Abstract

Despite constant improvements in machine translation quality, automatic poetry translation remains a challenging problem due to the lack of open-sourced parallel poetic corpora, and to the intrinsic complexities involved in preserving the semantics, style and figurative nature of poetry. We present an empirical investigation for poetry translation along several dimensions: 1) size and style of training data (poetic vs. non-poetic), including a zero-shot setup; 2) bilingual vs. multilingual learning; and 3) language-family-specific models vs. mixed-multilingual models. To accomplish this, we contribute a parallel dataset of poetry translations for several language pairs. Our results show that multilingual fine-tuning on poetic text significantly outperforms multilingual fine-tuning on non-poetic text that is 35X larger in size, both in terms of automatic metrics (BLEU, BERTScore, COMET) and human evaluation metrics such as faithfulness (meaning and poetic style). Moreover, multilingual fine-tuning on poetic data outperforms bilingual fine-tuning on poetic data. 1

1 Introduction

American poet Robert Frost once defined poetry as “that which gets lost out of both prose and verse in translation” (Frost, 1961). Indeed, the task is so complex that translators often have to “create a poem in the target language which is readable and enjoyable as an independent, literary text” (Jones, 2011). But even though poetry is destined to lose its accuracy, integrity, and beauty even in human translation, the process conceives new opportunities to stress-test the ability of machine translation models to deal with figurative speech.

While most computational work has focused on poetry generation (Hopkins and Kiela, 2017; Uthus et al., 2021; Van de Cruys, 2020; Ghazvininejad et al., 2016; Li et al., 2020; Hämäläinen and Alnajjar, 2019; Yi et al., 2020; Deng et al., 2019; Chen et al., 2019; Yang et al., 2018), research on poetry translation is in its infancy (Ghazvininejad et al., 2018; Genzel et al., 2010).

For example, Ghazvininejad et al. (2018) employs a constrained decoding technique to maintain rhyme in French to English poetry translation. However, while keeping the poetic style and fluency, the translation might diverge in terms of meaning w.r.t. the input. Table 1 shows how the system generates a semantically inconsistent word “masquerade” to rhyme with “laid”, whereas the original poem talks about “tomb”.

Meanwhile, state-of-the-art machine translation systems trained on large non-poetic data might preserve meaning and fluency, but not the poetic style (see Google translation in Table 1).

| Original | Human (2018) | Google |
|----------|-------------|--------|
| Il ny avait que sable et boue Où s'était ouverte la tombe. Le long des murs de la prison On ne voyait aucune tombe | For where a grave had opened wide, There was no grave at all: Only a stretch of mud and sand By the hideous prison-wall, | There was only sand and mud Where the grave was laid. Along the prison walls We saw no masquerade |
| | But there was only sand and mud. To where the grave was laid. Along the walls of prison wall. | Where the grave had opened. Along the prison walls No tomb could be seen. |

Table 1: A French poem accompanied by human translation, translation from (Ghazvininejad et al., 2018) system, and Google Translate. 2

---

1The italics part of the title is the translation of a poem by Pablo Neruda with the same name.

2The bold part of the table is the human translation provided by the authors.

---
more than 190,000 lines of poetry spanning over six languages. We try to tackle the second challenge by leveraging multilingual pre-training (e.g., mBART (Liu et al., 2020)) and multilingual fine-tuning (Tang et al., 2020; Aharoni et al., 2019) that have recently led to advances in neural machine translation for low-resource languages. Moreover, it has been shown that adaptive pre-training and/or fine-tuning on in-domain data always lead to improved performance on the end task (Gururangan et al., 2020).

Since poetry translation falls into the low-resource (no or little parallel data) and in-domain scenario, we present an empirical investigation on whether these advances bring us a step closer to poetry translation systems that don’t go far off in terms of faithfulness (i.e., keeping the meaning and poetic style of the input).

We make the following contributions:

• We release several parallel poetic corpora enabling translation from Russian, Spanish, Italian, Dutch, German, and Portuguese to English. We also release test sets for poetry translation from Romanian, Ukrainian and Swedish to evaluate the zero-shot performance of our models.

• We show that multilingual fine-tuning on poetic text significantly outperforms multilingual fine-tuning on non-poetic text that is 35X larger in size (177K vs 6M), both in terms of automatic and human evaluation metrics such as faithfulness. However, for the bilingual case the pattern is not so evident. Moreover, multilingual fine-tuning on poetic data outperforms bilingual fine-tuning on poetic data. The latter two results showcase the importance of multilingual fine-tuning for poetry translation.

• We also show that multilingual fine-tuning on languages belonging to the same language family sometimes leads to improvement over fine-tuning on all languages.

Beyond advancing poetic translation, our findings will be helpful for other figurative language or literary text translation tasks, and we hope that resources and experiments released will encourage further research in this area. Our code and data and can be found in https://github.com/tuhinjubcse/PoetryTranslationEMNLP2021 while our pre-trained models can be found at https://huggingface.co/TuhinColumbia.

2 Datasets

2.1 Poetic Training Data

Given the lack of available multilingual poetic corpora, we collect several medium-scale parallel datasets. We identify websites that provide English translations for Spanish (Es), Russian (Ru), Portuguese (Pt), German (De), Italian (It) and Dutch (Nl) poetry. Table 2 shows the number of parallel sentences for each language pair as well as the websites from which they have been collected. Given that most of the websites were specifically designed for poetry translation, where translations are typically written by experts (professional translators), we believe our data to be of high quality. We make a simplifying assumption and focus on line-by-line translation. Thus, during scraping from these websites, we discard translations that are different in

| Language Pair       | Source                                                                 | Train | Valid | Test |
|---------------------|------------------------------------------------------------------------|-------|-------|------|
| Spanish-English     | https://www.poesi.as/ https://lyricstranslate.com/                     | 37,746| 2059  | 536  |
| Russian-English     | http://www.poesisfromtheportuguese.org/ https://lyricstranslate.com/   | 50,001| 4186  | 548  |
| Portuguese-English  | http://www.poemswithoufrontiers.com/ https://www.poetryinternational.org/ https://lyricstranslate.com/ | 15,199| 699   | 140  |
| German-English      | http://www.poemswithoutfrontiers.com/ https://www.poetryinternational.org/ https://lyricstranslate.com/ | 17,000| 1,050 | 1295 |
| Italian-English     | https://digitaldante.columbia.edu/ https://www.poetryinternational.org/ https://lyricstranslate.com/ | 34,534| 1,997 | 528  |
| Dutch-English       | https://www.poetryinternational.org/ https://lyricstranslate.com/      | 23,403| 1,000 | 159  |
| Language       | Translation                                                                 |
|---------------|----------------------------------------------------------------------------|
| Russian       | Their love was so gentle, so long, and surprising.                          |
|               | With pining, so deep, and zeal, like a crazy uprising!                    |
| Spanish       | I can write the saddest lines tonight.                                    |
|               | I loved her, sometimes she loved me too.                                  |
| Portuguese    | To a garden luxuriously verdant and enamelled with countless flowers      |
| German        | We’ll sacrifice your chastity first, dearest and get the gift of language in return |
| Italian       | Round, frozen in its oceans, transparent like a cell under the microscope |
| Dutch         | Evening and the glow stick’s in your eyes and you are looking its orange snapped into your eyes the liquid light |

Table 3: Parallel Poetic translations written by humans from our multilingual datasets.

the number of lines from the original poems. We collect approximately 190K (with 177K in training) parallel poetic lines spanning 6 different languages (see Table 3 for examples). This data is further split into train and validation.

2.2 Non-Poetic Training Data

We also benchmark the quality of poetry translations obtained by models trained on non-poetic data. For this we rely on OPUS100 corpus (Tiedemann, 2012) as well as the ML50 corpus (Tiedemann et al., 2020) designed to demonstrate the impact of multilingual fine-tuning. Each of the language pairs in OPUS100 have 1 million parallel sentences in their training set, several orders of magnitude larger than our poetic parallel data. For example, Portuguese-English non-poetic data is 65 times larger than the poetic data, while the Russian-English non-poetic data is 18 times larger than the poetic data. The size of the smallest non-poetic parallel corpus is about 6 times larger than all our poetic parallel data combined. For the ML50 (Tang et al., 2020), benchmark data is collected across 50 languages from publicly available datasets such as WMT, IWSLT, WAT, TED. The size of parallel sentences in ML50 corresponding to the languages under study are: De (45.8M), Es (14.5M), Ru (13.9M), Ni (0.23M), It (0.2M), and Pt (0.04M).

2.3 Test Data

We create a high quality blind test set for every language independent of data mentioned in Table 2 by carefully hand-picking poems unseen in the training or validation set. Every line has a single reference. Our blind test consists of 3522 sentences spanning across 209 poems in 6 languages. Our combined multilingual test set consists of 548 lines in Russian, 536 lines in Spanish, 528 lines in Italian, 1295 lines in German, 140 lines in Portuguese and 159 lines in Dutch. We also test our models on 7 Ukrainian poems (100 lines), 8 Romanian poems (100 lines) and 7 Swedish poems (100 lines) in a zero-shot setting.

3 Methods

mBART (Liu et al., 2020) is a multilingual sequence-to-sequence (seq2seq) denoising autoencoder, which is trained by applying the BART objective (Lewis et al., 2019) to large-scale monolingual corpora across many languages. The input texts are noised by masking phrases and permuting sentences, and a single Transformer model is learned to recover the texts. Unlike other pre-training approaches for machine translation, mBART pre-trains a complete autoregressive seq2seq model. It is trained once for all languages, providing a set of parameters that can be fine-tuned for any of the language pairs for supervised machine translation without any task-specific or language-specific modifications or initialization schemes. For supervised sentence-level MT, mBART initialization leads to significant gains (up to 12 BLEU points) across low/medium-resource pairs (< 10M bi-text pairs). This makes mBART an ideal candidate for our task of poetry translation given the scale of our parallel corpora.

However, while mBART was trained on a variety of languages, the multilingual nature of the pre-training is not used during fine-tuning. To solve
this, Tang et al. (2020) propose multilingual fine-tuning of pre-trained models, and demonstrate large improvements compared to bilingual fine-tuning. They explore 3 configurations to create different versions of multilingual translation models: Many-to-one (N → 1), one-to-Many (1 → N), and Many-to-Many (N→N) via a pivot language. The Many-to-one model encodes N languages and decodes to English. Given that we are translating poems in various languages to English, we further fine-tune the Many-to-one model for our task.

3.1 Implementation Details

For bilingual fine-tuning on poetic data, we use the mbart-large-50 checkpoint from (Wolf et al., 2020), and fine-tune it for up to 8 epochs, saving the best checkpoint based on eval-BLEU scores. For bilingual fine-tuning on non-poetic data, we fine-tune the model for 3 epochs. For multilingual fine-tuning, we use the mbart-large-50-many-to-one-mmt. We perform multilingual fine-tuning for 3 epochs for both poetic/non-poetic data. We use the same hyperparameters as the standard huggingface implementation. We use (2-4) nvidia A100 GPUs for fine-tuning pretrained checkpoints. For fine-tuning mBART on non-poetic data, we set the gradient_accumulation_steps to 10 and batch size to 8 while for poetic fine-tuning we vary batch size between 24 and 32, and set gradient_accumulation_steps to 1.

To perform multilingual fine-tuning, we concatenate bitexts of different language pairs (i, j) into a collection \( B_{i,j} = (x_i, y_j) \) for each direction (i, j). Following mbart (Liu et al., 2020), we augment each bitext \((x_i, y_j)\) by adding a source and a target language token at the beginning of \(x\) and \(y\), respectively, to form a target language token augmented pair \((x_0, y_0)\). We then initialize transformer based seq-to-seq model by the pretrained mBART, and provide the multilingual bitexts \( B = \cup_{i,j} B_{i,j} \) to fine-tune the pretrained model.

4 Experimental Setting

We experiment with several systems to evaluate performance across several dimensions: poetic vs non-poetic data; multilingual fine-tuning vs. bilingual fine-tuning; language-family-specific models vs. mixed-multi language-models.

• Non-Poetic Bi (OPUS): fine-tuned mBART50 on Non-Poetic data from OPUS100 (Section 2.2) for respective languages bilingually.
• Non-Poetic Multi (ML50): mBART-large-50-many-to-one model implemented in the huggingface package. This is a multilingually fine-tuned model on 50 languages from the ML50 data that is 4 times larger than OPUS and created using all of the data that is publicly available (e.g., WMT, IWSLT, WAT, TED).
• Non-Poetic Multi (OPUS): multilingually fine-tuned mBART-large-50-many-to-one model on Non-Poetic data for 6 languages from OPUS100 (Section 2.2) (6M parallel sentences).
• Poetic: fine-tuned mBART50 bilingually (e.g., Ru-En, Es-En, It-En) on poetic data described in Section 2.1.
• Poetic All: multilingually fine-tuned mBART-large-50-many-to-one on all poetic data combined.
• Poetic LangFamily: multilingually fine-tuned mBART-large-50-many-to-one on poetic data for all languages belonging to the same language family. For instance, Pt, Es, It belong to the Romance language family, while De and Ni are both Germanic languages.

4.1 Automatic Evaluation Setup

For the automatic evaluation, we compare the performance of all the above mentioned models in terms of three metrics: BLEU, BERTScore and COMET.

BLEU (Papineni et al., 2002) is one of the most widely used automatic evaluation metrics for Machine Translation. We use the SacreBLEU (Post, 2018) python library to compute BLEU scores between the system output and the human written gold reference.

BERTScore (Zhang et al., 2019) has been used recently for evaluating text generation systems using contextualized embeddings, and it is said to somewhat ameliorate the problems with BLEU. BERTScore also has better correlation with human judgements (Zhang et al., 2019). It computes a similarity score using contextual embeddings for each token in the system output with each token in the reference. We report F1-Score of BERTScore. We use the latest implementation to date which replaces BERT with deberta-large-mnli, which is a
DeBERTa model (He et al., 2020) fine-tuned on MNLI (Williams et al., 2017).

Recently Kocmi et al. (2021) criticized the use of BLEU through a systematic study of 4380 machine translation systems and recommend use of a pre-trained metric COMET (Rei et al., 2020). COMET leverages recent breakthroughs in cross-lingual pre-trained language modeling resulting in highly multilingual and adaptable MT evaluation models that exploit information from both the source input and a target-language reference translation in order to more accurately predict MT quality. We rely on the recommended model \textit{wmt-large-da-estimator-1719}, which is trained to minimize the mean squared error between the predicted scores and the DA (Graham et al., 2013) quality assessments. Notice that these scores are normalized per annotator and hence not bounded between 0 and 1, allowing negative scores to occur; higher score means better translation.

4.2 Human-based Evaluation Setup

Even though arguably useful for evaluating meaning preservation, automatic metrics are not as suitable to measure other aspects of poetic translation such as the use of figurative language and style. We conduct human evaluation by recruiting three bilingual speakers as volunteers for each language. NMT systems are susceptible to producing highly pathological translations that are completely unrelated to the source input often termed as hallucinations (Raunak et al., 2021), e.g. see the word \textit{Lungs} in Table 10. To account for these effects, we use faithfulness as a measure that combines both meaning preservation and poetic style.

We presented the best translations from multilingual models trained on poetic and non-poetic data. Human judges were asked to evaluate on a binary scale: i) if the model introduces hallucinations or translates it into something arbitrary, a.k.a (Are they keeping the meaning of the input text?) and at the same time ii) if the syntax structure is poetic and whether the generations are rich in poetic figures of speech (e.g., metaphors, similes, personification).

In this evaluation we compare the multilingually fine-tuned models on Non-Poetic data (Non-Poetic Multi (OPUS) and Non-Poetic Multi (ML50)) vs. multilingually fine-tuned models on Poetic data (Poetic All and Poetic LangFamily). We chose the best model in each category based on the BERTScore in the automatic evaluation.

We chose a subset of the test set for human evaluation: 1044 sentences spanning across 80 poems in 6 languages (204 lines in Russian, 173 lines in Italian, 140 lines in Portuguese, 220 lines in Spanish, 148 lines in German, and 159 lines in Dutch with corresponding human translations). Human judges were also provided with gold translations to make the judgement easier. Agreement rates were measured using Krippendorff’s $\alpha$ and a moderate agreement of 0.61 was achieved.

5 Results

Our results based on automatic metrics are summarized in Table 4 and the human evaluation in Table 5.

The first insight is that multilingual fine-tuning on Poetic data (Poetic All and Poetic LangFamily) outperforms multilingual fine-tuning on Non-Poetic data (Non-Poetic Multi (ML50, Opus)) for all languages both in terms of automatic metrics (BLEU and BERTScore) and human evaluation based on faithfulness (Table 5). Between Non-Poetic Multi and Poetic-All we see at least 2.5 point improvement in BLEU scores as well as 1 point improvement in BertScore in translation of every language pair. For the recently developed metric COMET we see our best models are multilingually fine-tuned poetic models as well that is consistent with our results obtained on other two metrics.

However, when comparing the bilingually fine-tuned models (Poetic vs. Non-Poetic Bi(Opus)) the pattern is not so evident based on automatic metrics. We see comparable performance, but not a clear winner across languages and metrics. However, as with the multilingual case, the size of Poetic data is much smaller than the Non-Poetic data (20X to 50X smaller depending on language). We also mixed poetic and non-poetic data in equal proportion and fine-tuned mBART by framing it as a domain adaption problem, however it did not lead to significant improvements and degenerated in a few languages. We also try intermediate fine-tuning (Phang et al., 2018), where we first fine-tune a pre-trained mBART model on our Non-Poetic data and then fine-tune the best model checkpoint on our Poetic data. The results for this experiment also did not lead to any significant difference in performance.

The third insight is that language-family-specific multilingual fine-tuning (Poetic LangFamily) helps in some of the languages when compared to mul-
| Model                  | BLEU | BERTScore | COMET |
|-----------------------|------|-----------|-------|
| Non-Poetic Bi(OPUS) Ru-En | 12.4 | 65.4      | -47.83 |
| Non-Poetic Multi(ML50) Ru-En | 13.0 | 67.5      | -37.55 |
| Non-Poetic Multi(OPUS) Ru-En | 12.8 | 67.2      | -39.5  |
| Poetic Ru-En           | 11.9 | 64.3      | -55.14 |
| Poetic LangFamily      | -    | -         | -     |
| Poetic All             | 17.0 | 70.2      | -25.71 |
| Non-Poetic Bi(OPUS) Es-En | 26.9 | 74.6      | 1.43  |
| Non-Poetic Multi(ML50) Es-En | 5.1  | 58.9      | -60.98 |
| Non-Poetic Multi(OPUS) Es-En | 28.0 | 75.6      | 4.84  |
| Poetic Es-En           | 20.8 | 74.3      | -3.09  |
| Poetic LangFamily      | 30.9 | 77.2      | 12.14  |
| Poetic All             | 31.2 | 76.6      | 10.10  |
| Non-Poetic Bi(OPUS) Pt-En | 9.5  | 63.3      | -47.27 |
| Non-Poetic Multi(ML50) Pt-En | 7.3  | 62.7      | -53.48 |
| Non-Poetic Multi(OPUS) Pt-En | 9.2  | 64.0      | -42.86 |
| Poetic Pt-En           | 9.6  | 63.4      | -50.93 |
| Poetic LangFamily      | 12.5 | 66.4      | -39.36 |
| Poetic All             | 12.2 | 66.6      | -35.89 |
| Non-Poetic Bi(OPUS) It-En | 22.2 | 70.3      | -14.85 |
| Non-Poetic Multi(ML50) It-En | 17.0 | 68.7      | -24.53 |
| Non-Poetic Multi(OPUS) It-En | 22.9 | 71.1      | -8.87  |
| Poetic It-En           | 18.8 | 69.3      | -24.21 |
| Poetic LangFamily      | 25.4 | 72.2      | -7.35  |
| Poetic All             | 24.6 | 71.6      | -8.97  |
| Non-Poetic Bi(OPUS) De-En | 15.2 | 68.6      | -27.95 |
| Non-Poetic Multi(ML50) De-En | 20.1 | 73.4      | -5.88  |
| Non-Poetic Multi(OPUS) De-En | 17.8 | 70.9      | -16.77 |
| Poetic De-En           | 16.8 | 70.2      | -23.07 |
| Poetic LangFamily      | 20.5 | 73.6      | -4.22  |
| Poetic All             | 22.7 | 74.6      | -0.52  |
| Non-Poetic Bi(OPUS) Nl-En | 24.5 | 72.5      | -4.83  |
| Non-Poetic Multi(ML50) Nl-En | 23.8 | 72.2      | -6.73  |
| Non-Poetic Multi(OPUS) Nl-En | 26.1 | 72.9      | -4.83  |
| Poetic Nl-En           | 26.5 | 71.6      | -12.73 |
| Poetic LangFamily      | 32.1 | 74.3      | -3.74  |
| Poetic All             | 30.7 | 74.5      | -1.90  |

Table 4: Performance of mBART fine-tuned on different datasets in terms of automatic evaluation metrics on test data in various settings. Difference is significant, ($\alpha < 0.005$) via Wilcoxon signed-rank test.

We also ran a preliminary experiment where we tested if multilingual fine-tuning with a dissimilar language hurts the performance compared to fine-tuning with a language from the same language family (e.g., De and It vs. De and Nl). Our initial experiments show that fine-tuning on languages from the same language family helps compared to languages from different language family.

Last but not least, we notice that the multilingual fine-tuned model on poetic data (Poetic All) is consistently better than the bilingual fine-tuned model on poetic data (Poetic) across all languages.

While we show that multilingual fine-tuning is an effective way to improve performance on low resource poetic data, we believe techniques like iterative backtranslation (Hoang et al., 2018) with sophisticated techniques for data selection (Dou et al., 2020) or domain repair (Wei et al., 2020) could improve the performance of model trained on poetic data. We leave this for future work.

Zero-Shot Performance on Unseen Languages
We test the generalization capabilities of our model fine-tuned on poetic data using poetry written in languages not seen during fine-tuning. We compare

| Model                  | Non-Poetic Best | Poetic Best |
|-----------------------|-----------------|-------------|
| Ru-En                 | 20%             | 80%         |
| Es-En                 | 0%              | 100%        |
| Pt-En                 | 40%             | 60%         |
| De-En                 | 28%             | 72%         |
| Nl-En                 | 28%             | 72%         |
| Br-En                 | 0%              | 100%        |

Table 5: Human evaluation in terms of preference between multilingual fine-tuning on Non-Poetic data vs Poetic data, in terms of faithfulness. Significant difference ($\alpha < 0.05$) via Wilcoxon signed-rank test.
the zero-shot performance of our model fine-tuned multilingually on poetic data (excluding the unseen languages) to the Non-Poetic Multi (OPUS) and Non-Poetic Multi (ML50) model. We chose Ukrainian, Romanian and Swedish poetry given the fact that our model is fine-tuned on poetry belonging to languages from the Slavic, Romance, and Germanic families. Table 6 shows that our multilingually fine-tuned poetic model outperforms the other two multilingual models fine-tuned on Non-Poetic data, even though the languages were not contained in the fine-tuning data. This suggests that performance improvements of poetic fine-tuning are not only due to language-specific training data, but rather to multilinguality, presence of language family related data, as well as poetic style. These corroborate recent findings by Ko et al. (2021) who adapt high-resource NMT models to translate low-resource related languages without parallel data. They exploit the fact that some low-resource languages are linguistically related or similar to high-resource languages, and often share many lexical or syntactic structures.

6 Shortcomings of Style Transfer Techniques as a Post-Editing tool

We evaluate whether style transfer techniques could help attenuate the shortcomings of translation models trained on non-poetic data. We use the romantic poetry style transfer model provided by Krishna et al. (2020) to paraphrase our non-poetic translations. To control for faithfulness, we generate 20 outputs for each input (i.e., non-poetic translations) using nucleus sampling ($p = 0.6$), we then select the sentence that has the highest similarity score with input using the SIM model by Wieting et al. (2019).

The style transfer experiments decrease performance across all languages on both BLEU and BERTScore metrics as evaluated on the Multi(OPUS) model (see results in Table 7).

Qualitatively, this may happen due to errors cascading from incorrect translations by the non-poetic model, introduction of archaic language where it is not appropriate, and change in meaning. An example output is provided in Table 8.

Table 7: BLEU and BERTScore after style transfer applied to the Multi(OPUS) configuration. Value in parenthesis reports decrease from the score obtained just by using Multi(OPUS).

Table 8: Style transfer example. M=Multi(OPUS)

7 Analysis

It is well-known that occasionally NMT systems have a tendency to generate translations that are grammatically correct but unrelated to the source sentence particularly for low-resource settings (e.g., hallucinate words that are not mentioned in the source language) (Arthur et al., 2016; Koehn and Knowles, 2017). Pre-trained multilingual language models and techniques like multilingual training or fine-tuning are indeed effective for dealing with low-resource data such as poetry as seen in Figure 1 and 2 (More examples in Appendix A). However, it is surprising that even a model trained on 6M parallel lines from OPUS(100) performs worse than models trained on in-domain data that is 35X smaller.

Table 9 shows how model fine-tuned multilingually on non-poetic data suffer from loss of metaphoric expression in poetry, while a model fine-tuned multilingually trained on Poetic data is able to capture it. Table 10 shows how every model except our best poetic model fine-tuned multilingually suffer from hallucinations. The Non-Poetic model, while fluent to the reader, is not faithful to the original translation. Figures 1, 2, 3 show...
Gold:
I love the lavish withering of nature,
The gold and scarlet raiment of the woods,
The crisp wind rustling o’er their threshold,
The sky engulfed by tides of rippled glow,
The sun’s scarce rays, approaching frosts,
And gray-haired winter threatening from afar.

Poetic All:
I love the luxuriant decay of nature,
In their haylofts the wind’s noise and fresh breath,
And the heavens are covered with wavy mist,
And the rare rays of sun, and the first frosts,
And threats of the distant gray-haired winter.

Multi(OPUS):
I love the lush nature of decay,
And forests clothed in purple and gold,
In their shadows is the sound of the wind, and the breath of fresh air,
And the heavens are covered with clouds,
And the rarest ray of sunshine, and the first frosts,
And distant threats of the gray winter.

Gold:
Round, frozen in its oceans, transparent
like a cell under the microscope
or horizontally with mountains planted firmly above fields
with the tongue of rivers and the stretched out sea.
Every now and then I have an inkling of vertigo:
we’re turning faster. Asleep, I cry out “I’m falling”
and then I feel space, blackness, the stars at the nape of my neck,
and there is an incision into a thousand spheres.

Poetic All:
Round, icy of its oceans, transparent
like a cell under the microscope
yet horizontally with mountains steadily resting on the meadows
with the tongue of rivers and the rising sea.
Only at times do I fear the vertigo;
we turn more swiftly. As we sleep we cry out
and there is the space, blackness, stars on the nape
The fright that vomits itself into a thousand spheres.

Multi(OPUS):
Round, icy of its oceans, transparent
as a cell under the microscope
but horizontally with mountains set high above the meadows
with the tongue of the rivers and the low sea.
I only sometimes suspect vertigo:
We spin faster. Sleeping I scream “cado”
And there I feel space, black, stars on my neck.
The scare that vomits itself into a thousand balls.

---

Figure 1: Example Russian-English translation

| Gold | Of a temple rising up in the gloom. |
| Poetic All | Of a temple that rises in the dark. |
| Non-Poetic | A temple rising in the twinkling of an eye |
| Gold | stand sails of smoke |
| Poetic All | stand the sails of the smoke |
| Non-Poetic | There are sails of smoke |

Table 9: Examples where metaphoric expressions are lost when translated using model fine-tuned multilingually on Non-Poetic (OPUS) data.

8 Related work

Domain adaptation in neural machine translation Chu and Wang (2018) categorize domain adaptation for NMT in two groups: data centric and model centric. Data centric techniques mostly focus on data augmentation for limited parallel corpora of low-resource languages. For example, Currey et al. (2017) propose copying the target data to the source side to incorporate monolingual training data for low-resource languages. Back-translation has been used for synthetic parallel corpora generation (Sennrich et al., 2016). To improve performance on specific domains, Chu et al. (2017) augment corpora with tags to indicate specific domains. A conventional model-centric approach is fine-tuning on in-domain parallel corpora or on mixed in-domain and out-of-domain corpora (Chu and Wang, 2018). In our work, we deal with a model-centric approach where we leverage a multilingual pre-trained model (mBART) and then fine-tune it multilingually on in-domain corpus. Recently, Hu et al. (2019) introduced a domain adaptation technique using lexicon induction, where large amounts of monolingual data are leveraged to find translations of in-domain unseen words. However, word-level lexicon induction might not be the most useful augmentation technique in our case, since poetic text deals with multi-word unseen phenomena, e.g. metaphors.

Poetic and literary translation Jones and Irvine (2013) discuss the difficulties of faithful machine translation of literary text in terms of the competing objectives of staying faithful to the original text but, on the other hand, trying to convey the experience of reading a literary piece to the reader.
Met steeds speelser gemak
sla ik de aanvallen af
op mijn zwaarbevochten onverschilligheid.
De hemelen . . .

You bring wine and yourself: my final weaknesses.
I'll satiate myself till repetition is no longer wanted,
and wash the dirt from my hard hands once more.

Poetic All:

With ever more playful ease
I counter the attacks
on my heavily stressed indifference.
The skies are dripping, with splendour and all,
the trees languidly let their sparkling decay hang.
No eye, no ear. A broad smile.
You bring wine and yourself: my last weaknesses.
I will wallow until repetition becomes uncomfortable,
and once again spoil the filth from my hard hands.

Multi(OPUS):

Easier and easier to play
I'll stop the attacks.
on my hard-fought inflexibility.
The heavens are falling down with splendor,
Let the trees hang heedless of their blazing rage.
No eye, no ear, a wide smile.
You bring wine and yourself: my last weaknesses.
I'll be content until repetition is desired,
And wash the dirt off my hard hands one more time.

Gold:

With ever greater playful ease
I dismantle the attacks
on my hard-won indifference.
The heavens fall back, glory and all,
trees let dangle in dejection their sparkling leaves.
No eye, no ear. A broad smile.
You bring wine and yourself: my final weaknesses.
I'll be content until repetition is desired,
And wash the dirt off my hard hands one more time.

Table 10: Table showing hallucinations by other Non-Poetic models including Google Translate, the ubiquitous translation behemoth.

| language     | translation                                                                 |
|--------------|------------------------------------------------------------------------------|
| Russian      | Медуницы и осы тяжелую розу сосут. Человек умирает. Песок остывает согретый. |
| NonPoetic    | The vines and the leaves are the heavy roses. A man dies. The sand is boiled down. |
| Poetic       | Honeycombs and wasps suck the heavy rose. Man dies. The heated sand cools. |
| Gold         | Bees and wasps suck the heavy rose. Man dies. The heated sand cools. |
| Google Translate | Lungs and wasps suck a heavy rose. The man is dying. Warmed sand is cooling |

Figure 3: Example Dutch-English translation

Besacier and Schwartz (2015) conduct a pilot study of how suitable an MT+PE (machine translation + post-editing) pipeline would be for literary translation, concluding that their SMT approach could produce “acceptable and rather readable” translations. Matusov (2019) found that adapting NMT systems to literary content leads to improved automatic evaluation metrics on literary prose as compared to general domain NMT systems. Kuzman et al. (2019) found that Google NMT outperformed bespoke NMT models tailored to literature for English-Slovene literary translations. Toral et al. (2020) perform a comprehensive human and automatic evaluation (using BLEU) of NMT using Transformers (Vaswani et al., 2017) for English-Catalan translation of novels, finding that domain-specific models lead to performance improvements judging by all evaluation techniques. Fonteyne et al. (2020) conduct a document-level evaluation of the translation of Agatha Christie’s novel from English to Dutch using Google’s NMT system and found that most frequent issues were incorrect translation, coherence, style, and register.

Even though a lot of work has been done in the direction of automatic literary translation, automatic poetry translation is still in its infancy. Genzel et al. (2010) produce poetry translations with meter and rhyme using phrase-based statistical MT approaches. Ghazvininejad et al. (2018) present a neural poetry translation system that focuses on form rather than meaning. They also only focus on poetry translation from French to English, and their code or data is not publicly available. In our work, the focus is on faithfulness and the ability to preserve figurative language in translation across multiple languages.

9 Conclusion and future work

We release poetic parallel corpora for 6 language pairs. Our work shows the clear benefit of domain adaptation for poetry translation. It further shows that improvements can be achieved by leveraging multilingual fine-tuning, and that the improvements transfer to unseen languages. Future directions include addition of new languages and larger corpora, adapting low-resource machine translation techniques for poetry translation, translating to languages that are morphologically richer than English, as well as working on better evaluation metrics to detect hallucinations. While computational methods for poetry translation may never outperform the human standard, we hope our contributions spark interest in the machine translation community to take up this rather challenging task. Additionally, by open-sourcing our work we hope to provide a helpful resource for professional translators.
Ethical Considerations

Although we use language models trained on data collected from the Web, which have been shown to have issues with gender bias and abusive language (Sheng et al., 2019; Wallace et al., 2019) even in the multilingual space (Zhao et al., 2020), the inductive bias of our models should limit inadvertent negative impacts. Unlike model variants such as GPT, mBART is a conditional language model, which provides more control of the generated output. Our poetic parallel corpora are unlikely to contain toxic text and underwent manual inspection by the authors.

Technological advances in machine translation have had both positive and negative effects. Translation technology can diminish translators’ professional autonomy as well as endanger professional translators’ livelihood. Moreover, given the fact that most people resort to models trained on non-literary text for literary translation could harms us in many ways. One such example is the negative influence caused by ungrammatical or unidiomatic language on readers’ linguistic skills in the target language, especially in the case of child readers. The low quality of literary translations can prevent the transfer of literary ideas and repertoires from one culture to another. Our work on poetry translation with a focus on faithfulness tries to bridge that gap. We believe interactive, human-in-the-loop MT systems designed especially for literary or poetic translation such as ours might speed up literary translators’ work and make it more enjoyable.

Like (Petrilli, 2014) we believe too that automation in linguistic work and translation will not lead to the passivation and exclusion of humans. Rather, it will increase human-machine interaction and continue enhancing human creativity. Finally, we want to acknowledge all human translators who posted their work open-sourced on the websites we collected the data from. For our train and validation splits, the poems were broken down line by line and shuffled randomly. They do not contain any metadata and as such cannot distribute the creative value of the original poems.

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.

Philip Arthur, Graham Neubig, and Satoshi Nakamura. 2016. Incorporating discrete translation lexicons into neural machine translation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1557–1567, Austin, Texas. Association for Computational Linguistics.

Laurent Besacier and Lane Schwartz. 2015. Automated translation of a literary work: A pilot study. In Proceedings of the Fourth Workshop on Computational Linguistics for Literature, pages 114–122, Denver, Colorado, USA. Association for Computational Linguistics.

Huimin Chen, Xiaoyuan Yi, Maosong Sun, Wenhao Li, Cheng Yang, and Zhipeng Guo. 2019. Sentiment-controllable chinese poetry generation. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, pages 4925–4931. International Joint Conferences on Artificial Intelligence Organization.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of simple domain adaptation methods for neural machine translation.

Chenhui Chu and Rui Wang. 2018. A survey of domain adaptation for neural machine translation. In Proceedings of the 27th International Conference on Computational Linguistics, pages 1304–1319, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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, Copenhagen, Denmark. Association for Computational Linguistics.

Liming Deng, Jie Wang, Hangming Liang, Hui Chen, Zhiqiang Xie, Bojin Zhuang, Shaojun Wang, and Jing Xiao. 2019. An iterative polishing framework based on quality aware masked language model for Chinese poetry generation.

Zi-Yi Dou, Antonios Anastasopoulos, and Graham Neubig. 2020. Dynamic data selection and weighting for iterative back-translation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 5894–5904, Online. Association for Computational Linguistics.

Margot Fonteyne, Arda Tezcan, and Lieve Macken. 2020. Literary machine translation under the magnifying glass: Assessing the quality of an NMT-translated detective novel on document level. In Proceedings of the 12th Language Resources and Evaluation Conference, pages 3790–3798, Marseille, France. European Language Resources Association.
Piji Li, Haisong Zhang, Xiaojiang Liu, and Shuming Shi. 2020. Rigid formats controlled text generation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 742–751. Online. Association for Computational Linguistics.

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.

Evgeny Matusov. 2019. The challenges of using neural machine translation for literature. In Proceedings of the Qualities of Literary Machine Translation, pages 10–19, Dublin, Ireland. European Association for Machine Translation.

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 on association for computational linguistics, pages 311–318. Association for Computational Linguistics.

Susan Petrilli. 2014. Sign studies and semioethics: Communication, translation and values, volume 13. Walter de Gruyter GmbH & Co KG.

Jason Phang, Thibault Févry, and Samuel R Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. arXiv preprint arXiv:1811.01088.

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.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1172–1183, Online. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2685–2702. Online. Association for Computational Linguistics.

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

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. 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 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. arXiv preprint arXiv:2008.00401.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12), pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Antonio Toral, A. Oliver, and Pau Ribas Ballest’in. 2020. Machine translation of novels in the age of transformer. ArXiv, abs/2011.14979.

David Uthus, Maria Voitovich, and RJ Mical. 2021. Augmenting poetry composition with verse by verse. arXiv preprint arXiv:2103.17205.

Tim Van de Cruys. 2020. Automatic poetry generation from prosaic text. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 2471–2480.

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

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing NLP. 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 2153–2162, Hong Kong, China. Association for Computational Linguistics.

Hao-Ran Wei, Zhirui Zhang, Boxing Chen, and Weihua Luo. 2020. Iterative domain-repaired back-translation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 5884–5893, Online. Association for Computational Linguistics.

John Wieting, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig. 2019. Beyond BLEU: training neural machine translation with semantic similarity. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4344–4355, Florence, Italy. Association for Computational Linguistics.
Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Cheng Yang, Maosong Sun, Xiaoyuan Yi, and Wenhao Li. 2018. Stylistic Chinese poetry generation via unsupervised style disentanglement. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3960–3969, Brussels, Belgium. Association for Computational Linguistics.

Xiaoyuan Yi, Ruoyu Li, Cheng Yang, Wenhao Li, and Maosong Sun. 2020. Mixpoet: Diverse poetry generation via learning controllable mixed latent space.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Jieyu Zhao, Subhabrata Mukherjee, Saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. 2020. Gender bias in multilingual embeddings and cross-lingual transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2896–2907, Online. Association for Computational Linguistics.