging Technologies (ICET)*, pages 316–320. https://doi.org/10.1109/ICET.2010.5638469

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.704

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR.

Tianxiao Shen, Myle Ott, Michael Auli, and Marc'Aurelio Ranzato. 2019. Mixture models for diverse machine translation: Tricks of the trade. In *International Conference on Machine Learning*, pages 5719–5728. PMLR.

Raphael Shu, Hideki Nakayama, and Kyunghyun Cho. 2019. Generating diverse translations

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00645/2362224/tacl_a_00645.pdf by guest on 24 April 2024

with sentence codes. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1823–1827, Florence, Italy. Association for Computational Linguistics. `https://doi.org/10.18653/v1/P19-1177`

K. Sinha, R. Mahesh, and Anil Thakur. 2005. Translation divergence in English-Hindi mt. In *EAMT*.

Xabier Soto, Dimitar Shterionov, Alberto Poncelas, and Andy Way. 2020. Selecting backtranslated data from multiple sources for improved neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3898–3908, Online. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2020.acl-main.359`

Zewei Sun, Shujian Huang, Hao-Ran Wei, Xin-yu Dai, and Jiajun Chen. 2020. Generating diverse translation by manipulating multi-head attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8976–8983. `https://doi.org/10.1609/aaai.v34i05.6429`

Antonio Toral. 2019. Post-editese: An exacerbated translationese. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 273–281, Dublin, Ireland. European Association for Machine Translation.

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2203–2213, Online. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2021.eacl-main.188`

Eva Vanmassenhove, Dimitar Shterionov, and Andy Way. 2019. Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 222–232, Dublin, Ireland. European Association for Machine Translation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting PaLM for translation: Assessing strategies and performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427, Toronto, Canada. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2023.acl-long.859`

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. `https://doi.org/10.18653/v1/N18-1136`

Chao Wang, Michael Collins, and Philipp Koehn. 2007. Chinese syntactic reordering for statistical machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 737–745, Prague, Czech Republic. Association for Computational Linguistics.

Shira Wein and Nathan Schneider. 2021. Classifying divergences in cross-lingual amr pairs. *Proceedings of The Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*. `https://doi.org/10.18653/v1/2021.law-1.6`

Tak-sum Wong, Kim Gerdes, Herman Leung, and John S. Y. Lee. 2017. Quantitative comparative syntax on the cantonese-mandarin parallel dependency treebank. In *Proceedings of the fourth international conference on Dependency Linguistics (Depling 2017)*, pages 266–275.

Xuanfu Wu, Yang Feng, and Chenze Shao. 2020. Generating diverse translation from model distribution with dropout. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*,

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00645/2362224/tacl_a_00645.pdf by guest on 24 April 2024

pages 1088–1097. https://doi.org/10.18653/v1/2020.emnlp-main.82

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/K17-3001

Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545, Austin, Texas. Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1160

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1323

Chunting Zhou, Xuezhe Ma, Junjie Hu, and Graham Neubig. 2019. Handling syntactic divergence in low-resource machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1388–1394, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1143

Target	Word-based			Arc-based		
	HT	MT	Δ%	HT	MT	Δ%
			(i) Diversity			
Fr	2.22	1.85	−16.8	2.25	1.77	−21.6
De	2.24	1.95	−12.9	2.39	2.01	−16.0
Zh	2.92	2.76	−5.5	3.78	3.47	−8.3
			(ii) Convergence Rate			
Fr	37.4	43.8	17.1	45.4	52.0	14.5
De	44.4	49.4	11.3	49.6	54.7	10.3
Zh	20.9	22.0	5.4	23.0	24.5	6.6

Table 5: Aggregate diversity scores and convergence rates for the LaBSE-filtered subset. The Δ% columns show the relative change in percentage from HT to MT.

A Analysis Subset Filtered Using LaBSE Embeddings

The main results of the paper are obtained on a held-out subset of the WMT data. To remove some of the noise due to the automatic extraction pipeline that produced the WMT data, we resort to LaBSE embeddings (Feng et al., 2022) to further filter the original held-out subset. More specifically, we use the LaBSE model to derive the crosslingual embeddings for the source and the target of any sentence pair, and sort all pairs based on the cosine distance between the source and the target embeddings. The top half (i.e., lowest distance) is kept for analysis, resulting in 500K sentence pairs for each language pair.

Table 5 summarizes the aggregate diversity scores and convergence rates. The relative changes are slightly smaller than those in Table 3, but the overall trend is remarkably similar: For both word-based and arc-based divergences, MT produces less diverse outputs with more convergent patterns.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00645/2362224/tacl_a_00645.pdf by guest on 24 April 2024

B Percentage of Content Words and Their Alignments

Table 6 summarizes the percentage of content words and their alignments based on the held-out analysis subset. We only keep the alignments for the main results if both the source token and the target token are content words. The statistics show that around 40%–50% of the tokens (either on the source or the target side) are considered content words, and a similar percentage of alignments pass our criterion.

Lang	Source content words	Target content words	Alignments
Fr	13.7M / 28.7M = 47.9%	14.8M / 33.7M = 44.0%	11.3M / 27.2M = 41.7%
De	10.2M / 21.3M = 48.1%	8.8M / 20.1M = 43.8%	7.9M / 18.5M = 42.9%
Zh	12.6M / 26.2M = 48.2%	12.8M / 24.5M = 52.4%	10.1M / 22.3M = 45.3%

Table 6: Percentage of content words and their alignments for the held-out analysis subset.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00645/2362224/tacl_a_00645.pdf by guest on 24 April 2024