Exploring Geometric Representational Disparities Between Multilingual and Bilingual Translation Models

Neha Verma[1], **Kenton Murray**[1,2], **Kevin Duh**[1,2]
[1]Center for Language and Speech Processing
[2]Human Language Technology Center of Excellence
{nverma7, kenton}@jhu.edu, kevinduh@cs.jhu.edu

Abstract

Multilingual machine translation has proven immensely useful for both parameter efficiency and overall performance across many language pairs via complete multilingual parameter sharing. However, some language pairs in multilingual models can see worse performance than in bilingual models, especially in the one-to-many translation setting. Motivated by their empirical differences, we examine the geometric differences in representations from bilingual models versus those from one-to-many multilingual models. Specifically, we compute the isotropy of these representations using intrinsic dimensionality and IsoScore, in order to measure how the representations utilize the dimensions in their underlying vector space. Using the same evaluation data in both models, we find that for a given language pair, its multilingual model decoder representations are consistently less isotropic and occupy fewer dimensions than comparable bilingual model decoder representations. Additionally, we show that much of the anisotropy in multilingual decoder representations can be attributed to modeling language-specific information, therefore limiting remaining representational capacity.

Keywords: machine translation, multilinguality, isotropy

1. Introduction

Recent advances in multilingual machine translation have led to better parameter efficiency and language transfer by simultaneously modeling multiple language pairs (Firat et al., 2016; Ha et al., 2016). Some work has even proven the viability of performing zero-shot translation between language pairs for which there may be very little to no bitext (Johnson et al., 2017; Zhang et al., 2020). However, multilingual translation systems with complete parameter sharing can suffer from interference, or reduced performance for some language pairs versus a comparable bilingual baseline (Aharoni et al., 2019; Arivazhagan et al., 2019).

Previous work has hypothesized that limited modeling capacity is a major contributor to reduced performance in multilingual models (Aharoni et al., 2019; Zhu et al., 2021; Conneau et al., 2020). Some prior work shows this bottleneck phenomenon empirically by evaluating bilingual versus multilingual model performance across different model and data sizes (Zhang et al., 2020; Shaham et al., 2023). Besides capacity, the direction of translation can also dictate how much interference occurs in multilingual models; one-to-many translation systems suffer more from interference compared to multilingual translation model types (Wang et al., 2018; Arivazhagan et al., 2019; Fernandes et al., 2023). Therefore, in this work, we focus on one-to-many multilingual translation systems.

Despite trends pointing towards performance differences between bilingual and multilingual translation systems, especially in those with a multilingual decoder, it still unclear *how* these systems may be performing differently. To this end, we systematically compare the behavior of one-to-many translation models to their bilingual counterparts. Specifically, we examine the geometry of model

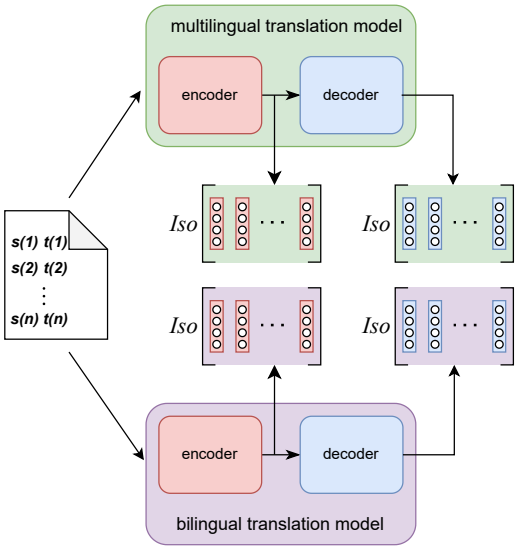


Figure 1: Schematic of our hidden space utilization comparisons. We extract final layer representations from both a bilingual model and a multilingual model on the same set of parallel sentences. We compute the isotropy of these representations (*Iso*), and compare the two models.

LREC-COLING 2024, pages 6909–6921
20-25 May, 2024.

representations from both types of models and compare them directly. We ask the following: (1) How does the ambient space utilization of model representations differ between bilingual models and one-to-many models? (2) If space utilization differs, what might be driving these differences?

We measure space utilization using IsoScore and intrinsic dimensionality (ID), which are two metrics that determine how uniformly a point cloud utilizes the dimensions of its underlying vector space, or its isotropy (Fukunaga and Olsen, 1971; Rudman et al., 2022).

We compute the isotropy of representations on the same set of sentence pairs across model types so that their scores are directly comparable, and summarize our method in Figure 1. We observe the following in our comparison:

- Across different data resource levels and different source-target language pairs, the isotropy of one-to-many decoder representations for a given source-target pair is reduced as contrasted with decoder representations in a comparable bilingual model.
- Source-side representation capacity improves slightly in one-to-many models over bilingual models. However, the extent of this encoder capacity improvement is smaller than the extent of the decoder capacity reduction.
- With further analysis, we find that reduced space utilization in multilingual decoder representations seems driven by language-specific information occupying much of the available representation space. Single language decoders, however, do not have to distinguish this language-specific information.

While most previous work has observed empirical differences between bilingual and multilingual models and some of its potential causes, our work characterizes the differences between bilingual and multilingual models in terms of their internal model representations. Our results could inform alternative approaches on current multilingual modeling design, especially in models that cover multiple target languages.

2. Analysis of Model Representations

2.1. Model Representation Space Utilization

In this work, we investigate the difference between our model types via the geometry of final and intermediate layer representations. Specifically, we are interested in how well these representations utilize the dimensions of the vector space they lie in. If a set of representations has very high variance across a few dimensions, and little to no variance spread across the remaining dimensions, this set is said to have low isotropy, or anisotropy.

Because a one-to-many model has to accommodate multiple languages in its decoder, we hypothesize that our multilingual models have less representational capacity than bilingual models for a given language pair. Therefore, we turn to examining the isotropy of representations produced from both a bilingual model and a multilingual model on a set of parallel sentences. Since our experiments keep the hidden dimension fixed across all models, and the representations are computed from the same data, these two sets of hidden vectors are directly comparable. In this setting, if one set of representations uses more ambient vector space compared to the other set, we can say that the first set is using more of its representational capacity.

2.2. Computing Isotropy

In computing the space utilization of model representations, we first compute the sequence of hidden states across tokens. For a given source target pair $(\mathbf{x}, \mathbf{y})$, a forward pass through the encoder gives $h_{\text{enc}}(\mathbf{x}) = (v_1, v_2, \ldots, v_{|\mathbf{x}|})$, and through the decoder gives $h_{\text{dec}}(\mathbf{x}, \mathbf{y}) = (w_1, w_2, \ldots, w_{|\mathbf{y}|})$

We compute the isotropy of these model representations at a sentence level. For converting encoder and decoder hidden state sequences into single vectors, we mean pool all non-padding tokens over the token dimension (Li et al., 2020; Kudugunta et al., 2019). Isotropy, formally, is a measure of how uniformly the variance of a dataset is spread across its vector dimensions.

The isotropy metrics used in this work are intrinsic dimensionality (ID) as computed by the PCA Fukunaga-Olsen algorithm (Fukunaga and Olsen, 1971) and IsoScore (Rudman et al., 2022). PCA Fukunaga-Olsen is a straightforward method to estimate the ID of a dataset based on a linear PCA decomposition of the data. This method is simple, robust to large samples, and handles high dimensionality, which is important for our hidden vector setting (Bac et al., 2021). The PCA-FO ID algorithm computes the following, for threshold $D_e \in [0, 1)$ and original dimensionality n:

1. Compute PCA of the dataset $X \subseteq \mathbb{R}^n$: $\text{cov}(X) = V\Lambda V^T$
2. Compute normalized eigenvalues $\lambda_i = \lambda_i / \lambda_1$
3. return $\text{count}(\lambda_i > D_e)/n$

In this work, we use $D_e = 0.05$.

IsoScore is a similar metric that uses the diagonal of the covariance matrix of PCA-transformed points in order to measure how many dimensions are used and how uniformly the dimensions are used. Previous works on representation isotropy have used other metrics, like average cosine similarity or partition scores (Mu and Viswanath, 2018;

Ethayarajh, 2019), but Rudman et al. (2022) found that these methods do not stand up to thorough validity testing, like mean agnosticism or rotational invariance.

More formally, IsoScore computes the following:

1. Reorient dataset $X \subseteq \mathbb{R}^n$ with PCA: X^{PCA}
2. Compute the diagonal covariance matrix of $X^{\text{PCA}} \in \mathbb{R}^n$, denoted as Σ_D. [1]
3. Normalize the variance diagonal to be: $\hat{\Sigma}_D := \sqrt{n}\frac{\Sigma_D}{\|\Sigma_D\|}$
4. Compute the distance between the covariance diagonal and the identity matrix, which reflects ideal isotropy: $\delta(X) := \frac{\|\hat{\Sigma}_D - \mathbf{1}\|}{\sqrt{2(n-\sqrt{n})}}$
5. Use $\delta(X)$ to compute the percentage of dimensions isotropically utilized. $\phi(X) = (n - \delta(X)^2(n - \sqrt{n}))^2/n^2$

The final range of $\phi(X)$ is linearly rescaled to span the interval $[0, 1]$, resulting in the IsoScore. More details and motivation behind the metric can be found in the original paper (Rudman et al., 2022). We detail an example of point clouds and their respective IsoScores and IDs in Figure 2.

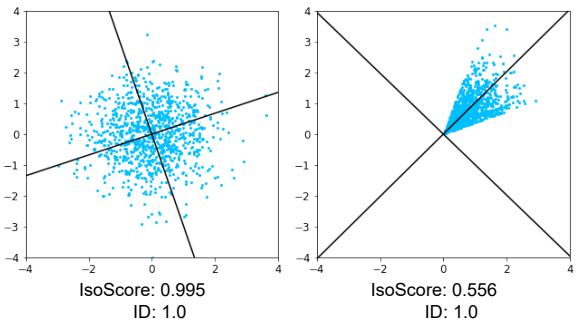


Figure 2: Depictions of 2D point clouds, their principal components, and their computed IsoScores and IDs. The left point cloud has high IsoScore due to even variance spread across principal components, but the right has lower IsoScore due to uneven variance spread. Both clouds have an ID of 1.0 as ID is less sensitive to variance spread.

The main difference between IsoScore and ID is that IsoScore accounts for evenness of variance spread among the dimensions, whereas ID only computes a variance threshold. In our Figure 2 example, the ID of these point clouds is both 1.0, meaning that all dimensions are utilized, but the IsoScore captures more fine-grained detail about *how* the dimensions are being used.

In our work, we compute IsoScores and the ID of several sets of model representations for comparison. We begin with a multilingual model that translates language pairs $s \rightarrow \{t_1, t_2, ..., t_n\}$, a bilingual model that translates only $s \rightarrow t_k$, and a set of sentences $\{s(i), t_k(i)\}$. For both models, we compute the isotropy using one of our metrics, of $X_{\text{enc}} = \{h_{\text{enc}}(s(i)) : \forall i\}$ and $X_{\text{dec}} = \{h_{\text{dec}}(s(i), t_k(i)) : \forall i\}$. These values are labelled $Iso(X^{\text{multi}}_{\text{enc}}(s, t_k)), Iso(X^{\text{multi}}_{\text{dec}}(s, t_k))$ and $Iso(X^{\text{bi}}_{\text{enc}}(s, t_k)), Iso(X^{\text{bi}}_{\text{dec}}(s, t_k))$. Additionally, to observe the overall behavior of our multilingual models, we compute the isotropy of hidden states from all covered language pairs, resulting in $Iso(X^{\text{multi}}_{\text{enc}}(s, \bigcup_j t_j)), Iso(X^{\text{multi}}_{\text{dec}}(s, \bigcup_j t_j))$.

3. Experimental Setup

3.1. Trilingual Models

In order to control for the effects of language similarity, we experiment with trilingual models that translate from English to two languages, keeping one of the target languages fixed (Xin et al., 2022; Fernandes et al., 2023; Shaham et al., 2023). Specifically, we look at trilingual models with English as a source language, and 2 target languages. We use Russian (ru) as a fixed target, and vary the 3 other target languages: Chinese (zh), German (de), and Ukrainian (uk). These three additional languages have differing degrees of language similarity with Russian; Ukrainian and Russian share a close language family and script, German and Russian share a distant language family and do not share a script, and Russian and Chinese do not share a language family or script. In summary, we experiment with en-{ru,zh}, en-{ru,de}, and en-{ru,uk} models.

3.2. Datasets

Our main experiments use data from previous WMT competitions on general translation. We use training and development data from the 2022 WMT General Machine Translation task, and describe our WMT data preparation pipeline in Appendix A. For validation on our en-{ru,uk} multilingual models, we subsample from the WMT22 Russian development set in order to match the size of the Ukrainian set for evenness. However, we perform our analysis on the whole development set.

We additionally use bitext from the Multitarget TED talks, which allow us to investigate the role of multiparallel data in MT representations (Duh, 2018). We filter the Multitarget TED talk training sets to be strictly multiparallel, like their dev and test sets, and henceforth refer to the dataset as multiparallel TED talks. To measure the effect of data availability as well as multiparallelism, we subsample our WMT data to match the size of the Multiparallel TED talks. This way, our small WMT

[1] PCA guarantees no off-diagonal covariance elements.

		WMT-large		WMT-small		Multiparallel TED	
dataset	lang	train	dev	train	dev	train	dev
en-{ru,de}	en-ru	98.2M	2993	149k	2993	149k	1958
	en-de	98.2M	2203	149k	2203	149k	1958
en-{ru,uk}	en-ru	31.5M	2993	67k	2993	67k	1958
	en-uk	31.5M	997	67k	997	67k	1958
en-{ru,zh}	en-ru	41.1M	2993	161k	2993	161k	1958
	en-zh	41.1M	3418	161k	3418	161k	1958

Table 1: Total sentences in each bitext used in our work. We train trilingual models that translate from English into two other languages. We force the WMT-small training split to be the same size as Multiparallel TED for comparability.

set and TED talks can help us study multiparallelism, and our small WMT set and large WMT set can help show the effect of scale on representational capacity. Statistics on our datasets are in Table 1.

3.3. Training Details

For our bilingual and multilingual translation models, we use the Transformer architecture as implemented by `fairseq` (Vaswani et al., 2017; Ott et al., 2019). For TED and WMT-small experiments, we use the `transformer_iwslt_de_en` configuration, and for WMT-large experiments, we use a transformer base configuration. We use weight tying between decoder input and output embeddings (Press and Wolf, 2017; Inan et al., 2016). For multilingual models, we incorporate target language id tokens prepended to the source sentence (Wicks and Duh, 2022). For all bilingual experiments, we use a joint source-target SentencePiece vocabulary of 16K tokens (Sennrich et al., 2016; Kudo and Richardson, 2018). For all multilingual experiments, we use a joint source-target vocabulary of 32K tokens. These vocabularies have high token overlap, where each multilingual vocabulary contains at least 93% of the bilingual vocabulary across all languages and datasets. This overlap leads to very similar tokenizations of the sentences in our comparisons.

For TED and WMT-small experiments, we select the best model checkpoint using validation on BLEU after training for up to 80 epochs. For WMT large experiments, we use average validation loss for selection after training up to 240k updates with a batch size of 32k tokens. All outputs are computed using a beam size of 5. We report BLEU scores on our dev sets computed with `sacrebleu` (Papineni et al., 2002; Post, 2018).

4. Results

4.1. Multilingual decoder capacity reduction

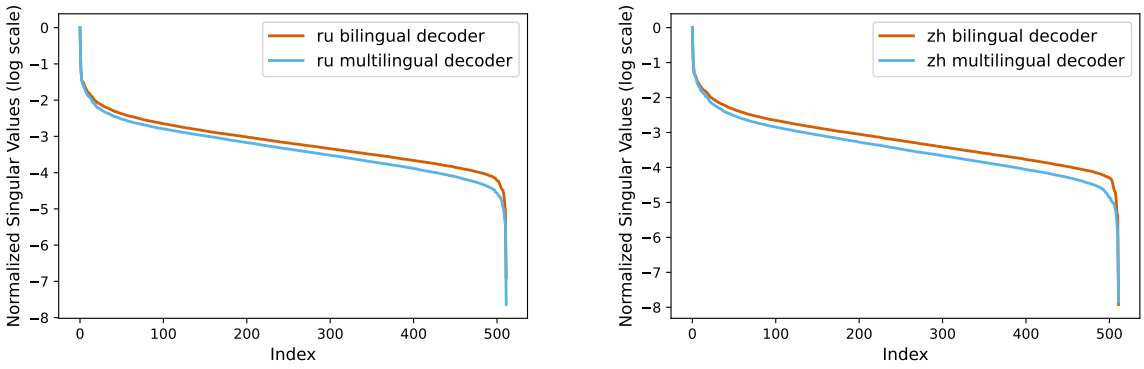
We find that across our language pair settings and across our dataset sizes, representations from bilingual model decoders are more isotropic than multilingual model decoder representations. In Table 2, we see that for all trilingual settings, and for both WMT-small and WMT-large, bilingual decoder isotropy scores are larger than those of multilingual models for the same language pair. For example, in the WMT-large en-{ru,zh} dataset, the IsoScore of multilingual decoder representations (iso-dec) for Russian is 0.164 and Chinese is 0.106, but in their respective bilingual models, these values jump to 0.192 for Russian and 0.142 for Chinese.

Additionally, we plot the singular values from the singular value decomposition (SVD) of the hidden states of one of our multilingual model decoders and its corresponding two bilingual model decoders in Figure 3. We see that the spectra of the bilingual model decoder hidden states are more balanced than those of from the multilingual model, as they do not drop off in value as quickly as the multilingual singular values. This additionally demonstrates that the bilingual decoder hidden states have better distribution of variance across its dimensions.

Because these representations are computed from the same set of source-target sentences, and only the model types differ, the multilinguality of the one-to-many decoder must be contributing its reduced representational capacity for the source-target pair. In this case, modeling language-specific information in each decoder pass may be occupying much of the multilingual decoder state space. We explore this hypothesis further in Section 4.5.

dataset	langs	type	WMT-large BLEU	WMT-large iso-enc	WMT-large ID-enc	WMT-large iso-dec	WMT-large ID-dec	WMT-small BLEU	WMT-small iso-enc	WMT-small ID-enc	WMT-small iso-dec	WMT-small ID-dec
en-{ru,zh}	en-ru	multi	23.7	**0.082**	**0.070**	0.164	0.145	20.0	**0.104**	**0.088**	0.208	0.250
		bi	23.7	0.075	0.057	**0.192**	**0.164**	19.1	0.074	0.057	**0.236**	**0.285**
	en-zh	multi	34.5	**0.051**	**0.045**	0.106	0.092	28.0	**0.070**	**0.057**	0.136	0.148
		bi	36.0	0.023	0.020	**0.142**	**0.199**	27.7	0.032	0.023	**0.185**	**0.201**
	both	multi	-	0.066	0.057	0.065	0.043	-	0.085	0.070	0.076	0.047
en-{ru,de}	en-ru	multi	23.8	**0.081**	**0.068**	0.161	0.145	19.1	**0.112**	**0.105**	0.230	0.293
		bi	23.8	0.074	0.064	**0.189**	**0.164**	18.9	0.109	0.104	**0.242**	**0.295**
	en-de	multi	26.4	**0.046**	**0.039**	0.141	0.164	18.0	**0.076**	**0.063**	0.171	0.227
		bi	28.0	0.037	0.029	**0.191**	**0.236**	15.3	0.056	0.037	**0.233**	**0.311**
	both	multi	-	0.049	0.037	0.037	0.021	-	0.070	0.049	0.070	0.039
en-{ru,uk}	en-ru	multi	23.7	**0.053**	**0.053**	0.161	0.168	17.3	0.118	0.115	0.242	**0.307**
		bi	23.8	0.031	0.029	**0.184**	**0.182**	15.7	**0.123**	**0.129**	**0.246**	0.305
	en-uk	multi	26.6	**0.086**	**0.080**	0.139	0.160	16.2	0.148	0.199	0.191	0.238
		bi	27.8	0.074	0.072	**0.195**	**0.238**	12.7	**0.160**	**0.213**	**0.221**	**0.281**
	both	multi	-	0.086	0.086	0.078	0.051	-	0.127	0.145	0.172	0.162

Table 2: Main isotropy results for models trained on WMT data. We report BLEU scores of each model on the appropriate validation set, and IsoScores and intrinsic dimensionalities (ID) for both encoder and decoder sentence representations. We report scores for both language pairs, and in both types of models, bilingual (bi) and multilingual (multi). We bold the higher IsoScore/ID value between each multilingual/bilingual comparison. We additionally report the IsoScore of multilingual model spaces on the entire development set, not separating by language pair (both).



(a) WMT-large, en-ru

(b) WMT-large, en-zh

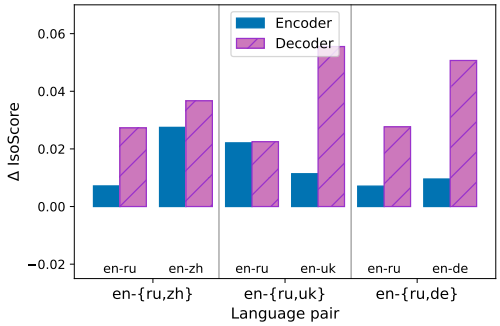
Figure 3: Semi-log plots of normalized singular values from SVD of bilingual decoder hidden states and multilingual decoder hidden states for the WMT-large en-{ru,zh} model. The spectra of bilingual decoder hidden states are better balanced than those of multilingual decoder hidden states. We use a semi-log scale for visibility.

4.2. Multilingual encoder capacity increase

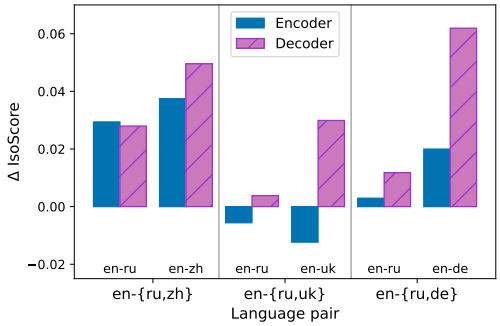
In encoder representation spaces, we see an opposite effect, although less pronounced. In both en-{ru,zh} and en-{ru,de} models, across small and large data availability, multilingual encoders tend to have greater isotropy among representations than bilingual model encoders. However, the one exception is the WMT-small en-{ru-uk} model. Results comparing this increase in encoder capacity to the decrease in decoder capacity in multilingual models, compared to their bilingual counterparts, are summarized in Figure 4.

Comparing multilingual encoder isotropy separated by language versus the isotropy of the whole multilingual encoder space (Table 2), we see that the difference in scores is not very large. This could indicate that the multilingual encoder

space is benefiting from sharing across the English sources from both language pairs in our multilingual dataset.



(a) WMT-large



(b) WMT-small

Figure 4: ΔIsoScore values comparing the extent of the observed encoder isotropy increase ($Iso(X^{\text{multi}}_{\text{enc}}) - Iso(X^{\text{bi}}_{\text{enc}})$) to the extent of the observed isotropy decrease ($Iso(X^{\text{bi}}_{\text{dec}}) - Iso(X^{\text{multi}}_{\text{dec}})$) in our multilingual models, compared to their bilingual counterparts. Overall, the extent of the decoder isotropy decrease is larger than that of the encoder increase.

4.3. Effects of training scale

In comparing IsoScore results on WMT-small vs WMT-large setups, we see that in a larger scale, there is consistently less space utilization in both multilingual and bilingual models. This occurs consistently in the decoder space, and in almost all settings in the encoder space. Both models have the same hidden dimension $d = 512$, and differ only in their feed-forward dimension and attention heads. Even among the overall multilingual isotropy scores (setting labeled 'both' in Table 2), WMT-large representations have smaller isotropy values than WMT-small representations in almost all language settings.

The observed increase in anisotropy with larger training scale is closely related to the representation degeneration problem reported in previous literature (Gao et al., 2019). This phenomenon describes a tendency towards anisotropy of the final softmax layer W in natural language generation models, due to a frequency bias affecting output token embedding updates. With more training updates, this frequency bias causes output token embeddings to become more anisotropic. In our case, we see a similar degeneration with final hidden states, which are closely related to the softmax layer given the output distribution computation $y = \text{softmax}(h^T W)$ where h is our final hidden vector.

In terms of performance, we note that only the WMT-large BLEU scores see a reduction or no improvement in the multilingual case; it is known that measurable interference does not generally occur much at a smaller data scale (Shaham et al., 2023).

4.4. Multi-way parallelism

We report results on the Multiparallel TED Talks in Table 3. In this setting, we find that our results on increased isotropy of multilingual source-side representations still holds in a majority of cases, even though the source-side sentences are identical across our two language pairs in the trilingual model. This is a strong indication that in one-to-many models, source-side representations benefit from a shared source embedding space, and do not separate much based on target language.

On the other hand, our results on decreased decoder capacity do not hold in all language settings in our multiparallel model. An isotropy increase occurs over bilingual models to a small extent for our en-{ru,de} model, and a larger one for our en-{ru,uk} model, where the target languages share a script. However, the isotropy of our entire decoder multilingual space is still relatively low. This indicates that although there is still separation in the decoder space by language, each language's representation cluster in the decoder space is still more locally isotropic than its bilingual counterpart.

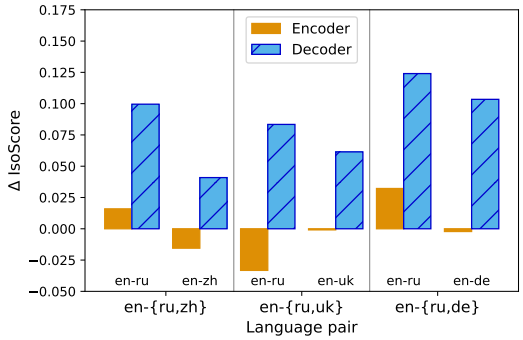
We test our TED model on our WMT test sets for direct comparability to our other models. Full results can be found in Appendix B. We see that results are mostly consistent for multilingual encoder isotropy improvement. For multilingual decoder isotropy, we see similar results with respect to language relatedness — bilingual decoder representations are more anisotropic than their multilingual counterparts for en-{ru,zh}, similar for en-{ru,de}, and the opposite for en-{ru,uk}, where the target languages are most related.

4.5. Decoder language separation

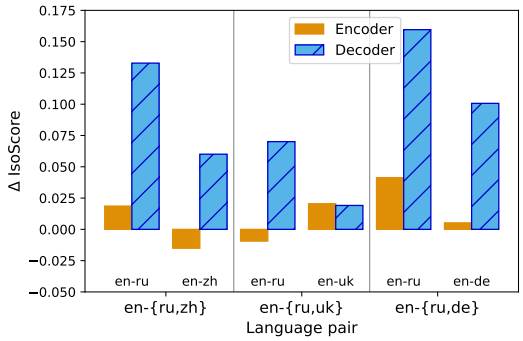
Across all three language settings, and in all of our data settings, we see that the isotropy of the overall multilingual decoder hidden space is much lower than either of the specific language portions of the

langs	type	Multiparallel TED BLEU	iso-enc	ID-enc	iso-dec	ID-dec
en-ru	multi	16.0	**0.135**	**0.133**	0.253	0.313
	bi	15.5	0.130	0.119	**0.284**	**0.348**
en-zh	multi	19.3	0.122	0.113	0.244	0.305
	bi	18.8	**0.125**	**0.125**	**0.277**	**0.338**
	both	-	0.138	0.137	0.104	0.063
en-ru	multi	15.8	**0.108**	**0.094**	**0.261**	**0.326**
	bi	15.3	0.097	0.098	0.250	0.309
en-de	multi	26.1	**0.104**	**0.088**	**0.258**	**0.320**
	bi	25.2	0.073	0.066	0.247	0.287
	both	-	0.108	0.094	0.116	0.072
en-ru	multi	13.5	**0.127**	0.139	**0.248**	**0.305**
	bi	12.2	0.124	**0.145**	0.222	0.260
en-uk	multi	16.8	**0.128**	0.141	**0.244**	**0.299**
	bi	15.6	0.124	**0.152**	0.201	0.238
	both	-	0.130	0.143	0.173	0.168

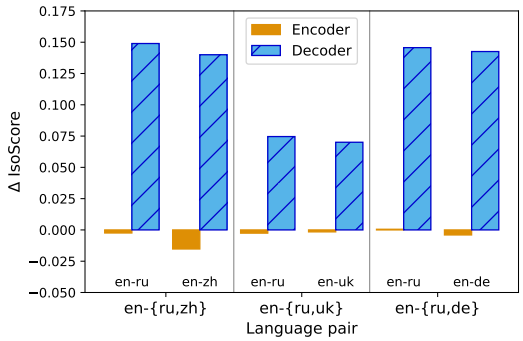
Table 3: Isotropy results on the encoder and decoder sentence representations from our Multiparallel TED model, tested on the Multiparallel TED development set.



(a) WMT-large



(b) WMT-small



(c) Multiparallel TED

Figure 5: ΔIsoScore values between language-specific multilingual representations separated by language and overall multilingual representations, for both the encoder and decoder ($Iso(X^{\text{multi}}(s, t_k) - Iso(X^{\text{multi}}(s, \cup_j t_j))$). Large ΔIsoScores between language-specific multilingual reps. and overall multilingual reps. indicate heavy encoding of language specificity in the decoder space.

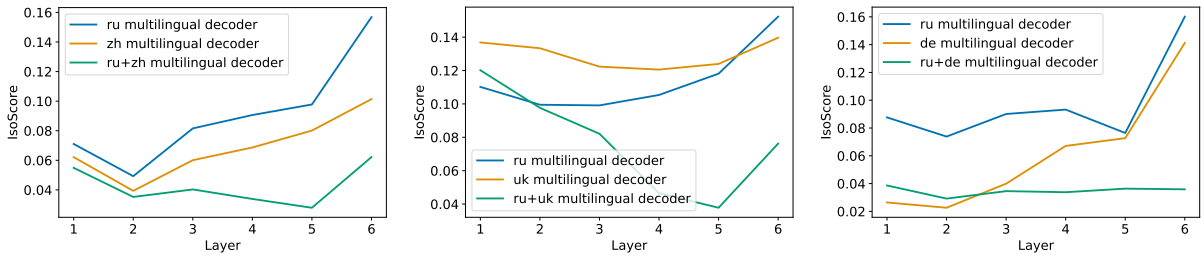
multilingual space. What this suggests, according to our metrics, is that there are some dimensions whose variance is heavily dictated by language information. When separating out these representations by language, the variance is reduced. This, however, is not the case when considering encoder language separation. We summarize this phenomenon in Figure 5.

In our multiparallel setting, tested on both our TED and WMT datasets, we see that this difference is smallest for en-{ru,uk}. We hypothesize that this difference is due to vocabulary sharing. Because Russian and Ukrainian share a script and subword units, shared output embedding vocabulary items would lead to closer hidden states. Their close typological relatedness could be contributing to their decoders state closeness as well. However, since Russian and German or Russian and Chinese share very few vocabulary units, their hidden states are further in the multilingual decoder space, as also seen in Figure 5.

4.5.1. Layerwise decoder behavior

We further investigate our claim that multilingual decoders use significant representational capacity to model language-specific information by observing how isotropy changes in multilingual decoder states across decoder layers. We show layerwise isotropy results for multilingual decoder states in Figure 6. We obtain hidden states according method described in Section 2.2, but instead at each layer boundary.

We find that throughout decoder layers, the overall isotropy of the entire set of decoder hidden states remains constant or decreases. However, for language-specific decoder states, we see that isotropy increases throughout the layers. Together, this implies that throughout the decoder layers, representations become more language specific. This suggests that earlier layers in the decoder benefit from some sharing, whereas later layers handle



(a) en-{ru,zh} (b) en-{ru,uk} (c) en-{ru,de}

Figure 6: Layerwise IsoScores on our WMT-large models. The divergence between the overall decoder isotropy and language-specific isotropy shows that hidden states become more language-specific throughout the decoder.

greater language specificity.

In summary, these results seem to suggest that decoders in multilingual translation models seem to separate out languages among the dimensions available in their hidden states. This finding could motivate the design and use of multilingual architectures that do not use complete sharing in their decoder parameters. Some prior work has already examined this approach (Sachan and Neubig, 2018; Kong et al., 2021; NLLB Team et al., 2022).

5. Related Work

5.1. Multilingual model capacity

Prior work has also examined the bottleneck phenomenon in multilingual machine translation. Much of this work observes the phenomenon empirically, and proposes methods to try to alleviate the parity. Sachan and Neubig (2018) also focus on one-to-many translation models, and propose partial sharing between language decoders in order to reduce the observed interference during full sharing. Tan et al. (2019) propose a knowledge distillation method to reduce the parity between bilingual and multilingual translation models by using bilingual models as multiple teachers and the multilingual model as a student. Other methods propose using a mix of language-specific and language-agnostic parameters, (Lin et al., 2021) and even automatically learning where to and where not to share across language pairs (Zhang et al., 2021). Wang et al. (2021) approach interference from a gradient viewpoint, and find that in En→ Any models, gradients become less similar in decoders, and hypothesize that this is due to the difference in decoder label spaces.

Kudugunta et al. (2019), like us, also investigate hidden representations to understand sharing in multilingual translation models. However, they focus on an in-house many-to-many translation model, and focus on representational similarities between languages, rather than representational capacity for language pairs.

Shaham et al. (2023) take an empirical approach to understanding interference in multilingual translation models, by investigate how scale and multilingual dataset ratios affect performance. They propose to both scale up models and adjust temperature sampling to reduce interference for simple models. However, this approach is largely empirical, and does not account for smaller scales and balanced datasets.

5.2. Isotropy of Representations

Recently, studies analyzing the geometry of Transformer representations have shown that they do not uniformly occupy many of the dimensions of the underlying space in which they lie. Ethayarajh (2019) show that many pretrained language models are anisotropic, where any two representations have very high cosine similarity. In addition to proposing a new metric, Rudman et al. (2022) also find that in their revised analysis, representations from language models use even fewer dimensions than previously reported. In the translation setting, Gao et al. (2019) show that embeddings from generation models, including MT models, tend to degenerate into an anisotropic distribution due to frequency bias. Yu et al. (2022) find a similar degeneration in generation models, and propose a gradient gating method that helps reduce the frequency bias causing embedding isotropy. They report improved MT results when controlling for anisotropy.

6. Conclusion

While previous work has empirically demonstrated performance differences in multilingual and bilingual models, in this work, we systematically compare the geometry of model representations in bilin-

gual and multilingual translation models in order to determine what might drive these differences. Using one-to-many models which are most prone to interference, we experiment with varying data sizes and source-target combinations.

We find for a given language pair, there is a consistent reduction in representational capacity in multilingual decoders versus comparable bilingual decoders. We additionally find a small increase in representational capacity for multilingual encoder spaces given the one-to-many task. Representational capacity decreases in a larger model and data paradigm, and results on multiparallel data show a strong improvement in multilingual encoder representational capacity and some improvement in multilingual decoder representational capacity. Finally, we find that reduced capacity in multilingual decoders can be attributed to language information occupying a significant portion of the available representation space.

7. Limitations

Our models cover at most 3 language families for the sake of controlled analysis when modern multilingual translation models cover many more. We think it is worthwhile to analyze models with larger coverage as future work. We focus on one-to-many models as they tend to fall behind other multilingual model types (Sachan and Neubig, 2018; Wang et al., 2018; Shaham et al., 2023). However, many-to-many models still have multilingual decoders but may have different behavior given their multilingual encoder state space.

Additionally, our conclusions focus on encoder-decoder models, but there is growing interest in decoder-only translation models whose isotropic behavior may differ.

Finally, our work focuses only on the characterization of representational capacity differences between model types, and not on the improvement of representational capacity of one-to-many models. However, we hope this work provides insight into the development of future modeling techniques for models with multilingual decoders.

8. Bibliographical References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. 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 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, et al. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*.

Mikel Artetxe and Holger Schwenk. 2019. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy. Association for Computational Linguistics.

Mikko Aulamo, Sami Virpioja, and Jörg Tiedemann. 2020. OpusFilter: A configurable parallel corpus filtering toolbox. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 150–156, Online. Association for Computational Linguistics.

Jonathan Bac, Evgeny M Mirkes, Alexander N Gorban, Ivan Tyukin, and Andrei Zinovyev. 2021. Scikit-dimension: a python package for intrinsic dimension estimation. *Entropy*, 23(10):1368.

Kyunghyun Cho. 2016. Noisy parallel approximate decoding for conditional recurrent language model. *ArXiv*, abs/1605.03835.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Kevin Duh. 2018. The multitarget ted talks task. http://www.cs.jhu.edu/~kevinduh/a/multitarget-tedtalks/.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. 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 55–65, Hong Kong, China. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22(1).

Patrick Fernandes, Behrooz Ghorbani, Xavier Garcia, Markus Freitag, and Orhan Firat. 2023. Scaling laws for multilingual neural machine translation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10053–10071. PMLR.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875, San Diego, California. Association for Computational Linguistics.

K. Fukunaga and D.R. Olsen. 1971. An algorithm for finding intrinsic dimensionality of data. *IEEE Transactions on Computers*, C-20(2):176–183.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tieyan Liu. 2019. Representation degeneration problem in training natural language generation models. In *International Conference on Learning Representations*.

Thanh-Le Ha, Jan Niehues, and Alex Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. In *Proceedings of the 13th International Conference on Spoken Language Translation*, Seattle, Washington D.C. International Workshop on Spoken Language Translation.

Hakan Inan, Khashayar Khosravi, and Richard Socher. 2016. Tying word vectors and word classifiers: A loss framework for language modeling. In *International Conference on Learning Representations*.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Xiang Kong, Adithya Renduchintala, James Cross, Yuqing Tang, Jiatao Gu, and Xian Li. 2021. Multilingual neural machine translation with deep encoder and multiple shallow decoders. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1613–1624, Online. 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.

Sneha Kudugunta, Ankur Bapna, Isaac Caswell, and Orhan Firat. 2019. Investigating multilingual NMT representations at scale. 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 1565–1575, Hong Kong, China. Association for Computational Linguistics.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130, Online. Association for Computational Linguistics.

Zehui Lin, Liwei Wu, Mingxuan Wang, and Lei Li. 2021. Learning language specific sub-network for multilingual machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 293–305, Online. Association for Computational Linguistics.

Jiaqi Mu and Pramod Viswanath. 2018. All-but-the-top: Simple and effective post-processing

for word representations. In *6th International Conference on Learning Representations, ICLR 2018*.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

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

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

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

Ofir Press and Lior Wolf. 2017. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 157–163, Valencia, Spain. Association for Computational Linguistics.

William Rudman, Nate Gillman, Taylor Rayne, and Carsten Eickhoff. 2022. IsoScore: Measuring the uniformity of embedding space utilization. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3325–3339, Dublin, Ireland. Association for Computational Linguistics.

Devendra Sachan and Graham Neubig. 2018. Parameter sharing methods for multilingual self-attentional translation models. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 261–271, Brussels, Belgium. Association for Computational Linguistics.

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

Uri Shaham, Maha Elbayad, Vedanuj Goswami, Omer Levy, and Shruti Bhosale. 2023. Causes and cures for interference in multilingual translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15849–15863, Toronto, Canada. Association for Computational Linguistics.

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. 2021. Whitening sentence representations for better semantics and faster retrieval. *ArXiv*, abs/2103.15316.

Xu Tan, Yi Ren, Di He, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2019. Multilingual neural machine translation with knowledge distillation. In *International Conference on Learning Representations*.

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

Yining Wang, Jiajun Zhang, Feifei Zhai, Jingfang Xu, and Chengqing Zong. 2018. Three strategies to improve one-to-many multilingual translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2955–2960, Brussels, Belgium. Association for Computational Linguistics.

Zirui Wang, Yulia Tsvetkov, Orhan Firat, and Yuan Cao. 2021. Gradient vaccine: Investigating and improving multi-task optimization in massively multilingual models. In *International Conference on Learning Representations*.

Rachel Wicks and Kevin Duh. 2022. The effects of language token prefixing for multilingual machine translation. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and*

the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 148–153, Online only. Association for Computational Linguistics.

Derrick Xin, Behrooz Ghorbani, Justin Gilmer, Ankush Garg, and Orhan Firat. 2022. Do current multi-task optimization methods in deep learning even help? In *Advances in Neural Information Processing Systems*.

Sangwon Yu, Jongyoon Song, Heeseung Kim, Seongmin Lee, Woo-Jong Ryu, and Sungroh Yoon. 2022. Rare tokens degenerate all tokens: Improving neural text generation via adaptive gradient gating for rare token embeddings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29–45, Dublin, Ireland. Association for Computational Linguistics.

Biao Zhang, Ankur Bapna, Rico Sennrich, and Orhan Firat. 2021. Share or not? learning to schedule language-specific capacity for multilingual translation. In *International Conference on Learning Representations*.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Yaoming Zhu, Jiangtao Feng, Chengqi Zhao, Mingxuan Wang, and Lei Li. 2021. Counter-interference adapter for multilingual machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2812–2823, Punta Cana, Dominican Republic. Association for Computational Linguistics.

A. WMT Data Preprocessing

We preprocess and filter the WMT training data in order to ensure a set of high quality bitext from the original crawled data provided by organizers. Steps 1-4 are reproduced from Fan et al. (2021).

1. Remove lines that are > 50% punctuation
2. Deduplicate training data
3. Language-specific filtering to remove sentences that are > 50% characters that are not identified as belonging to the given language.
4. Length ratio cleaning with ratio=3, and remove sentences with > 250 subwords.
5. Language identification filter such that both the source and target language ID must be correct. We use the `fasttext` LangID model: `lid.176.bin`. (Joulin et al., 2016, 2017).
6. Bitext filtering using LASER Embeddings as implemented by the `OpusFilter` toolkit (Aulamo et al., 2020; Artetxe and Schwenk, 2019).

B. TED Models on WMT dev set

langs	type	TED		
		BLEU	iso-enc	iso-dec
en-ru	multi	10.6	0.102	0.227
	bi	10.3	**0.107**	**0.264**
en-zh	multi	16.4	**0.084**	0.166
	bi	15.3	0.034	**0.194**
	multi	-	0.092	0.056
en-ru	multi	11.0	**0.097**	**0.243**
	bi	10.2	0.076	0.235
en-de	multi	16.5	**0.073**	0.223
	bi	15.2	0.040	**0.227**
	multi	-	0.079	0.085
en-ru	multi	7.7	**0.125**	**0.228**
	bi	7.1	0.103	0.213
en-uk	multi	11.2	**0.143**	**0.202**
	bi	10.0	0.131	0.188
	multi	-	0.130	0.174

Table 4: Isotropy results on our multiparallel TED model, tested on the WMT development set for direct comparison with our other models.

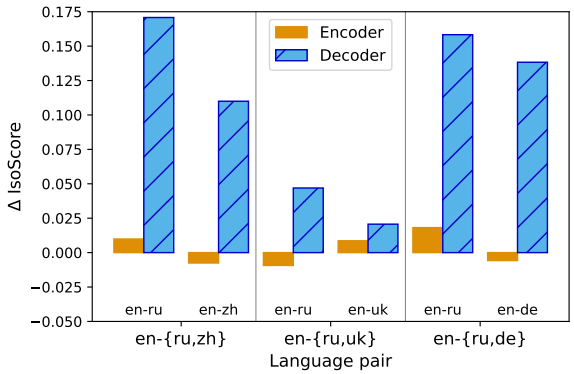


Figure 7: Δ IsoScore values between language-specific multilingual representations separated by language and overall multilingual representations, for both the encoder and decoder.