The Unreasonable Effectiveness of Few-shot Learning for Machine Translation

Xavier Garcia [1]   **Yamini Bansal** [1]   **Colin Cherry** [2]   **George Foster** [2]   **Maxim Krikun** [2]
Melvin Johnson [1]   **Orhan Firat** [1]

Abstract

We demonstrate the potential of few-shot translation systems, trained with unpaired language data, for both high and low-resource language pairs. We show that with only 5 examples of high-quality translation data shown at inference, a transformer decoder-only model trained solely with self-supervised learning is able to match specialized supervised state-of-the-art models as well as more general commercial translation systems. In particular, we outperform the best performing system on the WMT'21 English–Chinese news translation task by only using five examples of English–Chinese parallel data at inference. Furthermore, the resulting models are two orders of magnitude smaller than state-of-the-art language models. We then analyze the factors which impact the performance of few-shot translation systems, and highlight that the quality of the few-shot demonstrations heavily determines the quality of the translations generated by our models. Finally, we show that the few-shot paradigm also provides a way to control certain attributes of the translation — we show that we are able to control for regional varieties and formality using only five examples at inference, paving the way towards controllable machine translation systems.

1. Introduction

Current state-of-the-art machine translation systems are typically built by leveraging vast amounts of parallel data mined from the web. While this is practical for high-resource language pairs, it is unfeasible to obtain corpora of such sizes for the majority of languages in the world. Moreover, the reliance of mined parallel data has many potential downsides, such as allowing for poisoning attacks (Xu et al., 2021;

Wang et al., 2021a), memorization of low-quality examples (Raunak & Menezes, 2022), and biases towards generating text in language registers over-represented in the parallel data, such as language varieties (Lakew et al., 2018; Riley et al., 2022) or formality (Rippeth et al., 2022).

As an alternative, researchers began exploring the task of *unsupervised translation* (Ravi & Knight, 2011) i.e. building translation models without any parallel data at all. Unsupervised translation systems have demonstrated promising performance in recent years, able to match strong supervised baselines on academic benchmarks (Song et al., 2019; Garcia et al., 2020b; Han et al., 2021) by relying on a collection of tricks and techniques, such as multilinguality (Conneau et al., 2019; Garcia et al., 2020b;a; Lin et al., 2021), back-translation (Lample et al., 2017), and most recently through large-scale models with parameters in the hundreds of billions (Chowdhery et al., 2022; Vilar et al., 2022; Han et al., 2021). Despite such results, these systems are rarely compared to state-of-the-art supervised models which also leverage their own bag of tricks to improve performance.[1] Moreover, these models trained on large-scale monolingual data also suffer from similar issues as models trained on large-scale parallel data: variance in the quality of monolingual data and generation being biased towards the over-represented registers in the monolingual data.

It thus remains to explore what lies between these two research streams: *few-shot learning* (Brown et al., 2020). Recently, large language models have been shown capable of performing arbitrary tasks by exposing a few demonstrations of the task at inference time. The dependence of these models on demonstrations to perform the task allows us to overcome the aforementioned issues by carefully selecting a small set of high-quality translation pairs in the language register of interest as demonstrations.

In this work, we demonstrate that the few-shot translation paradigm allows us to build high-fidelity translation models at a smaller scale (8B parameters) than traditional large language models ($> 100B$ parameters) without the need for

[1]Google DeepMind [2]Google Translate. Correspondence to: Xavier Garcia <xgarcia@google.com>, Orhan Firat <orhanf@google.com>.

[1]For example, Tran et al. (2021) applies model ensembling, reranking strategies, language-specific post-processing and fine-tuning on in-domain data for all language pairs, as well as iterative back-translation for the Icelandic language pairs.

Language	Examples
English	69,813,325
German	69,813,325
Chinese	33,172,846
Icelandic	250,416

Table 1. **Number of processed examples per language.** Here by *example*, we mean the tokenized inputs that are being passed to the model. We describe the process used to create these examples in Section 2.2.

Language Pair	Forward	Backward
English-German	0.9%	1.5%
English-Chinese	3.0%	1.5%
English-Icelandic	0.3%	1.7%

Table 2. **Percent overlap between the references of each language pair and and the monolingual data.** We follow the same 15-gram protocol as Chowdhery et al. (2022).

back-translation or large-scale parallel text mining. We evaluate these models on the WMT'21 English–German and English–Chinese news translation task and show that they outperform commercial baselines, and show performance competitive with WMT'21 submissions, which themselves rely on many of the aforementioned techniques. We then verify that our approach works in low-resource scenarios by performing a similar study on the WMT'21 English–Icelandic language pair, where the amount of Icelandic monolingual data is two orders of magnitude smaller than Chinese or German. Furthermore, we show that constraining the demonstrations to be in a desired language register generally results in the output translation being part of the register. We show that this results in quantitative gains in translation benchmarks which account for language registers: the region-aware benchmark FRMT (Riley et al., 2022) and the IWSLT'22 Special Task on Formality Control for Spoken Language Translation (Anastasopoulos et al., 2022).

2. Experiments on high-resource languages & results

In this section, we outline the datasets, models, and initial set of experiments we perform in this work. We first discuss the composition of our monolingual data, as well as the evaluation datasets considered. Next, we describe the exact architecture we use for these studies, and how we train our models in terms of the particular objective being used and batching choice. Finally, we evaluate our few-shot translation models and compare their performance against a suite of state-of-the-art models.

All the experiments in this work were conducted using JAX (Bradbury et al., 2021), using the T5X framework (Roberts et al., 2022) and FLAX (Heek et al., 2020).

2.1. Datasets

Monolingual datasets   Our training data consists of a collection of language-specific corpora. For English, we use a similar mix of filtered web pages, Wikipedia, and books as done in Chowdhery et al. (2022). For every other language, we restricted ourselves to only high-quality webpages, using similar filters as the English data. The amount of data

obtained by this approach varies by language. We list the final number of examples after processing in Table 1. We provide an explicit description of the processing in Section 2.2. Collecting data in this process leads to much more English data than reported in our table. For simplicity, we artificially restricted the amount of data to be as much as the next highest-resource language, German.

Evaluation datasets   Most work in the unsupervised translation literature tends to focus on older WMT datasets e.g. WMT'14 English–French. This is problematic for us for a number of reasons: 1) the quality of WMT submissions has increased dramatically in recent years; 2) the quality of test sets has improved over the years; 3) because our pretraining data is derived from recent web crawls, there is a possibility of train / test overlap with previous years. For these reasons, we follow the recommendations outlined in Vilar et al. (2022) and use only recent test sets, coming from the WMT'21 news translation task. We primarily focus on English–German and English–Chinese language pairs, as these are typically high-resource language pairs, and thus we believe should have strong WMT submissions.

Train-test overlap   To account for potential train-test overlap, we follow the strategy used in previous work (Chowdhery et al., 2022; Vilar et al., 2022) to measure target-side test overlap based on n-gram matching. We use 15-grams, with the understanding that test sequences shorter than 15 tokens will count as a match if they are found as a substring in the training data. We report the degree of overlap in Table 2 and note that we do not see much overlap with the newer test sets.

2.2. Architecture and Training Procedure

Architecture   We use a Transformer (Vaswani et al., 2017) decoder-only architecture, using the same architecture modifications as in Chowdhery et al. (2022). We use the exact hyperparameter configurations as their 8 billion parameter model for our main experiments. In particular, we use 32 Transformer layers, with 16 heads, a hidden dimension of 4096, and multi-query attention. The feed-forward size is 16384 and the attention head size is 256.

Vocabulary   We use a Sentencepiece (Kudo & Richardson, 2018) model that was built in a similar fashion to the one used in Chowdhery et al. (2022), with the notable difference that we use 128,000 Sentencepieces instead 256,000. In particular, we use the same Sentencepiece model for all the models trained in this paper.

Data pre-processing and training objective   In this work, we use a variant of the UL2 objective (Tay et al., 2022) that has been specialized for decoder-only models. The original UL2 objective relies on a mixture of 6 difference instances of the span corruption objective (Raffel et al., 2020) using different hyperparameter configurations, coupled with a prefix language modeling (Raffel et al., 2020). The span corruption instances are completely determined by two hyperparameters: *noise density*, which controls how much of the input is corrupted and *mean noise span length*, which controls the average number of tokens corrupted per span. In this work, we use 2 (instead of 6) separate span corruption instances with (noise density, mean noise span length) given by (0.15, 3) and (0.5, 32) respectively. In addition to these two objectives and the prefix language modeling objective, we also include a standard causal language modeling objective. We mix these objectives randomly, sampling prefix language modeling 20% of the time, causal language modeling 60% of the time, and the remaining span corruption instances 20% of the time.[2]

Trilingual models   We will be primarily focused on models exclusively supporting two languages at a time. However, to access the potential value of multilinguality in a more controlled setting, we will also consider *trilingual models*, supporting three languages at a time. To establish a fair comparison between our trilingual and bilingual models, we need to enforce some constraints which should match practical use-cases. We enforce the constraint that the trilingual models must see the same amount of data per-language as the bilingual models i.e. the only difference in training data between the trilingual and bilingual models will be the additional data from the third language. While this choice results in more compute spent on training the trilingual models, it does not see any more data in any two languages compared to the analogous bilingual counterpart.

Training distribution   For each set of languages considered, we perform one epoch over the combined corpus. This results in longer training for the English–German model than the English–Chinese model, due to the presence of more German data than Chinese data. We will assess the

potential value of multi-epoch training in Section 2.4.

Training hyperparameters   We use a maximum sequence length of 2048, with a batch size of 1024. We use the Adafactor optimizer (Shazeer & Stern, 2018), without using the factorizing option. We use a cosine learning rate decay schedule (Hoffmann et al., 2022), starting at 0.01 and ending at 0.001 at the end of training. This results in 98000 steps for the English-Chinese model, 135000 steps for the English-German model, and 166000 steps for the trilingual model.

Evaluation   Our initial attempts at evaluating our models using BLEU (Papineni et al., 2002) heavily underestimated the performance of our models. Vilar et al. (2022) has shown that translations generated from few-shot translation through large language models produce qualitatively different translations, with different kinds of errors from the ones generated by traditional machine translations models. We believe such errors are overly punished by the brittleness of n-gram metrics such as BLEU. Recent work has shown that such n-gram metrics are sub-optimal for evaluating high-quality translations (Kocmi et al., 2021; Freitag et al., 2021). For this reason, we use the learnt metric BLEURT (Sellam et al., 2020) as our main metric to assess quality. We follow the recommendations from the publicly-available Github page and use the BLEURT-20 checkpoint.[3] For completeness and full transparency, we report BLEU scores in the appendix, in Section 4.

Few-shot generation   To condition the model to perform translation, we use the relevant development set for each language pair as a pool of demonstrations. For each source example `text` in a given test set, we randomly sample 5 sentence pairs $(x_1, y_1), ..., (x_5, y_5)$ from the development set and we insert these objects into the following template, where `source_language` and `target_language` are the names of source and target languages for the given language pair:

$$\{\texttt{source\_language}\} : \{x_1\}$$
$$\{\texttt{target\_language}\} : \{y_1\}$$
$$\cdots$$
$$\{\texttt{source\_language}\} : \{x_5\}$$
$$\{\texttt{target\_language}\} : \{y_5\}$$

$$\{\texttt{source\_language}\} : \{\texttt{text}\}$$
$$\{\texttt{target\_language}\} :$$

To generate the predictions, we use minimum Bayes risk (MBR) decoding (Eikema & Aziz, 2020; Freitag et al.,

---

[2]The exact hyperparameter choices for this preprocessor were chosen following recommendations from the authors of Tay et al. (2022). Early experiments showed that it performed better than just using causal language modeling and we did not explore the choice of hyperparameters any further.

[3]https://github.com/google-research/bleurt#checkpoints

Rank	$en \rightarrow zh$	$zh \rightarrow en$	$en \rightarrow de$	$de \rightarrow en$
1st	Zeng et al. (2021)	Wang et al. (2021b)	Tran et al. (2021)	Tran et al. (2021)
2nd	Tran et al. (2021)	Zhou et al. (2021)	Qian et al. (2021)	Online−W
3rd	Zhou et al. (2021)	Li et al. (2021)	Online−W	Subramanian et al. (2021)

Table 3. **WMT baselines for the high-resource language pairs we consider in this paper.** Note that although we cite the original paper, we recompute the metrics from the their textual outputs, rather than rely on the numbers provided in the paper to ensure fair comparison.

Model	$en \leftrightarrow zh$ newstest21		$en \leftrightarrow de$ newstest21	
Supervised baselines				
WMT'21 1st Place	**70.0**	66.6	**76.9**	**76.9**
WMT'21 2nd Place	69.7	66.3	76.3	76.7
WMT'21 3rd Place	69.7	65.8	76.0	76.4
Google Translate	69.5	65.0	76.4	75.7
Few-shot translation models				
PaLM	67.7	64.1	<u>75.9</u>	74.8
Bilingual LMs (Beam)	62.6	67.0	74.9	74.1
Bilingual LMs (MBR)	68.4	67.8	75.5	76.5
Trilingual LM (Beam)	65.3	65.3	74.5	74.4
Trilingual LM (MBR)	<u>68.9</u>	**68.3**	75.5	<u>76.8</u>

Table 4. **BLEURT scores from various models, both supervised and few-shot on some WMT *newstest21* sets.** We italicized the name of our baselines, and bolded the best performing results. We also underline the best performing few-shot results. We use the suffix *Beam* when using beam search, and *MBR* when using MBR decoding.

2022), using learnt metrics and 64 sampled predictions with vanilla ancestral sampling. We use BLEURT for the utility function in MBR. We also include results with beam search, with beam size 4 and α=0.6.

2.3. Main results and discussion

Baselines We consider the top three performing systems from WMT'21 for each language pair for comparisons. We provide a reference for each such system considered in Table 3. One potential pitfall of this approach is that WMT submissions have been hyper-specialized for this particular domain and evaluation procedure, and thus may overestimate the performance of general-purpose translation systems built to handle all domains. Since we are not focusing on making our translation models WMT-specific, we also include strong baselines of more general-purpose systems: *PaLM* (Chowdhery et al., 2022), a large multilingual (albeit English-centric) language model; *Google Translate*, as an example of a commercial system which was not fine-tuned for WMT or any other such competition.[4]

We list the performance of our models (labelled as *Bilingual*

LMs and *Trilingual LM*) as well as the baselines on our dataset in Table 4. To ensure fair comparison, we computed the BLEURT score directly from the text predictions of the baselines, as we did with our models. For PaLM, we used the predictions with the best performing metrics from Vilar et al. (2022).[5] We first note that with the exception of German-English, both our language models outperform PaLM in this task when using MBR, despite having less than 2% the number of parameters of PaLM. We believe this is in great part due to the larger quantity of monolingual data for the non-English languages seen by our models during training. This is especially highlighted in the out of English translation pairs (English–XX pairs.)

We also remark that most WMT baselines outperform the commercial system. As we noted earlier, this is most likely due to the fact that these baselines have been specialized for this WMT competition, while the commercial systems have to handle a more broader range of domains. When comparing the commercial system to our few-shot translation models, we note that our models excel in the English–XX directions despite also not being specialized for WMT. Moreover, our models are also able to outperform one of the strongest WMT baselines: for example, the trilingual model outperforms the best performing English–Chinese model from WMT'21, while being 0.1 BLEURT away from the best English–German system. The XX–English direction tells a different story: in this setting, we see our few-shot translations underperform against all our supervised baselines. Finally, we note that MBR consistently provides improvements in BLEURT. As such, we default to MBR for the remaining experiments.

2.4. Performance on a low-resource language

Finally, we consider whether this approach also yields high-quality translation models for low-resource languages. Building translation systems for such language pairs typically require leveraging data beyond the parallel data available for the language pair, whether in the form of mono-

---

[4]We re-used the translations from (Vilar et al., 2022) when available. For the Icelandic pair, we obtained the predictions using the public API at translate.google.com in December 2022.

[5]The PaLM numbers in Table 4 sometimes use different demonstrations from the ones we used. All our results draw demonstrations from the same development sets as in Vilar et al. (2022), which they refer to as *WMT-dev*. We only report PaLM numbers from Vilar et al. (2022) using different demonstrations if they are better than the ones obtained using WMT-dev as the source of demonstrations.

Rank	$en \to is$	$is \to en$
1st	Tran et al. (2021)	Tran et al. (2021)
2nd	Zhou et al. (2021)	Online−B
3rd	MANIFOLD	Zhou et al. (2021)

Table 5. **WMT baselines for the Icelandic language pairs.** Note that although we cite the original paper, we compute the metrics from the their textual outputs, since many of these works do not report BLEURT numbers.

Model	$en \leftrightarrow is$ newstest21	
Supervised models		
WMT'21 1st Place	**77.2**	**76.1**
WMT'21 2nd Place	74.3	72.3
WMT'21 3rd Place	74.3	70.4
Google Translate	76.8	71.1
Few-shot translation models		
PaLM	61.7	59.5
Bilingual LMs (MBR)	76.2	72.0

Table 6. **BLEURT scores from various models, both supervised and fewshot on the WMT** *newstest21* **test sets involving Icelandic** (*is*). To obtain these models, we fine-tuned the English-German models from Section 2 on a mixture on English and Icelandic, for a few epochs over the Icelandic monolingual data.

lingual data or parallel data for other language pairs. It is conceivable that the approach considered in this paper might not be as successful for languages with smaller amounts of monolingual data, as the value of parallel data could be magnified in these resource-constrained settings (Kirstain et al., 2021).

We could attempt to study this problem by artificially constraining the amount of monolingual data for German or Chinese. However, previous works (Kim et al., 2020; Marchisio et al., 2020) have shown that datasets for authentic low-resource languages offer additional difficulties that do not arise in datasets for high-resource languages, such as lower quality of data available, limited domain coverage, little language similarity with English, rich morphology, etc. To avoid this pitfall, we consider *Icelandic* as our low-resource language of study.[6] Icelandic is also convenient for us since it is one of the languages available in WMT'21, and thus we have strong baselines available in the form of WMT submissions. As can be seen from Table 1, Icelandic has two orders of magnitude less data than German or Chinese and hence poses a far more difficult challenge for our models.

Warm-starting Training models with low-resource languages can be a difficult endeavor due to issues regarding overfitting and imbalanced datasets. Previous work on transfer learning in language modeling shows that continued training on a new distribution yields similar results to training from scratch, albeit at a much faster rate (Hernandez et al., 2021). In this work, we follow this approach and fine-tune our English–German model. This allows us to leverage all the English knowledge our model has already accumulated during its original training run, which is critical due to the English-centric nature of our evaluation.

Multi-epoch training Previous literature on language modeling has found some success with repeated epochs over the training data (Taylor et al., 2022). We explore this question by training on a equal mixture of English and Icelandic until we go 8 epochs over the Icelandic dataset and evalu-

ate the current checkpoint at the end of every epoch. We evaluate both the English–Icelandic and Icelandic–English directions on the Flores (Goyal et al., 2022) *devtest* sets for Icelandic, average the two BLEURT scores, then take the best-performing checkpoint and evaluate it on the WMT *newstest21* dataset. We select the optimal checkpoint occurring after 6 epochs over the Icelandic dataset.

Training hyperparameters We re-use all the same hyperparameters as when training the high-resource language models, with the exception of the learning rate: we set it to constant at 0.001, following recommendations from previous works (Xue et al., 2020).

Comparison with baselines We list our WMT baselines in Table 5 and show the results on both directions in Table 6. We see that for both directions, the results are competitive with the WMT baselines, and in fact surpass the commercial baseline for the English-Icelandic direction. Moreover, we see that PaLM fails to provide high-quality translations as measured by BLEURT.

Results of multi-epoch training We present the results of evaluating at each epoch over the Icelandic datasets in Figure 1. We see a few patterns emerge. First, for both directions, doing a single-pass is suboptimal. Second, the Icelandic-English direction doesn't benefit beyond two epochs over the Icelandic dataset, while the reverse direction benefits for up to 6 epochs. This suggests that models develop the ability to extract meaning from text much earlier than when they can reliably generate fluent text. Moreover, it also suggests that we may be underestimating the potential performance of our bilingual models if we had extended their training to include multiple epochs over the non-English datasets. Finally, we note that our models remain fairly competitive with the WMT'21 submissions and in fact outperform the commercial system in the English–Icelandic direction.
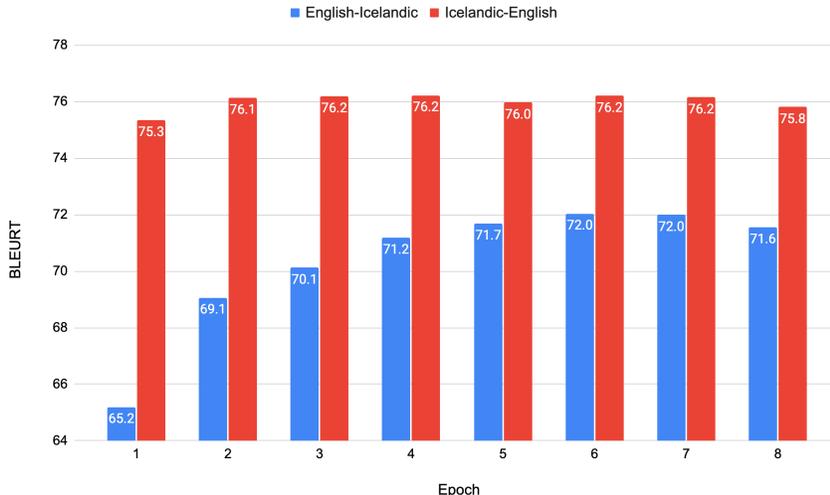
---

[6]Icelandic exhibits many of the properties typically associated with low-resource languages. For example, Joshi et al. (2020) places Icelandic in the same category as Hausa, which WMT'21 considers to be low-resource.

Figure 1. **BLEURT scores on the WMT'21 English–Icelandic and Icelandic–English task after several epochs over the Icelandic monolingual data.** We remark that while the Icelandic-English direction converges fairly quickly (after two epoch), the English-Icelandic direction benefits from further training.

3. Influence of few-shot demonstrations

From previous experiments, it is clear that few-shot translation models can be competitive with state-of-the-art supervised models. In this section, we dissect which aspects of the few-shot demonstrations impact the results. In this first subsection, we claim that one of the most influential predictors for the quality of our translations are the quality of the few-shot demonstrations. Next, we demonstrate that *style* of the few-shot demonstration also influence the style of the translation, granting users the ability to output arbitrary styles at inference by only requiring a few demonstrations exhibiting that style and with no additional fine-tuning costs. We additionally show that using the incorrect style can result in a measurable regression in BLEURT, further highlighting the need for such systems. To investigate this, we look at two specific use-cases: controlling for Mandarin regional varieties, and formality in German.

3.1. Quality of the demonstrations strongly influences quality of generated translations

Concurrent work (Agrawal et al., 2022; Vilar et al., 2022) has shown that the quality of the few-shot demonstration impacts the quality of the output. We provide a similar analysis of this claim, by using a direct quantitative signal of the quality of the exemplar.

For this experiment, we diverge from our typical protocol of using development sets as our source of exemplars. Instead, we use an English–German translation dataset, prepared in Bansal et al. (2022). This is a filtered version of the Paracrawl English–German dataset (Bañón et al., 2020),

where de-duplication has been applied, with both length and language identification filtering. This dataset also comes equipped with normalized Contrastive Data Selection (CDS) scores (Wang et al., 2018). These scores are at the example-level, and are calculated by computing the difference in cross-entropy scores between a translation model fine-tuned on a trusted dataset and one that was not. We should expect examples with higher CDS score to be increasingly different from the trusted data, and as such potentially of lower quality. The scores have been normalized to be between 0.0 and 1.0. We partition the examples into 3 buckets, where an example with a CDS score between 0.0 and 0.33 gets placed in the first bucket, an example with a CDS score between 0.33 and 0.66 gets placed in the second bucket, and the remaining examples go in the third bucket. By viewing CDS scores as a proxy for quality, we can then examine the impact of quality of the demonstrations by constraining our demonstrations to come from a given bucket. We use the same decoding strategy as in our previous experiments, and report the BLEURT scores for each bucket in Figure 2. We note a pattern emerges: in general, as we use demonstrations with increasing CDS scores, the quality of the translations produced by our model keeps degrading.

3.2. Controllability of output language variety through few-shot demonstrations

We now show that example quality is not the only attribute of the demonstrations which can heavily influence the quality of the outputs. We now claim that the *style* of the demonstrations also influences the output in a quantitively-measurable way. This is exciting since it establishes the possibility of
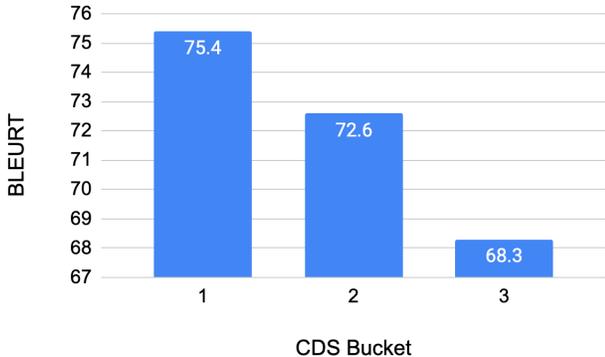
Figure 2. **BLEURT scores on the WMT'21 English–German task, using demonstrations drawn of varying quality.** For each CDS bucket, we evaluate the model on the WMT'21 English–German task.

Model	FRMT Score	Lexical Accuracy
FRMT Baselines		
PaLM 8B	58.3	69.0
PaLM 62B	65.1	70.8
PaLM 540B	68.4	83.6
Gold	–	94.4
Ours		
Bilingual LM (mismatched)	66.0	13.8
Bilingual LM (matched)	**71.0**	**86.2**

Table 7. **Performance of various models on the Mandarian regional varieties datasets from FRMT.** We use the term *matched* and *mismatched* for the cases when we use demonstrations from the same or different language variety as the references, respectively. We bold the best results which do not include the reference. Note that Gold has no FRMT score, since this score is computed using the Gold predictions as references.

generating translations in a given style, such as formality-aware controlled translation (Niu et al., 2017; 2018; Garcia et al., 2021) or targeting specific language varieties (Lakew et al., 2018; Kumar et al., 2021). Moreover, since the style is provided at inference, this allows us to condition the generation on arbitrary styles using only a very small set of high-quality translations exhibiting this attribute.

For this analysis, we use the FRMT dataset (Riley et al., 2022), which contains high-quality translations between English and two Mandarin regional varieties: Mainland and Taiwanese. We focus on two metrics: *FRMT Score*, which consists of a geometric mean across regional varieties of the arithmetic means of the bucket-level BLEURT scores. As a proxy to measure whether the correct language variety is being used, we use a metric introduced in Riley et al. (2022) known as *lexical accuracy*. To compute lexical accuracy, we follow the approach proposed by Riley et al. (2022) and use a special subset of the data (denoted in Riley et al. (2022) as the *lexical bucket*) which has examples containing English terms that have different translations in the different regional varieties and check how often do the translations match the target language variety.

As in the previous experiments, we use the corresponding development sets to draw demonstrations for the translation task.[7] To highlight the importance of using demonstrations in the right language variety, we evaluate our models in two ways: *matched*, in which we use the same language variety for the demonstrations as the reference; *mismatched*, where we use different regional varieties for the demonstrations and the references. We compare against PaLM at vari-

ous sizes, where the 540B model was the best-performing models in Riley et al. (2022). We also include the lexical accuracy as human performance, which we refer as *Gold*.

We report the results in Table 7. We first note that even in the mismatched settings, our models still outperform the PaLM-62B model in terms of FRMT score, confirming the fidelity of the translations. However, we see that the lexical accuracy for the mismatched setting is very low, suggesting that the model is outputting valid translations but not in the target language variety. Once we use demonstrations showcasing the appropriate language variety, both the FRMT score and lexical accuracy drastically improve, establishing a new state-of-the-art, despite being two orders of magnitude smaller than the previous state-of-the-art.

3.3. Controllability of formality through few-shot demonstrations

Finally, we consider the task of generating translations satisfying a given formality level. While the formality of text is in general a subjective matter in English, other languages have built-in rules for expressing formality. For example, many languages exhibit a form of the T–V distinction (Brown & Gilman, 1960), where the second person pronoun has two surface forms, an informal and a formal one. These pronouns usually also induce different verb conjugations, so even if they are not explicitly present in the sentence, one can still detect whether the intended style is informal or formal.

In particular, we focus on the English–German language pair of the IWSLT'22 Special Task on Formality Control for Spoken Language Translation.[8] We evaluate our models using both BLEURT and accuracy on whether the correct

---

[7]In addition to these demonstrations, we also use the names *Mainland Mandarin* and *Taiwanese Mandarin* for the respective regional varieties.

[8]https://iwslt.org/2022/formality

formality was used. To eliminate any discrepancies between our evaluation and the ones done by the organizers, we use the same evaluation scripts used by the official task for the formality accuracy.[9] As in the language variety experiments, we consider both the *matched* and *mismatched* setting, where the demonstrations exhibit the same (respectively, different) formality level as the target set. We draw our demonstrations from the *topical chat* split of the train set.[10]

Baselines   We consider Google Translate, which has no built-in feature for controlling formality. We also consider the `gold finetuned mBART-large` UMD submission (Rippeth et al., 2022) for the task.[11] We choose this submission from UMD as it has the highest translation quality compared with UMD's other submissions.[12]

We report the BLEURT scores and formality accuracy in Table 8. We first note that even though we took UMD's submission with the highest translation quality, it still attains lower quality than all the other models, including our mismatched models. However, UMD's submission exhibits higher formality accuracy compared to the commercial baseline, suggesting that there might be a trade-off in controlling formality versus quality. Our few-shot translation models achieve the best of both worlds: they exhibit high-quality translations, outperforming all available models, as well as achieving excellent formality control, obtaining higher accuracy than the UMD submission on the informal test set, despite only seeing a five examples at inference. Moreover, we do see increased BLEURT scores using the matched exemplars over the mismatched ones, suggesting that BLEURT is sensitive to formality.

4. Limitations and Future Work

Despite our strong results, we note that there are a few limitations of the approach described in this work which we would like to address in future work.

Low performance on n-gram based metrics   We list out the BLEU scores for the WMT'21 language pairs, computed

---

[10]The choice of using the *topical chat* split over *telophony* split of the training set was arbitrary.

[11]The explicit text predictions can be found here: https://github.com/amazon-science/contrastive-controlled-mt/tree/main/IWSLT2022/submissions/formality-control_UMD/EN-DE/unconstrained/formality/blind-test

[12]There are other submissions with better formality control than the submission we selected. However, we believe translation quality should be prioritized over control of stylistic features such as formality and hence we selected this submission over the alternatives.

Model	BLEURT		Accuracy	
	Formal	Informal	Formal	Informal
External Baselines				
Google Translate	72.2	72.5	62.7%	37.3%
UMD	69.4	69.3	**93.6%**	77.4%
Ours				
Mismatched	72.1	72.2	15.1%	15.5%
Matched	**73.5**	**74.7**	84.9%	**85.5%**

Table 8. **Performance of various models on the IWSLT2022 Formality Control for Spoken Language Translation.** We use the term *matched* (respectively, *mismatched*) for the cases when we use demonstrations from the same (respectively, different) formality as the references. We bold the best results and italicize the name of our models.

using SacreBLEU, in Table 9.[13] We note that there is a big gap between the WMT models and our models. On one hand, Freitag et al. (2022) reported that BLEU drops sharply when using MBR. However, we also find this pattern even when using beam search, suggesting that few-shot translation models naturally produce translations which are qualitatively different from traditional supervised models. Future work should further study the style of the translations obtained by few-shot models.

Model size   While we were able to outperform large language models like PaLM while using only a fraction of the number of parameters, our resulting models are still an order of magnitude larger than traditional machine translation systems, which usually have hundreds of millions of parameters. Preliminary experiments at the 1 billion parameter scale yielded unsatisfactory results, which is why we restricted our experiments to 8 billion parameter scale. Such size limitations may inhibit adoption of these techniques and future work should aim at lowering these size requirements.

Multilinguality   Our initial trilingual experiments showed that our models benefited from multilinguality. However, previous work in the multilingual literature has shown that adding more languages eventually leads to a decrease in per-language performance, a phenomenon known as the *curse of multilinguality* (Conneau et al., 2019). Future work should explore how many languages can such systems support until we encounter performance degradation.

Incorporating larger amounts of high-quality parallel data   In Section 3.1, we demonstrated that one of the critical properties of parallel data that we needed was *quality*, in contrast to the current popular approaches used in supervised models that mainly rely on *quantity*. However, our approach only utilizes five examples at inference, even

---

Models	*zh ↔ en* newstest21		*de ↔ en* newstest21		*is ↔ en* newstest21	
Supervised models						
WMT'21 1st Place	33.4	36.9	41.9	42	41.7	33.3
WMT'21 2nd Place	31.9	35.9	39.7	43.2	40	30.6
WMT'21 3rd Place	32.6	35.8	40	41.3	39.2	28.6
Google Translate	32.2	36.2	40.9	39.8	41.5	28.7
Few-shot translation models						
PaLM	25.8	29.6	38.8	32.9	19.1	16.87
Bilingual LM (MBR)	21.6	25.6	37.1	29.4	33.52	17.7
Bilingual LM (Beam)	20.4	29.2	35.5	32.8	36.2	19.2
Trilingual LM (MBR)	22.2	26.8	36.1	28.5	–	–
Trilingual LM (Beam)	20.45	25.5	36.2	31.8	–	–

Table 9. **BLEU scores from various models, both supervised and few-shot on some WMT *newstest21* sets.** We italicized the name of our baselines. We use the suffix *Beam* when using beam search, and *MBR* when using MBR decoding. Note that we have no scores for the trilingual models on the Icelandic pairs since we didn't train any such models.

though one could potentially get several hundreds of high-quality examples. If we were to obtain such a dataset and fine-tune our models using traditional approaches, we might obtain higher-quality translation models, but then we might also lose the exciting controllability that we get through the few-shot translation paradigm as shown in Sections 3.2 and 3.3. Future work should look into developing approaches that can leverage larger sets of high-quality data while still retaining the flexibility of the few-shot translation paradigm.

Better curation of the pre-training corpus Our current models are able to perform translation in large part due to the pre-training corpus. Such corpora, however, have potentially many issues which could impact translation. While we showed that we could mitigate the impact of many of these factors through careful selection of the few-shot demonstrations, future work should study more styles and domains to ensure the robustness of this approach and potential failure modes.

Decoding algorithm To get our best results, we relied on MBR decoding rather than beam search. We believe this to be a fair comparison with the WMT submissions, since they also rely on various reranking strategies. On the other hand, Vilar et al. (2022) did not rely on MBR decoding for their results, and we should expect PaLM's translation quality to improve when using MBR. Moreover, using these kind of decoding algorithms incurs a larger cost for serving than beam search, further inhibiting adoption. Future work should try to recover similar performance without the need for expensive decoding algorithms.

5. Conclusion

In this work, we investigated the potential value of few-shot translation models by examining their performance against the strongest supervised baselines we could find on traditionally high-resource language pairs. We show that these systems can be competitive with supervised models under the assumption that we have a small set of high-quality demonstrations. We analyze how the quality of these demonstrations heavily influences the quality of the translations generated by our models. Finally, we demonstrate that this paradigm also gives a way of controlling the style of the translations generated by these models.

References

Agrawal, S., Zhou, C., Lewis, M., Zettlemoyer, L., and Ghazvininejad, M. In-context examples selection for machine translation. *arXiv preprint arXiv:2212.02437*, 2022.

Anastasopoulos, A., Barrault, L., Bentivogli, L., Zanon Boito, M., Bojar, O., Cattoni, R., Currey, A., Dinu, G., Duh, K., Elbayad, M., Emmanuel, C., Estève, Y., Federico, M., Federmann, C., Gahbiche, S., Gong, H., Grundkiewicz, R., Haddow, B., Hsu, B., Javorský, D., Kloudová, V., Lakew, S., Ma, X., Mathur, P., McNamee, P., Murray, K., Nădejde, M., Nakamura, S., Negri, M., Niehues, J., Niu, X., Ortega, J., Pino, J., Salesky, E., Shi, J., Sperber, M., Stüker, S., Sudoh, K., Turchi, M., Virkar, Y., Waibel, A., Wang, C., and Watanabe, S. Findings of the IWSLT 2022 evaluation campaign. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pp. 98–157, Dublin, Ireland (in-person and online), May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.iwslt-1.10. URL `https://aclanthology.org/2022.iwslt-1.10`.

Bañón, M., Chen, P., Haddow, B., Heafield, K., Hoang, H., Esplà-Gomis, M., Forcada, M. L., Kamran, A., Kirefu, F., Koehn, P., et al. Paracrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual*

Meeting of the Association for Computational Linguistics, pp. 4555–4567, 2020.

Bansal, Y., Ghorbani, B., Garg, A., Zhang, B., Cherry, C., Neyshabur, B., and Firat, O. Data scaling laws in nmt: The effect of noise and architecture. In *International Conference on Machine Learning*, pp. 1466–1482. PMLR, 2022.

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., et al. Jax: Autograd and xla. *Astrophysics Source Code Library*, pp. ascl–2111, 2021.

Brown, R. and Gilman, A. The pronouns of power and solidarity. In Sebeok, T. A. (ed.), *Style in Language*, pp. 253–276. MIT Press, Cambridge, Mass, 1960.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.

Eikema, B. and Aziz, W. Is MAP decoding all you need? the inadequacy of the mode in neural machine translation. *CoRR*, abs/2005.10283, 2020. URL https://arxiv.org/abs/2005.10283.

Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q., and Macherey, W. Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474, 12 2021. ISSN 2307-387X. doi: 10.1162/tacl_a_00437. URL https://doi.org/10.1162/tacl_a_00437.

Freitag, M., Grangier, D., Tan, Q., and Liang, B. High quality rather than high model probability: Minimum bayes risk decoding with neural metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825, 2022.

Garcia, X., Foret, P., Sellam, T., and Parikh, A. P. A multilingual view of unsupervised machine translation. *arXiv preprint arXiv:2002.02955*, 2020a.

Garcia, X., Siddhant, A., Firat, O., and Parikh, A. P. Harnessing multilinguality in unsupervised machine translation for rare languages. *arXiv preprint arXiv:2009.11201*, 2020b.

Garcia, X., Constant, N., Guo, M., and Firat, O. Towards universality in multilingual text rewriting. *arXiv preprint arXiv:2107.14749*, 2021.

Goyal, N., Gao, C., Chaudhary, V., Chen, P.-J., Wenzek, G., Ju, D., Krishnan, S., Ranzato, M., Guzman, F., and Fan, A. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538, 2022.

Han, J. M., Babuschkin, I., Edwards, H., Neelakantan, A., Xu, T., Polu, S., Ray, A., Shyam, P., Ramesh, A., Radford, A., et al. Unsupervised neural machine translation with generative language models only. *arXiv preprint arXiv:2110.05448*, 2021.

Heek, J., Levskaya, A., Oliver, A., Ritter, M., Rondepierre, B., Steiner, A., and van Zee, M. Flax: A neural network library and ecosystem for JAX, 2020. URL http://github.com/google/flax.

Hernandez, D., Kaplan, J., Henighan, T., and McCandlish, S. Scaling laws for transfer. *arXiv preprint arXiv:2102.01293*, 2021.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., and Choudhury, M. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6282–6293, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.560. URL https://aclanthology.org/2020.acl-main.560.

Kim, Y., Graça, M., and Ney, H. When and why is unsupervised neural machine translation useless? *arXiv preprint arXiv:2004.10581*, 2020.

Kirstain, Y., Lewis, P., Riedel, S., and Levy, O. A few more examples may be worth billions of parameters. *arXiv preprint arXiv:2110.04374*, 2021.

Kocmi, T., Federmann, C., Grundkiewicz, R., Junczys-Dowmunt, M., Matsushita, H., and Menezes, A. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the*

Sixth Conference on Machine Translation, pp. 478–494, Online, November 2021. Association for Computational Linguistics. URL `https://aclanthology.org/2021.wmt-1.57`.

Kudo, T. and Richardson, J. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*, 2018.

Kumar, S., Anastasopoulos, A., Wintner, S., and Tsvetkov, Y. Machine translation into low-resource language varieties. *arXiv preprint arXiv:2106.06797*, 2021.

Lakew, S. M., Erofeeva, A., and Federico, M. Neural machine translation into language varieties. *arXiv preprint arXiv:1811.01064*, 2018.

Lample, G., Conneau, A., Denoyer, L., and Ranzato, M. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*, 2017.

Li, Z., Utiyama, M., Sumita, E., and Zhao, H. MiSS@WMT21: Contrastive learning-reinforced domain adaptation in neural machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pp. 154–161, Online, November 2021. Association for Computational Linguistics. URL `https://aclanthology.org/2021.wmt-1.12`.

Lin, X. V., Mihaylov, T., Artetxe, M., Wang, T., Chen, S., Simig, D., Ott, M., Goyal, N., Bhosale, S., Du, J., et al. Few-shot learning with multilingual language models. *arXiv preprint arXiv:2112.10668*, 2021.

Marchisio, K., Duh, K., and Koehn, P. When does unsupervised machine translation work? *arXiv preprint arXiv:2004.05516*, 2020.

Niu, X., Martindale, M., and Carpuat, M. A study of style in machine translation: Controlling the formality of machine translation output. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2814–2819, 2017.

Niu, X., Rao, S., and Carpuat, M. Multi-task neural models for translating between styles within and across languages. *arXiv preprint arXiv:1806.04357*, 2018.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Qian, L., Zhou, Y., Zheng, Z., Zhu, Y., Lin, Z., Feng, J., Cheng, S., Li, L., Wang, M., and Zhou, H. The volctrans glat system: Non-autoregressive translation meets wmt21. *arXiv preprint arXiv:2109.11247*, 2021.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P. J., et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020.

Raunak, V. and Menezes, A. Finding memo: Extractive memorization in constrained sequence generation tasks. *arXiv preprint arXiv:2210.12929*, 2022.

Ravi, S. and Knight, K. Deciphering foreign language. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 12–21, 2011.

Riley, P., Dozat, T., Botha, J. A., Garcia, X., Garrette, D., Riesa, J., Firat, O., and Constant, N. Frmt: A benchmark for few-shot region-aware machine translation. *arXiv preprint arXiv:2210.00193*, 2022.

Rippeth, E., Agrawal, S., and Carpuat, M. Controlling translation formality using pre-trained multilingual language models. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pp. 327–340, Dublin, Ireland (in-person and online), May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.iwslt-1.30. URL `https://aclanthology.org/2022.iwslt-1.30`.

Roberts, A., Chung, H. W., Levskaya, A., Mishra, G., Bradbury, J., Andor, D., Narang, S., Lester, B., Gaffney, C., Mohiuddin, A., et al. Scaling up models and data with t5x and seqio. *arXiv preprint arXiv:2203.17189*, 2022.

Sellam, T., Das, D., and Parikh, A. P. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*, 2020.

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

Song, K., Tan, X., Qin, T., Lu, J., and Liu, T.-Y. Mass: Masked sequence to sequence pre-training for language generation. *arXiv preprint arXiv:1905.02450*, 2019.

Subramanian, S., Hrinchuk, O., Adams, V., and Kuchaiev, O. NVIDIA NeMo's neural machine translation systems for English-German and English-Russian news and biomedical tasks at WMT21. In *Proceedings of the Sixth Conference on Machine Translation*, pp. 197–204, Online, November 2021. Association for Computational Linguistics. URL `https://aclanthology.org/2021.wmt-1.18`.

Tay, Y., Dehghani, M., Tran, V. Q., Garcia, X., Bahri, D., Schuster, T., Zheng, H. S., Houlsby, N., and Metzler, D. Unifying language learning paradigms. *arXiv preprint arXiv:2205.05131*, 2022.

Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Saravia, E., Poulton, A., Kerkez, V., and Stojnic, R. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022.

Tran, C., Bhosale, S., Cross, J., Koehn, P., Edunov, S., and Fan, A. Facebook AI's WMT21 news translation task submission. In *Proceedings of the Sixth Conference on Machine Translation*, pp. 205–215, Online, November 2021. Association for Computational Linguistics. URL `https://aclanthology.org/2021.wmt-1.19`.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Vilar, D., Freitag, M., Cherry, C., Luo, J., Ratnakar, V., and Foster, G. Prompting palm for translation: Assessing strategies and performance. *arXiv preprint arXiv:2211.09102*, 2022.

Wang, J., Xu, C., Guzmán, F., El-Kishky, A., Tang, Y., Rubinstein, B. I., and Cohn, T. Putting words into the system's mouth: A targeted attack on neural machine translation using monolingual data poisoning. *arXiv preprint arXiv:2107.05243*, 2021a.

Wang, L., Li, M., Liu, F., Shi, S., Tu, Z., Wang, X., Wu, S., Zeng, J., and Zhang, W. Tencent translation system for the WMT21 news translation task. In *Proceedings of the Sixth Conference on Machine Translation*, pp. 216–224, Online, November 2021b. Association for Computational Linguistics. URL `https://aclanthology.org/2021.wmt-1.20`.

Wang, W., Watanabe, T., Hughes, M., Nakagawa, T., and Chelba, C. Denoising neural machine translation training with trusted data and online data selection. *arXiv preprint arXiv:1809.00068*, 2018.

Xu, C., Wang, J., Tang, Y., Guzmán, F., Rubinstein, B. I., and Cohn, T. A targeted attack on black-box neural machine translation with parallel data poisoning. In *Proceedings of the Web Conference 2021*, pp. 3638–3650, 2021.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*, 2020.

Zeng, X., Liu, Y., Li, E., Ran, Q., Meng, F., Li, P., Xu, J., and Zhou, J. WeChat neural machine translation systems for WMT21. In *Proceedings of the Sixth Conference on Machine Translation*, pp. 243–254, Online, November 2021. Association for Computational Linguistics. URL `https://aclanthology.org/2021.wmt-1.23`.

Zhou, S., Zhou, T., Wei, B., Luo, Y., Mu, Y., Zhou, Z., Wang, C., Zhou, X., Lv, C., Jing, Y., Wang, L., Zhang, J., Huang, C., Yan, Z., Hu, C., Li, B., Xiao, T., and Zhu, J. The NiuTrans machine translation systems for WMT21. In *Proceedings of the Sixth Conference on Machine Translation*, pp. 265–272, Online, November 2021. Association for Computational Linguistics. URL `https://aclanthology.org/2021.wmt-1.26`.