

Translationese as a Language in “Multilingual” NMT

Parker Riley^{△*}, Isaac Caswell[▽], Markus Freitag[▽], David Grangier[▽]

[△]University of Rochester

[▽]Google Research

Abstract

Machine translation has an undesirable propensity to produce “translationese” artifacts, which can lead to higher BLEU scores while being liked less by human raters. Motivated by this, we model translationese and original (i.e. natural) text as separate languages in a multilingual model, and pose the question: can we perform zero-shot translation between original source text and original target text? There is no data with original source and original target, so we train sentence-level classifiers to distinguish translationese from original target text, and use this classifier to tag the training data for an NMT model. Using this technique we bias the model to produce more natural outputs at test time, yielding gains in human evaluation scores on both accuracy and fluency. Additionally, we demonstrate that it is possible to bias the model to produce translationese and game the BLEU score, increasing it while decreasing human-rated quality. We analyze these models using metrics to measure the degree of translationese in the output, and present an analysis of the capriciousness of heuristically-based train-data tagging.

1 Introduction

“Translationese” is a term that refers to artifacts present in text that was translated into a given language that distinguish it from text originally written in that language (Gellerstam, 1986). These artifacts include lexical and word order choices that are influenced by the source language (Gellerstam, 1996) as well as the use of more explicit and simpler constructions (Baker et al., 1993).

These differences between translated and original text mean that the direction in which parallel data (bitext) was translated is potentially important for machine translation (MT) systems. Most

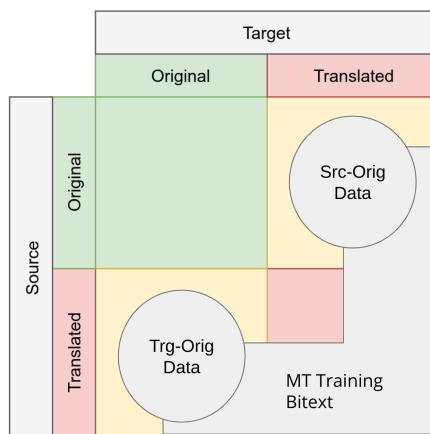


Figure 1: Representation of MT training and test data. The quadrants represent whether the source or target side of a parallel sentence pair is translated or original. The gray regions represent MT train+test data, illustrating which of the four sentence pair types are used.

parallel data is either source-original (the source was translated into the target; “forward” data) or target-original (the target was translated into the source; “reverse” data), though sometimes neither side is original because both were translated from a third language.

Figure 1 illustrates the four possible combinations of translated and original source and target data. Recent work has examined the impact of translationese in MT evaluation, using the WMT evaluation campaign as the most prominent example. From 2014 through 2018, WMT test sets were constructed such that 50% of the sentence pairs are source-original (upper right quadrant of Figure 1) and the rest are target-original (lower left quadrant). Toral et al. (2018), Zhang and Toral (2019), and Graham et al. (2019) have all examined the effect of this testing setup on MT evaluation, and have all argued that target-original test data should not be included in future evaluation campaigns because the translationese source is too easy to translate. While target-original test data does have the down-

*Work done while at Google Research.

side of a translationese source side, recent work has shown that human raters prefer MT output that is closer in distribution to original target text than translationese (Freitag et al., 2019). This indicates that the target side of test data should also be original (upper left quadrant of Figure 1); however, it is unclear how to produce high-quality test data (let alone training data) that is simultaneously source- and target-original.

Because of this lack of original-to-original sentence pairs, we frame this as a zero-shot translation task, where translationese and original text are distinct languages or domains. We adapt techniques from zero-shot translation with multilingual models (Johnson et al., 2016), where the training pairs are tagged with a reserved token corresponding to the domain of the target side: translationese or original text. Tagging is helpful when the training set mixes data of different types by allowing the model to 1) indicate each pair’s type at training to preserve distinct behaviors and avoid the model regressing to a mean/dominant prediction across data types, and 2) to elicit different behavior in inference, i.e. providing a tag at test time yields predictions resembling a specific data type. We then investigate what happens when the input is an original sentence in the source language and the model’s output is also biased to be original, a scenario never observed in training.

Tagging in this fashion is not trivial, as most MT training sets do not annotate which pairs are source-original and which are target-original, so in order to distinguish them we train binary classifiers to distinguish original and translated target text.

Finally, we perform several analyses of tagging the “languages” of translationese and demonstrate that tagged back-translation (Caswell et al., 2019), can be framed as a simplified version of our method, and thereby improved by targeted decoding.

Our contributions are as follows:

1. We propose two methods to train translationese classifiers using only monolingual text, coupled with synthetic text produced by machine translation.
2. Using only original→translationese and translationese→original training pairs, we apply techniques from zero-shot multilingual MT to enable original→original translation.
3. We demonstrate with human evaluations that this technique improves translation quality,

both in terms of fluency and accuracy.

4. We show that biasing the model to instead produce translationese outputs inflates BLEU scores while harming quality as measured by human evaluations.

2 Classifier Training + Tagging

Motivated by prior work detailing the importance of distinguishing translationese from original text (Kurokawa et al., 2009; Lembersky et al., 2012; Toral et al., 2018; Zhang and Toral, 2019; Graham et al., 2019; Freitag et al., 2019; Edunov et al., 2019) as well as work in zero-shot translation (Johnson et al., 2016), we hypothesize that performance on the source-original translation task can be improved by distinguishing target-original and target-translationese examples in the training data and constructing an NMT model to perform zero-shot original→original translation.

The problem is that most machine translation training sets do not annotate the original language for the sentence pairs¹; to address this, we train a binary classifier that predicts whether the target side of a sentence pair is original text in that language or translated from the source language. This follows several prior works attempting to identify translated text (Kurokawa et al., 2009; Koppel and Ordan, 2011; Lembersky et al., 2012).

For the same reasons that we need a translationese classifier (i.e. labeled data is scarce), we require a large corpus of target-language text annotated by whether it is original or translated. We use News Crawl data from WMT² as target-original data. It consists of news articles crawled from the internet, and so we assume that most of them are not translations. Getting translated data is trickier; most human-translated pairs where the original language is annotated are only present in test sets, which are generally small. To sidestep this, we adopt the following argument:

Premises

1. **[known]** Human translators produce translationese.
2. **[known]** Machine translation also produces translationese.

¹Europarl (Koehn, 2005) is a notable exception, but we found it to be too small for our purposes, and it is not in the news domain, which is what is used for the WMT translation task.

²<http://www.statmt.org/wmt18/translation-task.html>

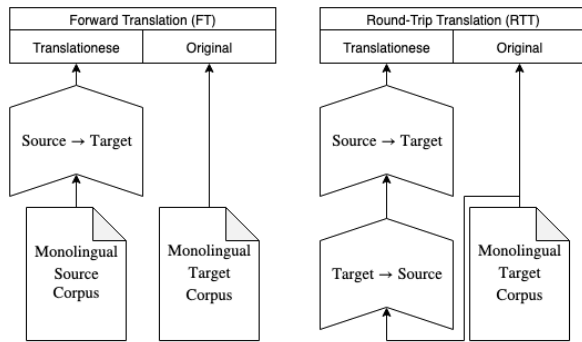


Figure 2: Illustration of data set creation for the FT and RTT translationese classifiers. The Source→Target and Target→Source nodes represent NMT systems.

3. **[our assumption]** Human translationese and machine translationese are similar.

Conclusion

1. Therefore, we can model human translationese with machine translationese.

As a result, we choose to use machine translation output as a proxy for human translationese, meaning we construct a classifier training data set using only unannotated monolingual data. We propose two ways of doing this: using forward translation (FT) or round-trip translation (RTT). Both are illustrated in Figure 2.

To generate FT data, we take source-language News Crawl data and translate it into the target language using a machine translation model trained on WMT training bitext. We can then train a classifier to distinguish the FT target-language data and monolingual target-language data.

One potential problem with the FT data set is that the original and translated pairs may differ not only in the respects we care about (i.e. translationese), but also in content. Taking English→French as an example language pair, one could imagine that certain topics are more commonly reported on in original English language news than in French, and vice versa, e.g. news about American or French politics, respectively. The words and phrases representing those topics could then act as signals to the classifier to distinguish the original language.

To address this, we also experiment with RTT data. For this approach we take target-language monolingual data and round-trip translate it with two machine translation models (target→source and then source→target), resulting in another target-language sentence that should contain the same content as the original sentence, alleviating

the concern with FT data. Here we hope that the noise introduced by round-trip translation will be similar enough to human translationese to be useful for our downstream task.

In both settings, we use the trained binary classifier to detect and tag training bitext pairs where the classifier predicted that the target side is original.

3 Experimental Set-up

3.1 Data

We perform our experiments on WMT18 English→German bitext and WMT15 English→French bitext. We use WMT News Crawl for monolingual data (2007-2017 for German and 2007-2014 for French). We filter out sentences longer than 250 subwords and remove pairs whose length ratio is greater than 2. This results in about 5M pairs for English→German. We do not filter the English→French bitext, resulting in 41M sentence pairs.

For monolingual data, we deduplicate and filter sentences with more than 70 tokens or 500 characters. For the experiments described later in Section 5.3, this monolingual data is back-translated with a target-to-source translation model; after doing so, we remove any sentence pairs where the back-translated source is longer than 75 tokens or 550 characters. This results in 216.5M sentences for English→German (of which we only use 24M) and 39M for English→French.

The classifiers were trained on the target language monolingual data in addition to either an equal amount of source language monolingual data machine-translated into the target language (for the FT classifiers) or the same target sentences round-trip translated through the source language with MT (for the RTT classifiers). In both cases, the MT models were trained only with WMT bitext.

The models used to generate the synthetic data have BLEU³ performance as follows on newstest2014/full: German→English 31.8; English→German 28.5; French→English 39.2; English→French 40.6.

3.2 Architecture

Our NMT models use the transformer-big architecture (Vaswani et al., 2017) implemented in *lingvo* (Shen et al., 2019) with a shared source-target byte-pair-encoding (BPE) vocabulary (Sen-

³BLEU + case.mixed + lang.LANGUAGE_PAIR + num-refs.1 + smooth.exp + test.SET + tok.13a + version.1.2.15

Language Pair	Test Set	Classifier Type	Content-Aware?	F1
English→French	Hansard	Kurokawa et al. (2009) A*	✓	0.77
		Kurokawa et al. (2009) B*	✗	0.69
English→French	newstest2014/full	Forward-translation (FT)	✓	0.81
		Round-trip translation (RTT)	✗	0.68
English→German	newstest2015	Forward-translation (FT)	✓	0.85
		Round-trip translation (RTT)	✗	0.65

Table 1: Performance of our own translationese classifiers and those reported by [Kurokawa et al. \(2009\)](#). All classifiers only saw the target side of each sentence pair. See Figure 2 for an overview of data used for the FT and RTT classifiers. The “Content-Aware?” column indicates whether the classifier was able to use content differences between the two classes to identify translations, or whether it was restricted to intrinsic textual features. *[Kurokawa et al. \(2009\) A](#) - word-based; [Kurokawa et al. \(2009\) B](#) - mixed word/POS-based.

[nrich et al., 2016b](#)) of 32k types. To stabilize training, we use exponentially weighted moving average (EMA) decay ([Buduma and Locascio, 2017](#)).

The English→French models were trained on 16 TPUs. The development set was newstest2013; for all models where some training data was tagged, the development set also included an additional set of copies of the target-original (i.e. original French) sentence pairs with the tag prepended to the source side. This was chosen as a way to select a model that balanced tagged and untagged decode performance. The models were trained for approximately 400,000 steps, and we choose the checkpoint with the highest development BLEU. The English→German models were trained on 32 GPUs with newstest2015 as the development set.

For the translationese classifier, we trained a three-layer CNN-based classifier optimized with Adagrad. The classifier checkpoint with the highest area under the precision-recall curve (AUCPR) on a development set was retained for evaluation; newstest2015 was the development set for the English→German classifiers and a subset of newstest2013 containing 500 English-original and 500 French-original sentence pairs for the English→French classifiers. The confidence threshold for the resulting binary classifiers was the one that achieved the highest F1 score on the development set. We found that the choice of architecture (RNN/CNN) and hyperparameters did not make a substantial difference in classifier accuracy.

3.3 Evaluation

We report BLEU ([Papineni et al., 2002](#)) scores with SacreBLEU ([Post, 2018](#))⁴ and run human evaluations for the best performing systems (Section 4.3).

⁴BLEU + case.mixed + lang.LANGUAGE_PAIR + numrefs.1 + smooth.exp + test.SET + tok.intl + version.1.2.15

	Classifier Type	Bitext %Trg-Orig	BT %Trg-Orig
French	FT	47%	47%
	RTT	30%	68%
German	FT	22%*	82%
	RTT	29%*	70%

Table 2: Percentage of training data where the target side was classified as original. The English→German bitext was much smaller than the English→French bitext, so we upsampled the pairs with predicted original German (marked with a * here) to make both subsets of the bitext the same size.

4 Results and Discussion

4.1 Classifier Accuracy

Before evaluating the usefulness of our translationese classifiers for the downstream task of machine translation, we can first evaluate how accurate they are at distinguishing original text from human translations. We use WMT test sets for this evaluation, because they consist of source-original and target-original sentence pairs in equal number.

The F1 scores for the various classifiers are shown in Table 1. Table 2 reports how much of the training data was classified as having original text on the target side. The back-translation (BT) classification proportions pertain to the experiments (Section 5.3). We note that while the classifiers trained to distinguish forward translations from original text perform reasonably well, the ones trained to identify round-trip translations are less effective. This result is in line with prior work by [Kurokawa et al. \(2009\)](#), who trained an SVM classifier on French sentences to detect translations from English. They used word n-gram features for their classifier, but were worried about a potential content effect and so also trained a classifier

Test set →	Src-Orig		Trg-Orig		Combined
Decode as if →	Natural	Transl.	Transl.	Natural	Domain match
∴ Domain match? →	✗	✓	✗	✓	✓
Train data tagging ↓					
Untagged	39.5	39.5	44.5	44.5	42.0
FT clf.	37.7	40.0	42.5	45.0	42.5
RTT clf.	38.0	39.4	43.2	44.1	41.8

Table 3: Average BLEU (newstest2014/full and newstest2015) for models trained on WMT 2014 English→French bitext, with tags added according to classifier predictions on the target side. We see that the tag allows us to control the domain (translationese vs. original text), and selectively achieve the highest domain-specific BLEU. By using in-domain decoding only on each half of the test set, we can achieve the highest combined BLEU.

Test set →	Src-Orig		Trg-Orig		Combined
Decode as if →	Natural	Transl.	Transl.	Natural	Domain match
∴ Domain match? →	✗	✓	✗	✓	✓
Train data tagging ↓					
Untagged	36.3	36.3	30.0	30.0	34.0
FT clf.	28.3	36.0	29.4	29.8	33.6
RTT clf.	32.3	36.2	30.0	30.2	33.9

Table 4: Average BLEU over newstest20{14/full,16,17,18} for models trained on WMT 2018 English→German bitext, with tags added according to classifier predictions on the target side. Our baseline has no tags, and is just the transformer-big architecture of Vaswani et al. (2017). Newstest2015 was used as development data.

where nouns and verbs were replaced with corresponding part-of-speech (POS) tags. Note that they tested on the Canadian Hansard corpus (containing Canadian parliamentary transcripts in English and French) while we tested on WMT test sets, so the numbers are not directly comparable, but it is interesting to see the similar trends in comparing content-aware and content-unaware versions of the same method. We also point out that Kurokawa et al. (2009) both trained and tested with human-translated sentences, while we trained our classifiers with machine-translated sentences while still testing on human-translated data.

4.2 NMT with Translationese-Classified Bitext

Table 3 shows the BLEU scores of three models all trained on WMT 2014 English→French bitext. The difference is in how the pairs with original target sides vs. translationese targets were distinguished from each other: either no distinction was made, or tags were applied to those sentence-pairs with a target side that a classifier predicted to be original French. We first note that the model trained on data tagged by the round-trip translation (RTT) classifier slightly underperforms the baseline. However, the model trained with data tagged by the for-

ward translation (FT) classifier is able to achieve an improvement of 0.5 BLEU on both halves of the test set when biased toward translationese on the source-original half and original text on the target-original half. This, coupled with the observation that the BLEU score on the source-original half sharply drops when adding the tag, indicates that the two halves of the test set represent quite different tasks, and that the model has learned to associate the tag with some aspects specific to generating original text as opposed to translationese.

However, we were not able to replicate this positive result on the English→German language pair; those results are reported in Table 4. Interestingly, in this scenario the relative ordering of the FT and RTT classifiers is reversed, with the German RTT classifier outperforming the FT classifier. This is also interesting because the German classifiers achieved higher F1 scores than the French ones, indicating that a classifier’s performance alone is not a sufficient indicator of its effect on translation performance. One possible explanation for the negative result is that the English→German bitext only contains 5M pairs, as opposed to the 41M for English→French, so splitting the data into two portions could make it difficult to learn both portions’ output distributions properly.

Test set → Tagging ↓	Src-Orig			Test set → Tagging ↓	Src-Orig		
	Decode	BLEU	Fluency		Decode	BLEU	Fluency
Untagged	-	43.9	26.7%	FT clf.	Translationese	44.6	24.2%
FT clf.	Natural	41.5	32.3%	FT clf.	Natural	41.5	30.8%

Table 5: Fluency side-by-side human evaluation for WMT English→French newstest2014/full (Table 3). We do the evaluation only on the source-original half of the test set because it corresponds to our goal of original→original translation. Although we see a drop of performance in BLEU, humans rate the output of the “produce-original-output” decode on average as more fluent when comparing to both the bitext model output and the same model with the “produce-translationese-output” decode. Percents represent the percent of raters who preferred that system.

4.3 Human Evaluation Experiments

In the previous subsection, we saw that BLEU for the source-original half of the test set went down when the model trained with FT classifications (*FT clf.*) was decoded it as if it were target-original (Table 3). Prior work has shown that BLEU has a low correlation with human judgments when the reference contains translationese but the system output is biased toward original/natural text (Fritag et al., 2019). This is the very situation we find ourselves in now. Consequently, we run a human evaluation to see if the output truly is more natural and thereby preferred by human raters, despite the loss in BLEU. We run both a fluency and an accuracy evaluation for English→French to compare the quality of this system when decoding as if source-original vs. target-original. We also compare the system with the *Untagged* baseline. All evaluations are conducted with bilingual speakers whose native language is French. Our two evaluations are as follows:

- **Accuracy:** We do a direct assessment for all outputs to measure translation accuracy. Each output was scored on a 6-point scale by 3 different raters, with the average taken as the final score. Raters were shown only the source sentence and the model output.
- **Fluency:** We do a side-by-side evaluation to measure fluency. As in the accuracy evaluation, each output is rated by 3 different humans. However, in this evaluation the source sentence is not shown, and raters are shown two different model outputs for the same (hidden) source. Raters were asked to select which output was more fluent, or whether they were equally good.

Fluency human evaluation results are shown in Table 5. In both evaluations, humans show a preference for the tagged model output when it is decoded

Test set → Tagging ↓	Src-Orig		
	Decode	BLEU	Accuracy
Untagged	-	43.9	4.51
FT clf.	Translationese	44.6	4.67
FT clf.	Natural	41.5	4.72

Table 6: Human evaluation of accuracy (direct assessment) for WMT English→French on the source-original half of newstest2014/full. Humans were asked to rate each output separately on a 6 point scale. The “produce-original-output” decode scores the best, just as with fluency (Table 5), despite the lower BLEU.

as natural text (with the tag). This shows that the natural decodes are more fluent than both the translationese (untagged) decodes from the same model and the baseline tagless model, despite the drop in BLEU compared to each. Accuracy human ratings are summarised in Table 6. Humans prefer decoding as natural text, demonstrating that the model does not suffer a loss in accuracy by generating more fluent output.

5 Supplemental Experiments

5.1 Measuring Translationese

Translationese tends to be simpler, more standardised and more explicit (Baker et al., 1993) compared to original text and can retain typical characteristics of the source language (Touy, 2012). Toral (2019) proposed metrics attempting to quantify the degree of translationese present in a translation. Following their work, we quantify lexical simplicity with two metrics: lexical variety and lexical density. We also calculate the length ratio between the source sentence and the generated translations to measure interference from the source.

5.1.1 Lexical Variety

An output is simpler when it uses a lower number of unique tokens/words. By generating output closer

Test set → Tagging ↓	Src-Orig			
	Decode	lexical variety	lexical density	length ratio
Untagged	-	0.258	0.393	0.246
FT clf.	Translationese	0.255	0.396	0.264
FT clf.	Natural	0.260	0.397	0.245

Table 7: Measuring the degree of translationese for WMT English→French newstest2014/full on the source-original half. Higher lexical variety, higher lexical density, and higher length ratio indicates less translationese output. The least translationese value for each metric is bolded.

to original target text, our hope is to increase lexical variety. Lexical variety is calculated as the type-token ratio (TTR):

$$TTR = \frac{\text{number of types}}{\text{number of tokens}} \quad (1)$$

5.1.2 Lexical Density

Scarpa (2006) found that translationese tends to be lexically simpler and have a lower percentage of content words (adverbs, adjectives, nouns and verbs) than original written text. Lexical density is calculated as follows:

$$\text{lex_density} = \frac{\text{number of content words}}{\text{number of total words}} \quad (2)$$

5.1.3 Length Ratio

Both MT and humans tend to avoid restructuring the source sentence and stick to sentence structures popular in the source language. This results in a translation with similar length to that of the source sentence. By measuring the length ratio, we measure interference in the translation because its length is guided by the source sentence’s structure. We compute the length ratio at the sentence level and average the scores over the test set of source-target pairs (x, y) :

$$\text{length ratio} = \frac{||x| - |y||}{|x|} \quad (3)$$

5.1.4 Results

Results for all three different translationese measurements are shown in Table 7.

- *Lexical Variety*

Using the tag to decode as natural text (i.e. more like original target text) increases lexical variety. This is expected as original sentences tend to use a larger vocabulary than simpler outputs.

- *Lexical Density*

We also increase lexical density when decoding as natural text. In other words, the model has a higher percentage of content words in its output, which is an indication that it is more like original target-language text.

- *Length Ratio*

Unlike the previous two metrics, decoding as natural text does not lead to a more “natural” distribution of length-ratios. However, there is a clear difference between this metric and the previous two: length ratio also takes the source sentence into account. This hints at an interesting subtlety of our tagging scheme: since all of our training pairs have translationese on either the source or the target side, the model never fully learns to model the aspects of translationese that relate the source and target sentences, and can only model the naturalness of the target language in isolation. This comes back to the problem of the lack or original→original training data noted in the introduction.

5.2 Tagging using Translationese Heuristics

Rather than tagging training data with a trained classifier, as explored in the previous sections, it might be possible to tag using much simpler heuristics, and achieve a similar effect. We explore two options here:

5.2.1 Simple-length-ratio tagging

We tag all sentence-pairs (x, y) in the training data where the ratio of token lengths $\frac{|x|}{|y|}$ is greater than the empirical ratio of token lengths $\hat{\rho}_{length}$ in two monolingual corpora, M_x and M_y :

$$\hat{\rho}_{length} = \frac{\frac{1}{|M_x|} \sum_{x_i \in M_x} |x_i|}{\frac{1}{|M_y|} \sum_{y_i \in M_y} |y_i|} \quad (4)$$

In this case, the tag indicates that the output is shorter than expected. For English→French, we

Test set →	Src-Orig	Src-Orig	Trg-Orig	Trg-Orig
Decode as if →	Natural	Transl.	Transl.	Natural
∴ Domain match? →	✗	✓	✗	✓
Train data tagging type ↓				
All Untagged	39.5	39.5	44.5	44.5
Tagged by FT clf.	37.7	40.0	42.5	45.0
Tagged by Length Ratio	38.2	36.1	43.6	36.2
Tagged by Lex. Density	36.9	36.7	41.2	43.4

Table 8: Comparing heuristic-based tagging with classifier-based tagging. BLEU scores are averaged for newstest2014/full and newstest2015 English→French. Neither heuristic performs as well as the trained classifier, and length-ratio tagging actually has the reverse effect from what we expect.

found $\hat{\rho}_{length} = 0.8643$, meaning that original French sentences tend to have more tokens than English. This tagging scheme is more to understand the effect that a somewhat meaningful tagging technique might have, rather than to capture a complicated phenomenon like translationese. That said, since translationese texts tend to have similar length to the source (see Section 5.1.3), i.e. $\frac{|x|}{|y|} \approx 1.0$, tagging at $\frac{|x|}{|y|} = \hat{\rho}_{length} = 0.8643$ means that the majority of the translationese training sentences are likely to get tagged. We find that 49.8% of the training bitext falls above this value and is tagged.

5.2.2 Lexical Density Tagging

We tag examples with a target-side lexical density of greater than 0.5, which means that the target is more likely to be original than translationese. Please refer to Section 5.1.2 for an explanation of this metric.

5.2.3 Results

Table 8 shows the results for this experiment, compared to the untagged baseline and the classifier-tagged model from Table 3. This table specifically looks at the effect of decoding on the subsets of the test set with- and without the tag, e.g. controlling whether the output should feature more or less translationese. We see that like the classifier-based tagging, the lexical density tagging approach yields expected results, in that the tag can be used to effectively increase BLEU on the target-original portion of the test set. The length-ratio tag, however, has the opposite effect: producing shorter outputs (“decode as if translationese”) produces higher BLEU against the target-original test set and lower BLEU against the source-original subset. We speculate that this data partition has accidentally picked up on some artifact of the data.

Two interesting observations from Table 8 are that 1) both heuristic tagging methods perform much more poorly than the classifier tagging

method on both halves of the test set, and 2) all varieties of tagging produce large changes in performance (up to -7.2 BLEU). This second observation highlights how powerful tagging can be – and how dangerous it can be to partition training data when the partitions do not correspond well to the desired feature (translationese in our case).

5.3 Back-Translation Experiments

We also investigated whether using a classifier to tag training data improved model performance in the presence of back-translated (BT) data. Caswell et al. (2019) introduced tagged back-translation (TBT), where all back-translated pairs are tagged and no bitext pairs are. They experimented with decoding the model with a tag (“as-if-back-translated”) but found it harmed BLEU score. However, in our early experiments we discovered that doing this actually *improved* the model’s performance on the target-original portion of the test set, while harming it on the source-original half. Thus, we frame TBT as a heuristic method for identifying target-original pairs: the monolingual data used for the back-translations is assumed to be original, and the target side of the bitext is assumed to be translated. We wish to know whether we can find a better tagging scheme for the combined BT and bitext data, based on a classifier or some other heuristic. We are confident that at least *some* of the training bitext is target-original, and it is even possible that some of the monolingual target data is *not* original, as it could have been translated from another language. Because of this, we perform experiments with MT models trained with bitext and BT data tagged by a classifier or heuristic.

Results for English→French models trained with BT data are presented in Table 9. While combining the bitext classified by the FT classifier with all-tagged BT data yields a minor gain of 0.2 BLEU over the TBT baseline of Caswell et al. (2019), the other methods do not beat the baseline. This indi-

Test set →		Src-Orig		Trg-Orig		Combined
Decode as if →		Natural	Transl.	Transl.	Natural	Both
∴ Domain match? →		✗	✓	✗	✓	✓
Bitext tagging ↓	BT tagging ↓					
Untagged	All Tagged	38.4	40.8	47.5	49.8	45.5
FT clf.	All Tagged	38.8	40.8	47.3	50.3	45.7
FT clf.	FT clf.	38.2	40.9	45.5	49.0	45.2
RTT clf.	RTT clf.	38.3	40.1	49.4	49.5	45.1

Table 9: Average BLEU (newstest2014/full and newstest2015) scores for models trained on WMT 2014 English→French bitext and 39M back-translated monolingual sentences, with tags added according to heuristics and/or classifier predictions on the target (French) side. The first result row represents tagged back-translation (Caswell et al., 2019) and is our baseline. All data was filtered by an automatic language identification tool, retaining only pairs with English source and French target. Newstest2013 was used as development data.

Test set →		Src-Orig		Trg-Orig		Combined
Decode as if →		Natural	Transl.	Transl.	Natural	Both
∴ Domain match? →		✗	✓	✗	✓	✓
Bitext tagging ↓	BT tagging ↓					
Untagged	All Tagged	33.5	37.3	36.7	37.1	37.6
FT clf.	All Tagged	33.4	37.2	36.2	37.2	37.5
RTT clf.	All Tagged	33.6	37.4	36.6	37.1	37.6
RTT clf.	RTT clf.	31.6	35.7	36.8	36.7	36.4
FT clf.	FT clf.	30.5	35.5	36.5	37.0	36.5

Table 10: Average BLEU scores (newstest20{14/full,16,17,18}) for models trained on WMT 2018 English→German bitext and 24 million back-translated monolingual sentences, with tags added according to heuristics and/or classifier predictions on the target (German) side. The first result row represents tagged back-translation (Caswell et al., 2019) and is our baseline. All data was filtered by an automatic language identification tool, retaining only pairs with English source and German target. Newstest2015 was used as development data. For the last two rows, the back-translations consisted of 12 million untagged and 12 million tagged pairs.

cates that assuming all of the target monolingual data to be original is not as harmful as the error introduced by the classifiers.

English→German results are presented in Table 10. Combining the bitext classified by the RTT classifier with all-tagged BT data matched the performance of the TBT baseline, but none of the models outperformed it. This is expected, given the poor performance of the bitext-only models for this language pair.

6 Example Output

In Table 11, we show example outputs for WMT English→French comparing the *Untagged* baseline with the *FT clf.* inference output when decoding with tags (force the model to generate natural output). In the first example, *avec suffisamment d’art* is an incorrect word-for-word translation, as the French word *art* cannot be used in that context. Here the word *habilement*, which is close to “skillfully” in English, sounds more natural. In the second example, *libre d’impôt* is the literal translation of “tax-free”, but French documents rarely use it, they prefer *pas imposable*, meaning “not taxable”.

In the last example, *arriéré* in French can be used only for backlog in the sense of a late payment and cannot be used in this context; *retard*, which means lateness/delay, is appropriate here.

7 Related Work

7.1 Translationese

The effects of translationese on MT training and evaluation have been investigated by many prior authors (Kurokawa et al., 2009; Lembersky et al., 2012; Toral et al., 2018; Zhang and Toral, 2019; Graham et al., 2019; Freitag et al., 2019; Edunov et al., 2019). Training classifiers to detect translationese has also been done (Kurokawa et al., 2009; Koppel and Ordan, 2011). Similarly to this work, Kurokawa et al. (2009) used their classifier to preprocess MT training data; however, they merely filtered out target-original pairs. In contrast, Lembersky et al. (2012) used both types of data (without explicitly distinguishing them with a classifier), and used entropy-based measures to cause their phrase-based system to favor phrase table entries with target phrases that are more similar to a corpus

source	Sorry she didn't phrase it artfully enough for you.
Untagged	Désolée, elle ne l'a pas formulé avec suffisamment d'art pour vous.
FT clf.	Désolé elle ne l'a pas formulé assez habilement pour vous.
source	Your first 10,000 is tax free .
Untagged	Votre première tranche de 10 000 est libre d'impôt .
FT clf.	La première tranche de 10 000 n'est pas imposable .
source	Exactly, when I graduated in 2011 there was already a three year backlog.
Untagged	Exactement, lorsque j'ai obtenu mon diplôme en 2011, il y avait déjà un arriéré de trois ans.
FT clf.	Exactement, lorsque j'ai obtenu mon diplôme en 2011, il y avait déjà un retard de trois ans.

Table 11: Example output for WMT English→French comparing the untagged baseline output with the *FT clf* decoding output with tags.

of translationese than original text. In this work, we combine aspects from each of these: we train a classifier to partition the training data, and use both subsets to train a single model with a mechanism allowing control over the degree of translationese to produce in the output. We also use more modern neural MT methods instead of a phrase-based system, and show with human evaluations that source-original test sentence pairs result in BLEU scores that do not correlate well with translation quality when evaluating models trained to produce more original output.

7.2 Training Data Tagging for NMT

In addition to tagging methods as in Caswell et al. (2019), inserting tags in training data and using them to control output is a technique that has been growing in popularity. Tags on the source sentence have been used to indicate target language in multilingual models (Johnson et al., 2016), formality level in English→Japanese (Yamagishi et al., 2016), politeness in English→German (Sennrich et al., 2016a), gender from a gender-neutral language (Kuczmarski and Johnson, 2018), as well as to produce domain-targeted translation (Kobus et al., 2016). Shu et al. (2019) use tags at training and inference time to increase the syntactic diversity of their output while maintaining translation quality; similarly, Agarwal and Carpuat (2019) and Marchisio et al. (2019) use tags to control the reading level (e.g. simplicity/complexity) of the output. Overall, tagging can be seen as domain adaptation (Freitag and Al-Onaizan, 2016; Luong and Manning, 2015).

8 Conclusion

We have demonstrated that translationese and original text can be treated as separate target languages

in a “multilingual” model, distinguished by a classifier trained using only monolingual and synthetic data. The resulting model has improved performance in the ideal, zero-shot scenario of original→original translation, as measured by human evaluation of accuracy and fluency. However, this is associated with a drop in BLEU score, indicating that better automatic evaluation is needed.

References

- Swetha Agarwal and Marine Carpuat. 2019. [Controlling Text Complexity in Neural Machine Translation](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.
- Mona Baker, Gill Francis, and Elena Tognini-Bonelli. 1993. *Corpus Linguistics and Translation Studies: Implications and Applications*, chapter 2. John Benjamins Publishing Company, Netherlands.
- Nikhil Buduma and Nicholas Locascio. 2017. *Fundamentals of deep learning: Designing next-generation machine intelligence algorithms*. “O’Reilly Media, Inc.”.
- Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. [Tagged back-translation](#). In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 53–63, Florence, Italy. Association for Computational Linguistics.
- Sergey Edunov, Myle Ott, Marc’Aurelio Ranzato, and Michael Auli. 2019. On the evaluation of machine translation systems trained with back-translation. *arXiv preprint arXiv:1908.05204*.
- Markus Freitag and Yaser Al-Onaizan. 2016. [Fast domain adaptation for neural machine translation](#).
- Markus Freitag, Isaac Caswell, and Scott Roy. 2019. [APE at Scale and Its Implications on MT Evaluation Biases](#). In *Proceedings of the Fourth Conference on Machine Translation*, pages 34–44, Florence, Italy. Association for Computational Linguistics.

- Martin Gellerstam. 1986. Translationese in Swedish novels translated from English. In Lars Wollin and Hans Lindquist, editors, *Translation Studies in Scandinavia*, page 8895. CWK Gleerup.
- Martin Gellerstam. 1996. Translations as a source for cross-linguistic studies. *Lund studies in English*, 88:53–62.
- Yvette Graham, Barry Haddow, and Philipp Koehn. 2019. [Translationese in machine translation evaluation](#). *CoRR*, abs/1906.09833.
- Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. [Google’s Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation](#). *CoRR*, abs/1611.04558.
- Catherine Kobus, Josep Maria Crego, and Jean Senelart. 2016. [Domain Control for Neural Machine Translation](#). *CoRR*, abs/1612.06140.
- Philipp Koehn. 2005. [Europarl: A Parallel Corpus for Statistical Machine Translation](#). In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.
- Moshe Koppel and Noam Ordan. 2011. [Translationese and its dialects](#). In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT ’11, pages 1318–1326, Stroudsburg, PA, USA. Association for Computational Linguistics.
- James Kuczumski and Melvin Johnson. 2018. [Gender-aware natural language translation](#). *Technical Disclosure Commons*.
- David Kurokawa, Cyril Goutte, and Pierre Isabelle. 2009. Automatic detection of translated text and its impact on machine translation. In *Proceedings of MT-Summit XII*, pages 81–88.
- Gennadi Lembersky, Noam Ordan, and Shuly Wintner. 2012. [Adapting translation models to translationese improves SMT](#). In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL ’12, pages 255–265, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Minh-Thang Luong and Christopher D Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 76–79.
- Kelly Marchisio, Jialiang Guo, Cheng-I Lai, and Philipp Koehn. 2019. [Controlling the reading level of machine translation output](#). In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 193–203, 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.
- Matt Post. 2018. [A Call for Clarity in Reporting Bleu Scores](#). *arXiv preprint arXiv:1804.08771*.
- Federica Scarpa. 2006. Corpus-based quality-assessment of specialist translation: A study using parallel and comparable corpora in English and Italian. *Insights into specialized translation—linguistics insights*. Bern: Peter Lang, pages 155–172.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. [Controlling politeness in neural machine translation via side constraints](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40, San Diego, California. Association for Computational Linguistics.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. [Neural machine translation of rare words with subword units](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.
- Jonathan Shen, Patrick Nguyen, Yonghui Wu, Zhifeng Chen, Mia X. Chen, Ye Jia, Anjali Kannan, Tara N. Sainath, and Yuan Cao et al. 2019. [Lingvo: a Modular and Scalable Framework for Sequence-to-Sequence Modeling](#). *CoRR*, abs/1902.08295.
- Raphael Shu, Hideki Nakayama, and Kyunghyun Cho. 2019. [Generating Diverse Translations with Sentence Codes](#). In *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics.
- Antonio Toral. 2019. [Post-editeese: an exacerbated translationese](#). *CoRR*, abs/1907.00900.
- Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. [Attaining the unattainable? Reassessing claims of human parity in neural machine translation](#). In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, Belgium, Brussels. Association for Computational Linguistics.
- Gideon Toury. 2012. *Descriptive translation studies and beyond: Revised edition*, volume 100. John Benjamins Publishing.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention Is All You Need](#). In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Hayahide Yamagishi, Shin Kanouchi, Takayuki Sato, and Mamoru Komachi. 2016. Controlling the voice of a sentence in Japanese-to-English neural machine translation. In *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, pages 203–210.

Mike Zhang and Antonio Toral. 2019. [The effect of translationese in machine translation test sets](#). *CoRR*, abs/1906.08069.