Penalizing Divergence: Multi-Parallel Translation for Low-Resource Languages of North America

Garrett Nicolai **Changbing Yang** **Miikka Silfverberg**
University of British Columbia
first.last@ubc.ca

Abstract

This paper explores a special case in multilingual machine translation: so called multi-parallel translation, where the target data for all language pairs are identical. While multi-parallelism offers benefits which are not available in a standard translation setting, translation models can easily overfit when training data are limited. We introduce a regularizer, the divergence penalty, which penalizes the translation model when it represents source sentences with identical target translations in divergent ways. Experiments on very low-resourced Indigenous North American languages show that an initially deficient multilingual translator can improve by 4.9 BLEU through mBART pre-training, and 5.5 BLEU points with the strategic addition of monolingual data, and that a divergence penalty leads to further increases of 0.4 BLEU. Further experiments on Germanic languages demonstrate a improvement of 0.5 BLEU when applying the divergence penalty. An investigation of the neural encoder representations learned by our translation models shows that the divergence penalty encourages models to learn a unified neural interlingua.

1 Introduction

Bilingual neural translation models typically require millions of parallel sentences to achieve adequate quality. A vast majority of the world's languages lack sufficient parallel corpora, and translation efforts for low-resource languages have turned to multilingual methods, leveraging linguistic similarity to augment a deficient signal with plentiful, albeit sometimes noisy, data from related languages (Aharoni et al., 2019; Goyal et al., 2020). In this paper, we explore a very specific multilingual translation setting: multi-parallel translation. Here, models are trained on documents, such as the proceedings of the European Parliament, collections of subtitles, and the Bible. Each sentence has translations in many languages, providing not just a bilingual signal, but one that is bilingual in many directions.

We explore a massively multi-parallel document: the Bible, which has hundreds of translations. While it represents a very restricted domain, the Bible is the only parallel document available for many languages and multi-parallel translation is, therefore, of key importance for low-resource NLP.

Earlier work shows that multi-parallel translation systems can in practice deliver poor results when available training data for individual languages are very small (Mueller et al., 2020). In this setting, the performance of the translation models degrades when the number of source languages is increased. Adapting pre-trained multilingual models such as mBART has also not led to significant progress (Lee et al., 2022). For languages that are not closely-related to the languages in the model, high-quality translation remains an unsatisfied goal.

We hypothesize that this drop in performance is a consequence of an inability of the model to learn a neural interlingua (Johnson et al., 2017), that is, a shared semantic representation for source languages. This prevents knowledge transfer between languages, degrading translation performance. To counteract this tendency, we present the divergence penalty—a modification to the standard loss of a multilingual translation system, which encourages identical encoder and decoder representations for parallel source sentences.

Our experiments on Indigenous North American languages show several avenues for improving the quality of the embedding space. We are able to stabilize the embeddings through mBART pre-training[1] and the addition of monolingual corpora, with gains of up to 5.5 BLEU in the super-low setting. It can be supplemented, however, with our di-

[1] In contrast to Lee et al. We hypothesize that the multi-parallel setting may be responsible, but it is beyond the scope of this paper.

Proceedings of the 29th International Conference on Computational Linguistics, pages 4292–4298
October 12–17, 2022.

vergence penalty, which further coerces the embeddings to adopt interlingual representations. Both visual inspection of t-SNE plots (Hinton and Roweis, 2002) and quantitative examination of learned representations show that the divergence penalty encourages the encoder and decoder representations to cluster according to semantics of the source sentence, regardless of source language.

This is not the first work to explore approaches to strengthening the embedding space of a multilingual translation model. Mullov et al. (2021) use cross-lingual word embeddings. Others have used explicit neural interlinguas (Zhu et al., 2020), and multiple encoders with tied attention (Vázquez et al., 2019). More closely related to our work, Yang et al. (2021) apply an agreement objective which encourages similar representations for artificially code-switched sentence variants and the original sentences. Pan et al. (2021) use contrastive learning to encourage shared representations for semantically similar sentences which resembles our divergence penalty. Finally, Arivazhagan et al. (2019) introduce an auxiliary loss which enforces multilingual similarity to improve zero-shot translation results. All of the aforementioned works, however, investigate translation in a substantially higher-resourced setting and none of them investigate multi-parallel translation.

A second class of related research falls into the broad category of data augmentation for low-resource translation. Sennrich and Zhang (2019) demonstrate that neural translation can learn in low-resource settings without significant modifications, but that these systems can be very sensitive to hyper-parameter tuning. Currey et al. (2017) demonstrate that the expedient method of copying source data to the target can improve low-resource translation quality, while Madaan and Sadat (2020) further leverage back-translations to boost the signal of low-resource translations. Likewise, Rubino et al. (2020) extol the virtues of monolingual data and back-translations in low-resource settings.

2 Methods

In a multi-parallel translation setting, each sentence in the training data has been translated into several languages. A standard multilingual transformer model learns from sentences in isolation, and does not leverage co-dependencies between source sentences in a multi-parallel scenario. We introduce the *divergence penalty* as an auxiliary loss which penalizes models which do not learn similar encoder and decoder representations for parallel sentences. Within each training batch, we identify sentences that are parallel with each other (via their targets), and compute a pairwise cosine comparison of their representation vectors at each position in the sentence. In a true interlingua, these representations would be identical.

The transformer modifies the embedding space at several points, and we compare three variants that utilize a snapshot of the embeddings at a specific point. First, we calculate the cosine distance after context-attention has been applied to the final layer of the encoder (EP). Secondly, we calculate the distance on the output distributions (DP). Finally, we sum the two together (BOTH). Given a sequence $\mathbf{r} = \mathbf{r}_1, ..., \mathbf{r}_n$ of encoder, decoder or joint encoder and decoder representations, the divergence penalty takes the form:

$$\mathcal{L}_{DIV}(\mathbf{r}) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(1 - \mathbf{r}_i^\top \mathbf{r}_j)}{n^2}$$

We then weight this distance (via a tunable hyper-parameter $\alpha_{DIV} \in [0, 1]$), and add it to the batch loss. Batches with parallel sources propagate higher loss if the model has learned divergent representations. Batches that have no multi-parallel sentences see no modification. During training, all parallel translations of the same target sentence are added into the same batch.

Training is performed using the Fairseq (Ott et al., 2019) implementation of transformers, with 3 layers and 4 attentional heads. Embedding dimensions are set at 512, while the feed-forward size is 1024. The model optimizes a label-smoothed cross entropy using Adam(0.9, 0.98), and an inverse square-root learning rate schedule (5e-4 - 1e-9). The model is run for 50 epochs, with the best model chosen via validation loss. These settings closely follow Nicolai et al. (2021). α_{DIV} is tuned for each model, with values in [0, 0.3] typically leading to the best results.

3 Data and Architectures

Our experiments are conducted on Bible data (Nicolai et al., 2021) for three Indigenous language families of North America: Algic, Athabaskan, and Inuit-Aleut. The target language in most of our experiments is English although we also train many-to-many translation systems. Family data sets are constructed by concatenating individual language

Family	Language	Train	Test
Algic	Algonquin	7133	-
	Arapaho	1024	-
	Cree	30269	394
	Mikmaq	7133	394
	Ojibwa	9795	-
	Potawatomi	1870	-
	Siksika	965	-
Algic	Apache	7131	-
	Carrier	8667	-
	Dane-zaa	616	-
	Gwich'in	7132	-
	Navajo	30276	394
	Tlicho	8667	394
	Tsilhqot'in	602	-
Inuit-Aleut	Inuinnaqtun	4289	-
	Inuktitut	30275	394
	Inupiatum	7132	394
	Yupik	30276	394
Germanic	Afrikaans	30249	394
	Bokmål	6719	365
	Danish	30276	394
	Dutch	30216	394
	German	30107	394
	Icelandic	7000	390
	Low German	7116	394
	Nynorsk	6719	365
	Swedish	29870	394
	Swiss German	7120	393

Table 1: The number of Bible verses used for training and testing for different languages. The New Testament consists of approximately 7130 verses, while a full translation is approximately 30275 verses. Numbers are approximate due to verse-splitting techniques.

Bibles, and prepending a language tag. We additionally perform experiments on Germanic Bibles (McCarthy et al., 2020).

Some languages have a full Bible translation available, while most only have a subset. See Table 1 for details. Evaluation is performed on languages with full Bible translations, as well as one ultra low-resource language for each family: Inupiatum (Inuit-Aleut), Mikmaq (Algic), and Tlicho (Athabaskan). For the sake of this paper, we consider "low-resource" to represent languages that have complete Bible translations of 30,000 sentences. "Ultra low-resource" languages are those that have only 7,000 sentences in the New Testament.

For development and testing, we sample each book of the New Testament at a rate of 1% for test, and a further 1% for development. All datasets are segmented using a joint source-target Byte Pair Encoding with a vocabulary size of 16,000. Explorations varying the vocabulary size suggested that this was a reasonable, stable choice for these data sets.[2]

Models are learned using a modified version of the Fairseq (Ott et al., 2019) implementation of transformers. Each model is trained with 3 layers and 4 attentional heads, with an embedding size of 512, and a feed-forward size of 1024. Models are optimized using Adam(0.9, 0.98), and an inverse square root learning schedule starting at 1e-7. Models are trained for a maximum of 50 epochs, with a batch size of 2000 tokens. Dropout is 50%, while attentional dropout is 30%.

4 Experiments

In our experiments, we investigate the performance of different multi-parallel configurations on our Indigenous and Germanic data. We start by training baseline bilingual X-to-English (2L) and multilingual F-to-English systems (M2E), where X is an Indigenous or Germanic language and F is one of the Indigenous language families or the Germanic family. We also continue training on mBART (Liu et al., 2020) for a maximum of 50 epochs.[3]

Indigenous languages Table 2 shows that apart from Mikmaq, Indigenous M2E translators see a sharp decrease in performance from their bilingual analogues–even for ultra low-resource languages. The average BLEU score drops from 10.5 points for bilingual translations models to 6.1 for multilingual models. For Indigenous languages, mBART shows the importance of the language model in multilingual translation. A strong target language model, even in another domain, is enough to learn a model that improves notably over the bilingual one.

Turning to the second sub-table (Raw) in Table 2, we observe that instituting a divergence penalty on the encoder (EP) restores almost all of the quality of the higher-resource languages, and improves ultra low-resource translation performance by 1 BLEU. The decoder penalty (DP) results in a slight 0.2

[2]Our code and datasets are available at `anonymized/for/review`.

[3]Since mBART does not contain any of our Indigenous languages, we tokenize them with the English tokenizer, and use a language identifier to guide the source-to-English translation. Inuktitut is written in its own script, which likely explains its underperformance.

	Baselines				Raw			+E2E			
Lang	2L	M2E	M2M	mBART	EP	DP	BOTH	M2E	+EP	+DP	+BOTH
Cree	17.7	13.7	3.4	**19.4**	13.7	13.7	13.8	16.1	16.1	17.1	16.3
Navajo	12.1	4.1	2.7	12.0	10.0	4.6	10.8	12.6	12.6	**12.9**	12.6
Yupik	12.7	4.1	2.4	16.8	13.6	4.1	13.7	**14.8**	14.3	14.3	14.7
Inuktitut	13.0	3.7	2.6	1.6	14.0	3.7	13.3	14.7	14.7	**15.0**	14.9
Average Full	13.9	6.4	2.8	12.5	12.8	6.5	12.9	14.6	14.4	**14.8**	14.6
Mikmaq	4.2	9.8	3.0	7.9	8.4	9.8	8.4	9.8	9.8	9.6	**10.4**
Tlicho	7.6	3.7	2.5	9.4	9.1	4.2	9.8	11.3	11.3	**12.1**	11.3
Inupiatum	5.9	3.5	2.3	14.4	11.5	4.1	11.8	12.4	12.2	**13.0**	12.3
Average NT	5.9	5.7	2.6	10.6	9.7	6.0	10.0	11.2	11.1	**11.6**	11.3
Average All	10.5	6.1	2.7	11.6	11.5	6.3	11.7	13.1	13.0	**13.4**	13.2

Table 2: Translation results. The baselines compare other data augmentation strategies - 2L is a bilingual model; M2E is plain many-to-English translation; M2M is many-to-many translation, and mBART is the 25 language mBART model. The Raw columns apply the divergence penalties to the encoder (EP), decoder (DP) and a combination of both (BOTH) of the baseline M2E model. +E2E adds in a source-target copy of the English Bible.

BLEU improvement over the initial multilingual model. This makes sense, as the decoder requires a strong encoder representation to learn successful translations. Combining encoder and decoder penalties (BOTH) gives a 5.6 BLEU improvement over the M2E model and a 1.2 BLEU improvement over bilingual models. The improvement over bilingual models for ultra low-resource languages is substantial at 4.1 BLEU. Furthermore, for two of our ultra low-resource languages, BOTH also outperforms mBART, suggesting that while a strong target language model is important, focused embedding space modification can also lead to improvements.

Rather than augment the mBART architecture with our penalty, we instead mimic the language model through the addition of monolingual English-to-English (E2E) examples: from each English target sentence in our training data, we generate a new translation example with identical source and target sentence, and append the example to our multilingual training set.[4] When applying the divergence penalty, English is treated as an additional source language in the training set.

Adding monolingual data mimics mBART, pushing both low and higher-resource languages beyond the divergence penalty alone. Adding the encoder penalty on top of E2E does not improve results. However, the decoder penalty leads to a further average improvement of 0.3 BLEU, while the combination BOTH fares slightly worse. We hypothesize that monolingual data and the encoder penalty behave similarly, anchoring multilingual representations in encoder space and allowing translations to cluster around the English representations. We explore this further in Section 5.

[4]We also experimented with back-translation, but the quality of the back-translated training data were very poor, and hurt model quality, overall. We hypothesize that this is a result of the small size of our training sets, which do not allow us to learn a back-translation model of sufficient quality.

Lang.	2L	M2E	+EP	+DP	+BOTH
Afrikaans	27.4	25.7	25.7	25.9	25.7
Danish	26.5	26.5	26.5	25.9	26.5
German	25.5	25.9	25.9	25.8	25.9
Dutch	26.4	25.9	25.9	26.2	25.9
Swedish	23.7	23.9	23.9	24.7	23.9
Ave. Full	**25.9**	25.6	25.6	25.7	25.6
Swiss German	12.5	21.4	21.4	21.3	21.4
Low German	12.0	19.2	19.2	19.7	19.2
Nynorsk	12.6	22.3	22.3	22.9	22.3
Bokmål	12.2	22.4	22.4	23.2	22.4
Icelandic	11.3	18.9	18.9	19.2	18.9
Ave. NT	12.1	20.8	20.8	**21.3**	20.8
Ave. All	19.0	23.2	23.2	**23.5**	23.2

Table 3: Germanic results.

Germanic languages For our Germanic language experiments, all models are trained with E2E data because this strategy was found to be universally beneficial for Indigenous languages. In contrast to Indigenous languages, none of the Germanic languages see a drop in performance from bilingual to multilingual translation models as demonstrated in Table 3. Multilingual models substantially improve performance for the ultra low-resource Germanic languages where average performance improves by 8.7 points BLEU. For Germanic languages, EP does not result in improvements but DP improves performance for ultra low-resource languages by an average 0.5 points BLEU. A combination of EP and DP does not provide further gains in BLEU.

5 Analysis

To understand how the encoder and decoder penalties affect multi-parallel translation models, we use t-SNE to plot the encoder representations of source

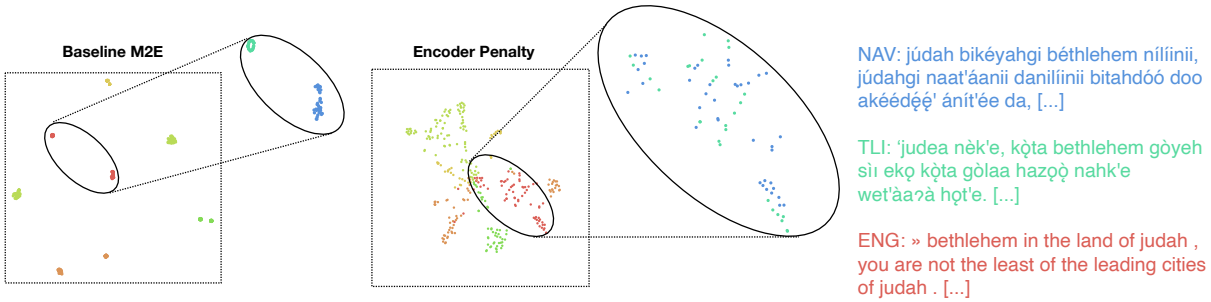

Figure 1: Example t-SNE plots of multilingual encoder representations for baseline M2E models and models with encoder penalty. The square plot represents encoder representations for 4 different sentences - each point represents a single encoder representation, and each sentence is represented by a different color. When we zoom in on the red cluster (representing the sentence shown on the right), we show Navajo representations in light blue and Tlicho in green.

sentences for our baseline multilingual Athabaskan model and a model trained with the encoder penalty in Figure 2.[5] Sentences are color-coded. For example, red dots include representations for one Tlicho and one Navajo source sentence with identical English translations. When we Zoom into the red cluster, showing Navajo representations in light blue and Tlicho in green, we can see that representations for the baseline multilingual system (**Baseline M2E**) form very tight clusters. However, in many cases, these cluster according to source language. This indicates that the model has failed to learn truly multilingual representations, which would instead cluster according to semantics of the source *sentence*, rather than source *language*.

When we add in the encoder penalty (**Encoder Penalty**), we see a significant correction. Tlicho and Navajo representations for equivalent source sentences now intermingle in a joint red cluster. Models supplemented by monolingual data also seem to cluster representations by meaning (included in Appendix A), demonstrating a similar tendency which supports our conclusion that both monolingual data augmentation and the encoder penalty strengthen learning of shared multilingual representations. The decoder penalty does not seem to have a similar effect (see appendix A).

The visual interpretation of t-SNE plots is confirmed via mathematical analysis. We calculate the centroid of each cluster (formed by representations for a particular source language and sentence ID), and determine the average cosine distance between centroids for clusters having the same sentence ID across languages. Formally, let $R(i, l)$ be the set of representations for language l and sentence ID i, and let $\mu(i, l)$ be the centroid of $R(i, l)$. We then compute: $d = \sum_{i=0}^{n} \text{dist}(\mu(i, l_1), \mu(i, l_2))/n$, where dist is cosine distance and n is our number of sentences. We compute these numbers over the entire test set for language l_1 and l_2. Comparing Navajo and Tlicho for the Athabaskan languages family, the baseline M2E model has an average distance of 0.1950, while the EP decreases the value dramatically, to 0.0018.

[5] We first use PCA to project into $\mathbb{R}^{10}$, and then use t-SNE to further project the results to $\mathbb{R}^2$. t-SNE plots ran for a maximum of 150,000 iterations, with perplexity of 30.

6 Conclusion

Multi-parallel translation has the ability to leverage cross-lingual information to supplement a weak low-resource translation signal, but not all languages can benefit. We have introduced a divergence penalty that forces multi-parallel models to learn shared embedding spaces that improve the quality of the translation. On its own, the penalty improves the quality of ultra low-resource Indigenous translation by 4.1 BLEU over a bilingual model, and by more than 4.3 BLEU over a deficient multilingual alternative. Furthermore, monolingual data provides a strong target for multilingual embeddings, but is complemented by our penalty for a further increase of 0.4 BLEU. This trend continues even when the the number of large translation corpora is increased.

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.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat,

Roee Aharoni, Melvin Johnson, and Wolfgang Macherey. 2019. The missing ingredient in zero-shot neural machine translation. *arXiv preprint arXiv:1903.07091*.

Anna Currey, Antonio Valerio Miceli-Barone, and Kenneth Heafield. 2017. Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 148–156.

Vikrant Goyal, Sourav Kumar, and Dipti Misra Sharma. 2020. Efficient neural machine translation for low-resource languages via exploiting related languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 162–168.

Geoffrey Hinton and Sam T Roweis. 2002. Stochastic neighbor embedding. In *NIPS*, volume 15, pages 833–840. Citeseer.

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

En-Shiun Annie Lee, Sarubi Thillainathan, Shravan Nayak, Surangika Ranathunga, David Ifeoluwa Adelani, Ruisi Su, and Arya D McCarthy. 2022. Pre-trained multilingual sequence-to-sequence models: A hope for low-resource language translation? *arXiv preprint arXiv:2203.08850*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Pulkit Madaan and Fatiha Sadat. 2020. Multilingual neural machine translation involving Indian languages. In *Proceedings of the WILDRE5–5th Workshop on Indian Language Data: Resources and Evaluation*, pages 29–32.

Arya D McCarthy, Rachel Wicks, Dylan Lewis, Aaron Mueller, Winston Wu, Oliver Adams, Garrett Nicolai, Matt Post, and David Yarowsky. 2020. The Johns Hopkins University Bible corpus: 1600+ tongues for typological exploration. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2884–2892.

Aaron Mueller, Garrett Nicolai, Arya D. McCarthy, Dylan Lewis, Winston Wu, and David Yarowsky. 2020. An analysis of massively multilingual neural machine translation for low-resource languages. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3710–3718, Marseille, France. European Language Resources Association.

Carlos Mullov, Ngoc-Quan Pham, and Alexander Waibel. 2021. Unsupervised transfer learning in multilingual neural machine translation with cross-lingual word embeddings. *arXiv preprint arXiv:2103.06689*.

Garrett Nicolai, Edith Coates, Ming Zhang, and Miika Silfverberg. 2021. Expanding the JHU Bible Corpus for machine translation of the indigenous languages of north america. In *Proceedings of the Workshop on Computational Methods for Endangered Languages*, volume 1, pages 1–5.

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 NAACL-HLT 2019: Demonstrations*.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. *arXiv preprint arXiv:2105.09501*.

Raphael Rubino, Benjamin Marie, Raj Dabre, Atushi Fujita, Masao Utiyama, and Eiichiro Sumita. 2020. Extremely low-resource neural machine translation for Asian languages. *Machine Translation*, 34(4):347–382.

Rico Sennrich and Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221.

Raúl Vázquez, Alessandro Raganato, Jörg Tiedemann, and Mathias Creutz. 2019. Multilingual NMT with a language-independent attention bridge. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 33–39.

Jian Yang, Yuwei Yin, Shuming Ma, Haoyang Huang, Dongdong Zhang, Zhoujun Li, and Furu Wei. 2021. Multilingual agreement for multilingual neural 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 2: Short Papers)*, pages 233–239, Online. Association for Computational Linguistics.

Changfeng Zhu, Heng Yu, Shanbo Cheng, and Weihua Luo. 2020. Language-aware interlingua for multilingual neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1650–1655.

A Encoder Space Plots

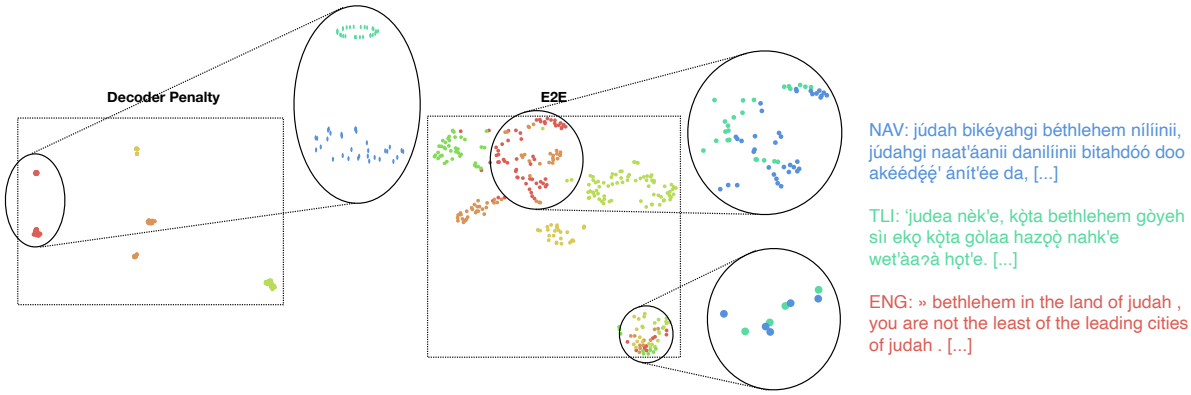

Figure 2: Example t-SNE plots of multilingual encoder representations for M2E models with decoder penalty (on the left) and augmented with English monolingual data (on the right). Each color in the original plot encodes representations for parallel source sentences both in Tlicho and Navajo. When we zoom in on the red cluster, we show Navajo representations in light blue and Tlicho in green. We can see that the decoder penalty does not help the model to learn shared encoder representations. Instead the representations of the Tlicho and Navajo sentences form distinct clusters. When we instead apply mononlingual data augmentation, the representations for the Tlicho and Navajo sentences cluster by meaning and we get shared multilingual representations. For a reason unknown to us, function words, punctuation and language tags form a tight cluster in the lower right corner of the plot when using monolingual data augmentation.