On the Language Coverage Bias for Neural Machine Translation

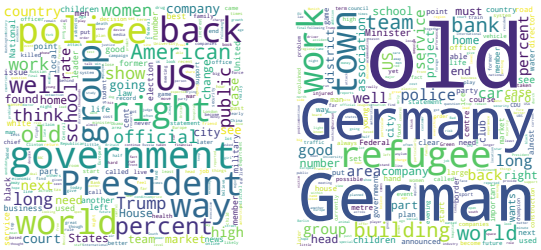
Shuo Wang[⋆] **Zhaopeng Tu**[⊤] **Zhixing Tan**[⋆] **Shuming Shi**[⊤] **Maosong Sun**[⋆†] **Yang Liu**[⋆†‡]
[⋆]Dept. of Comp. Sci. & Tech., Institute for AI, BNRist Center, Tsinghua University
[⊤]Tencent AI Lab
[†]Beijing Academy of Artificial Intelligence
[‡]Institute for AIR, Tsinghua University
[⋆]{wangshuo2018, zxtan, sms, liuyang2011}@tsinghua.edu.cn
[⊤]{zptu, shumingshi}@tencent.com

Abstract

Language coverage bias, which indicates the content-dependent differences between sentence pairs originating from the source and target languages, is important for neural machine translation (NMT) because the target-original training data is not well exploited in current practice. By carefully designing experiments, we provide comprehensive analyses of the language coverage bias in the training data, and find that using only the source-original data achieves comparable performance with using full training data. Based on these observations, we further propose two simple and effective approaches to alleviate the language coverage bias problem through explicitly distinguishing between the source- and target-original training data, which consistently improve the performance over strong baselines on six WMT20 translation tasks. Complementary to the translationese effect, language coverage bias provides another explanation for the performance drop caused by back-translation (Marie et al., 2020). We also apply our approach to both back- and forward-translation and find that mitigating the language coverage bias can improve the performance of both the two representative data augmentation methods and their tagged variants (Caswell et al., 2019).

1 Introduction

In recent years, there has been a growing interest in investigating the effect of original languages in parallel data on neural machine translation (Barrault et al., 2020; Edunov et al., 2020; Marie et al., 2020). Several studies have shown that target-original test examples[1] can lead to distortions in automatic and human evaluations, which should be omitted from machine translation test sets (Barrault et al., 2019; Zhang and Toral, 2019; Graham et al., 2020). Another branch of studies report that target-original test data leads to discrepant conclusions: back-translation only benefits the translation of target-original test data while harms that of source-original test data (Edunov et al., 2020; Marie et al., 2020). They attribute these phenomena to the reason that human-translated texts (i.e., *translationese*) exhibit formal and stylistic differences that set them apart from the texts originally written in that language (Baker et al., 1993; Volansky et al., 2015; Zhang and Toral, 2019).



(a) English-Original (b) German-Original

Figure 1: Example of language coverage bias illustrated by word clouds that are plotted at the English side of sentence pairs in the En-De test sets from WMT10 to WMT18. The test sets consist of English-original and German-original sentence pairs.

Complementary to the translationese bias, which is *content-independent* (Volansky et al., 2015), we identify another important problem, namely *language coverage bias*, which refers to the *content-dependent* differences in data originating from different languages. These differences stem from the diversity of regions and cultures. While the degree of the translationese bias varies across different translators (Toral, 2019), language coverage bias is an intrinsic bias between the source- and target-original data, which is hardly affected by the ability of the translator. Figure 1 shows an example, where the contents in English- and German-original texts differ significantly due to language coverage bias.

[1]Target-original test examples are sentence pairs that are translated from the target language into the source language.

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 4778–4790
August 1–6, 2021.

To investigate the effect of language coverage bias in the training data on NMT models, we propose an automatic method to identify the original language of each training example, which is generally unknown in practical corpora. Experimental results on three large-scale translation corpora show that there exists a significant performance gap between NMT models trained on the source- and target-original data, which have different vocabulary distributions, especially for content words. Since the target-original training data performs poorly in translating content words, using only the source-original data achieves comparable performance with using full training data. These findings motivate us to explore other data utilization methods rather than indiscriminately mixing the source- and target-original training data.

We propose to alleviate the language coverage bias problem by explicitly distinguishing between the source- and target-original training data. Specifically, two simple and effective methods are employed: *bias-tagging* and *fine-tuning*. Experimental results show that both approaches consistently improve the performance on six WMT20 translation tasks. Language coverage bias also provides another explanation for the failure of back-translation on the source-original test data, complementary to the translationese effect (Marie et al., 2020). We further validate our approach in the monolingual data augmentation scenario, where the language coverage bias problem would be more severe due to the newly introduced monolingual data.

Contributions The main contributions of our work are listed as follows:

- We demonstrate the necessity of studying the language coverage bias for NMT, and identify that using the target-original data can cause poor translation adequacy on content words.
- We address the language coverage bias induced by the target-original data by explicitly distinguishing the original languages, which can significantly improve the translation performance on six WMT20 translation tasks.
- We show that alleviating the language coverage bias also benefits monolingual data augmentation, which can improve both back- and forward-translation and their tagged variants (Caswell et al., 2019).

2 Experimental Setup

Data We conducted experiments on six WMT20 benchmarks (Barrault et al., 2020), including English⇔German (En⇔De), English⇔Chinese (En⇔Zh), and English⇔Japanese (En⇔Ja) news translation tasks. The preprocessed training corpora contain 41.0M, 21.8M, and 13.0M sentence pairs for En⇔De, En⇔Zh, and En⇔Ja, respectively. We used the monolingual data that is publicly available in WMT20 to train the proposed original language detection model (Section 3.1) and data augmentation (Section 4.2). The Appendix lists details about the data preprocessing.

For En⇔De and En⇔Zh, we used newstest2019 as the validation sets. For En⇔Ja, we split the official validation set released by WMT20 into two parts by the original language and only used the corresponding part for each direction. We used newstest2020 as the test sets for all the six tasks. We reported the Sacre BLEU (Post, 2018), as recommended by WMT20.

Model We used the Transformer-Big (Vaswani et al., 2017) model, which consists of a 6-layer encoder and a 6-layer decoder, and the hidden size is 1024. Recent studies showed that training on large batches can further boost model performance (Ott et al., 2018; Wu et al., 2018). Accordingly, we followed their settings to train models with batches of approximately 460k tokens. Please refer to the Appendix for more details about model training. We followed Ng et al. (2019) to use the Transformer-Big decoder as our language models, which are used to detect the original language and measure translation fluency. Language models are also trained with large batches (Ott et al., 2018).

3 Observing Language Coverage Bias

In this study, we first establish the existence of language coverage bias (Section 3.2), and show how the bias affects NMT performance (Section 3.3). To this end, we propose an automatic method to detect the original language of each training example (Section 3.1), which is often not available in large-scale parallel corpora (Riley et al., 2020).

3.1 Detecting Original Languages

Detection Method Intuitively, we use a large-scale monolingual dataset to estimate the distribution of the contents covered by each language. For each training example, we compare its similarities

Method	En-Zh	En-Ja	En-De
FT	83.6	83.7	86.6
Ours	**84.4**	**91.5**	**88.7**

Table 1: F1 scores of detecting original languages in the test sets. "FT" denotes the forward translation classifier proposed by Riley et al. (2020).

to the distributions of source and target languages, based on which we determine its original language.

Formally, let $\mathcal{D}_s$ and $\mathcal{D}_t$ denote the source-side and target-side distributions of the covered contents. Given a training example $\langle \mathbf{x}, \mathbf{y} \rangle$, the probability that it is covered by one language (represented as $\mathcal{D}_s$ and $\mathcal{D}_t$) can be expressed as

$$P(\mathcal{D}_s|\langle \mathbf{x}, \mathbf{y} \rangle) = \frac{P(\mathcal{D}_s)P(\langle \mathbf{x}, \mathbf{y} \rangle|\mathcal{D}_s)}{P(\langle \mathbf{x}, \mathbf{y} \rangle)},$$
$$P(\mathcal{D}_t|\langle \mathbf{x}, \mathbf{y} \rangle) = \frac{P(\mathcal{D}_t)P(\langle \mathbf{x}, \mathbf{y} \rangle|\mathcal{D}_t)}{P(\langle \mathbf{x}, \mathbf{y} \rangle)}.$$

We use a score function to denote the difference between the two probabilities:

$$\begin{aligned} score &= \log P(\mathcal{D}_s|\langle \mathbf{x}, \mathbf{y} \rangle) - \log P(\mathcal{D}_t|\langle \mathbf{x}, \mathbf{y} \rangle), \\ &= \log P(\langle \mathbf{x}, \mathbf{y} \rangle|\mathcal{D}_s) - \log P(\langle \mathbf{x}, \mathbf{y} \rangle|\mathcal{D}_t) + c, \end{aligned}$$

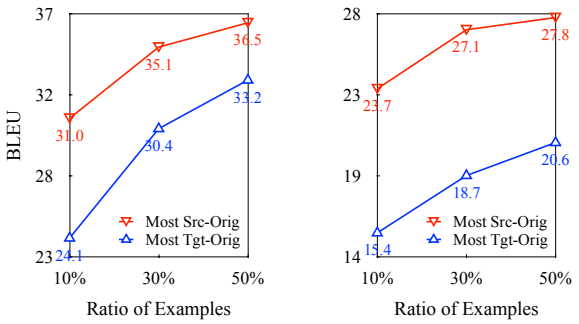
where $c = \log P(\mathcal{D}_s) - \log P(\mathcal{D}_t)$, which has a constant value when the source and target monolingual datasets are given. Intuitively, examples with higher score values are more likely to be source-original while those with lower score values are more likely to be target-original data. We train language models $\boldsymbol{\theta}_s^{lm}$ and $\boldsymbol{\theta}_t^{lm}$ on the source- and target-language monolingual data to estimate the conditional probabilities:

$$P(\langle \mathbf{x}, \mathbf{y} \rangle|\mathcal{D}_s) = P(\mathbf{x}|\boldsymbol{\theta}_s^{lm}),$$
$$P(\langle \mathbf{x}, \mathbf{y} \rangle|\mathcal{D}_t) = P(\mathbf{y}|\boldsymbol{\theta}_t^{lm}).$$

Accordingly, the score can be rewritten as

$$score = \log P(\mathbf{x}|\boldsymbol{\theta}_s^{lm}) - \log P(\mathbf{y}|\boldsymbol{\theta}_t^{lm}) + c. \quad (1)$$

We label examples as source-original if their score values are positive, and the other examples as target-original. To find a specific constant for each language pair, we tune the value of c to obtain the best classification performance on the validation sets, where the original languages are known.



(a) En⇒Zh (b) Zh⇒En

Figure 2: Translation performance on the validation sets of the En⇔Zh translation task for different ratios of most source- and target-original training examples.

Detection Accuracy We evaluated the detection method on the mixture of the test sets of bidirectional translation tasks in WMT20 for each language pair. For comparison, we re-implemented the CNN-based forward-translation (FT) classifier proposed by Riley et al. (2020). The FT classifier and the language models used in our method were trained on the same monolingual data sets. Table 1 shows that our method outperforms the FT classifier in all language pairs. In addition, our model also outperforms the FT approach on detecting noisy training data, which leads to an improvement in translation performance (please refer to Table 11 in the Appendix for more results).

3.2 Existence of Language Coverage Bias

In this section, we validate the existence of language coverage bias by (1) comparing the performance of NMT models trained on data with different original languages, and (2) directly calculating the divergence between the vocabulary distributions of the source- and target-original data.

Translation Performance Once all the training examples are assigned a score by the detection method (Eq. (1)), we regard $R\%$ of examples with the highest scores as the most source-original examples, and $R\%$ of examples with the least scores as the most target-original examples. We investigate the effect of $R\%$ on translation performance, as shown in Figure 2. Clearly, using the most source-original examples significantly outperforms using its target-original counterparts, demonstrating that the source- and target-original data indeed differ greatly from each other. To rule out the effect of data scale, we treat 50% of data with the highest scores as source-original data, and the same amount

Data	JS ($\times 10^{-5}$)		
	All	*Content*	*Function*
Random	4	10	0
S vs T	745	1503	261

Table 2: JS divergence of the vocabulary distributions between the source- and target-original data ("S vs T") on the training set of WMT20 En⇔Zh. "*All*", "*Content*", and "*Function*" denote all words, content words, and function words respectively. For reference, we also report the JS divergence between randomly selected 50% examples and the others ("Random").

Data	En-Zh		En-Ja		En-De	
Origin	⇒	⇐	⇒	⇐	⇒	⇐
Target	33.2	20.6	30.5	15.4	39.3	37.4
Source	36.5	**27.8**	**35.3**	17.9	41.7	**42.5**
Both	**36.6**	27.5	34.9	**18.5**	**42.3**	42.2

Table 3: Sacre BLEU of using different sets of training data on validation sets. We highlight the **highest** score in bold and the second-highest score with underlines.

Data	En⇒Zh			En⇐Zh		
Origin	*noun*	*verb*	*adj*	*noun*	*verb*	*adj*
Target	67.6	52.0	64.3	53.8	38.0	57.0
Source	69.7	54.0	**66.2**	**61.8**	**44.1**	**63.9**
Both	**69.9**	**54.1**	65.9	61.2	43.8	63.4

Table 4: Translation adequacy of different types of content words measured by F-measure (Neubig et al., 2019). The results are reported on the validation sets.

of data with the least scores as target-original data in the following experiments by default.

Since some recent works find that BLEU might be affected by the translationese problem (Edunov et al., 2020; Freitag et al., 2020), we have also conducted a side-by-side human-evaluation on the Zh⇒En development set, where 500 randomly sampled examples were evaluated by six persons (agreement (Fleiss, 1971): Fleiss' Kappa=0.46). 37.0% of outputs using the source-original data are better than using target-original data, and 21.0% are worse. By manually checking the outputs, we find using only the target-original data tends to omit important source contents (e.g., named entities) either by totally ignoring some contents or by using pronouns instead. The human-evaluation shows the same trend with the BLEU score presented in Figure 2. Given that conducting human-evaluation on all the six translation tasks is time-consuming and labor-intensive, we use automatic measures to further investigate this problem in Section 3.3.

Vocabulary Distributions Complementary to previous studies that focus on the *content-independent* stylistic difference (Volansky et al., 2015) between translationese and original texts (Riley et al., 2020; Edunov et al., 2020; Marie et al., 2020), we investigate the *content-dependent* language coverage bias between the source- and target-original data in this experiment. Intuitively, if the language coverage bias exists, the vocabulary distributions of the source- and target-original data should differ greatly from each other, since the covered issues tend to have different frequencies between them (D'Alessio and Allen, 2000). We use the Jensen-Shannon (JS) divergence (Lin, 1991) to measure the difference between two vocabulary distributions p and q:

$$\mathrm{JS}\left(p||q\right) = \frac{1}{2}\left(\mathrm{KL}(p||\frac{p+q}{2}) + \mathrm{KL}(q||\frac{p+q}{2})\right),$$

where $\mathrm{KL}(\cdot||\cdot)$ is the KL divergence (Kullback and Leibler, 1951) of two distributions.

Table 2 shows the JS divergence of the vocabulary distributions between the source- and target-original data. We also divide the words into content words and functions words based on their POS tags, since content words are more related to the language coverage bias, while the function words are more related to the stylistic and structural differences between the translationese and original texts (Lembersky et al., 2011; Volansky et al., 2015). The JS divergence between the source- and target-original data are 186× larger than that between randomly split data, which is mainly due to the difference between content words. Results for different ratios $R\%$ and other language pairs can be found in Appendix (Tables 12 and 13), where the trend holds in all cases, supporting our claim of the existence of language coverage bias.

3.3 Effect of Language Coverage Bias

In this section, we investigate the effect of language coverage bias on NMT models.

Using only the source-original data achieves comparable performance with using full data. Table 3 lists the translation performances of NMT

models trained on only the source- or target-original data and on both of them. The results show that using only the source-original data significantly outperforms using the target-original data in all language pairs, which reconfirm the necessity of studying the language coverage bias for NMT. It should be emphasized that using only the source-original data (i.e. 50% of the whole training set) achieves translation performances on par with using full training data. In the following experiments, we investigate why using target-original data together cannot further improve the performance.

Using additional target-original data does not consistently improve translation adequacy. To rule out the effect of translationese and focus on the content-dependent difference caused by the language coverage bias, we examine the translation adequacy of content words in Table 4 [2]. We follow Raunak et al. (2020) to use F-measure (Neubig et al., 2019) to quantify the translation accuracy of specific types of words.

Compared with the source-original data, using only the target-original data greatly reduces the translation accuracy of content words, which we attribute to the divergence of the content word distributions between the source- and target-original data. The results also indicate that indiscriminately using all the training data can not consistently improve the translation adequacy of content words over using only source-original data, and in some cases using all the data is even harmful to the adequacy on content words. Table 5 shows an example, which suggests that using only the target-original data tends to omit content words. This problem is potentially caused by that some content words at the source-side are less or even not visible in the target-original data, and indiscriminately adding target-original data induces a distribution shift on the content word distribution.

Using additional target-original data only slightly improves the structural fluency. Recently, Edunov et al. (2020) claim that using additional back-translated data can improve translation fluency. Target-original bilingual data is similar to back-translated data since both of them are constructed by translating sentences from the target language into the source language. One question

[2]We only list the results on En⇔Zh due to space limit. Please refer to Table 14 in the Appendix for the translation quality on other language pairs.

Input	大闸蟹是巴城最为知名的形象代言人。
Refer.	The hairy crab is the most famous image spokesperson in Bacheng.
Target Orig.	It is one of the city's most well-known image spokesmen.
Source Orig.	Hairy crabs are the most well-known image spokesmen of Bacheng.
Both	It is the best-known icon of Bacheng.

Table 5: An example of the outputs of NMT models trained on different sets of data. Using the target-original data tends to omit content words.

naturally arises: *can target-original bilingual data improve the fluency of NMT models?*

To answer the above question, we measure the fluency of outputs with language models trained on the monolingual data as described in Section 2. Previous study finds that different perplexities could be caused by specific contents rather than structural differences (Lembersky et al., 2011). Specifically, some source-original contents are of low frequency in the target-language monolingual data (e.g., "Bacheng" in Table 5), thus the language model trained on the target-language monolingual data tends to assign higher perplexities to outputs containing more source-original content words. To rule out this possibility and check whether the outputs are structurally different, we follow Lembersky et al. (2011) to abstract away from the content-specific features of the outputs to measure their fluency at the syntactic level. Table 6 shows the results. Although using only the source-original data results in high perplexities measured by vanilla language models, the perplexities of NMT models trained on different data are close to each other at the syntactic level. Using additional target-original data only slightly reduces the perplexity at the syntactic level over using only the source-original data.

4 Addressing Language Coverage Bias

In Section 3 we show that the target-original data performs poorly in translating content words due to the language coverage bias. Accordingly, simply using the full training data without distinguishing the original languages is sub-optimal for model training. Based on these findings, we propose to address the language coverage bias by explicitly distinguishing between the source- and the target-original data (Section 4.1). We then investigate whether the performance improvement still holds in

Data	No Abs.		Cont. Abs.	
Origin	*PPL*	*Diff.*	*PPL*	*Diff.*
WMT20 En⇒Zh				
Target	38.4	−6.3%	14.0	−2.8%
Source	44.0	+7.3%	14.6	+1.4%
Both	41.0	-	14.4	-
WMT20 En⇐Zh				
Target	25.9	−5.8%	13.4	−3.6%
Source	31.0	+12.7%	14.2	+2.2%
Both	27.5	-	13.9	-

Table 6: Translation fluency measured by the perplexities (i.e., PPL) of language models with different levels of lexical abstraction, "Diff." means the relative change with respect to "Both". "No Abs." denotes no abstraction (i.e., vanilla LM), "Cont. Abs." denotes abstracting all content words with their corresponding POS tags. The results are reported on the validation sets.

the monolingual data augmentation scenario (Section 4.2), where the language coverage bias problem is more severe due to the newly introduced dataset in source or target language.

4.1 Bilingual Data Utilization

In this section, we aim to improve bilingual data utilization through explicitly distinguishing between the source- and target-original training data.

Methodology We distinguish original languages with two simple and effective methods:

- *Bias-Tagging*: Tagging is a commonly-used approach to distinguishing between different types of examples, such as different languages (Aharoni et al., 2019; Riley et al., 2020) and synthetic vs authentic examples (Caswell et al., 2019). In this work, we attach a special tag to the source side of each target-original example, which enables NMT models to distinguish it from the source-original ones in training.

- *Fine-Tuning*: Fine-tuning (Luong and Manning, 2015) is a useful method to help knowledge transmit among data from different distributions. We pre-train NMT models on the full training data that consists of both the source- and target-original data, and then fine-tune them on only the source-original data. For fair comparison, the total training steps of the pre-training and fine-tuning stages are the same as the baseline.

Translation Performance Table 7 depicts the results on the benchmarking datasets. For comparison, we also list the results of several baselines using the vanilla Transformer architecture trained on the constrained bilingual data in the WMT20 competition (Barrault et al., 2020). Clearly, both the bias tagging and fine-tuning approaches consistently improve translation performance on all benchmarks, which confirms our claim of the necessity of explicitly distinguishing target-original examples in model training.

Analysis Recent studies have shown that generating human-translation like texts as opposed to original texts can improve the BLEU score (Riley et al., 2020). To validate that the improvement is partially from alleviating the content-dependent language coverage bias, we examine the translation adequacy of content words on the test sets, as listed in Table 8. The results indicate that explicitly distinguishing between the source- and target-original data improves the translation of content words (e.g., nouns), which is closely related to the language coverage bias problem. Table 9 lists the translation fluency at the syntactic level, where the proposed approaches maintain the syntactic fluency.

4.2 Monolingual Data Augmentation

In this section, we aim to provide some insights where monolingual data augmentation improves translation performance, and investigate whether our approach can further improve model performance in this scenario that potentially suffers more from the language coverage bias problem.

For fair comparison across language pairs, we augment NMT models with the same English monolingual corpus as described in Section 2. We down-sample the large-scale monolingual corpus to the same amount as that of the bilingual corpus in each language pair, in order to rule out the effect of the scale of synthetic data (Edunov et al., 2018; Fadaee and Monz, 2018). We use back-translation (Sennrich et al., 2016a) to augment the English monolingual data for the task of translating from another language to English ("X⇒En"), and use forward-translation for the task in the opposite translation direction ("En⇒X"). Table 10 lists the results, where several observations can be made.

Explaining Data Augmentation with Language Coverage Bias Concerning the monolingual data augmentation methods (Rows 3-4), the vanilla

Method	En-Zh		En-Ja		En-De		Average
	⇒	⇐	⇒	⇐	⇒	⇐	
WMT20 Systems							
Shi et al. (2020)	38.6	28.8	-	-	-	-	-
Zhang et al. (2020)	40.8	-	34.8	20.4	-	-	-
Molchanov (2020)	-	-	-	-	31.9	39.6	-
Our Implemented Systems							
Baseline	42.3	28.4	35.8	20.9	32.3	41.4	33.5
Tag	**43.4**$^{\Uparrow}$	29.2$^{\uparrow}$	36.3	**21.9**$^{\Uparrow}$	32.7	**42.5**$^{\Uparrow}$	34.3
Tune	43.3$^{\Uparrow}$	**29.7**$^{\Uparrow}$	**36.6**$^{\uparrow}$	21.8$^{\uparrow}$	**32.9**$^{\uparrow}$	42.2$^{\uparrow}$	**34.4**

Table 7: Sacre BLEU reported on the WMT20 test sets. "Tag" and "Tune" denote the bias-tagging and fine-tuning, respectively. We highlight the **highest** score in bold and the second-highest score with underlines. "↑/⇑" denotes significantly better than the baseline with $p < 0.05$ and $p < 0.01$, respectively. For comparison, we list three systems that use vanilla Transformer models trained on the bilingual data in the WMT20 competition.

Method	En⇒Zh			En⇐Zh		
	noun	*verb*	*adj*	*noun*	*verb*	*adj*
Baseline	70.7	61.0	67.9	60.2	43.6	61.4
Tag	**72.3**	**62.3**	67.9	60.7	43.9	**62.2**
Tune	71.8	61.9	**68.4**	**61.1**	**44.1**	**62.2**

Table 8: F-measure of different types of content words on the WMT20 En⇔Zh test sets. Results on other languages can be found in Appendix (Table 15).

Data	En-Zh		En-Ja		En-De	
Origin	⇒	⇐	⇒	⇐	⇒	⇐
Baseline	**13.7**	**13.4**	**17.4**	**15.4**	**16.3**	**17.8**
Tag	**13.7**	**13.4**	17.5	**15.4**	16.4	**17.8**
Tune	13.8	**13.4**	**17.4**	**15.4**	**16.3**	17.9

Table 9: PPL at the syntactic level on the test sets. We abstract the content words to rule out the language coverage bias when measuring fluency.

back-translation (Row 3) harms the translation performance on average, while the vanilla forward-translation improves the performance, which is consistent with the findings in previous studies (Edunov et al., 2020; Marie et al., 2020). Caswell et al. (2019) have shown that the tagging strategy works for back-translation while fails for forward-translation, and our results confirm these findings. Both phenomena can be attributed in part to the language coverage bias problem. Back-translated data originates from the target language, and thus suffers more from the language coverage bias problem. Accordingly, directly using the back-translated data is sub-optimal, while tagged back-translation recovers translation performance by distinguishing training examples with different origins, which is consistent with our results in Table 7. In contrast, the language coverage bias problem does not exist for source-side monolingual data (i.e. the same original language). Therefore, the vanilla forward-translation can improve translation performance, while tagged forward-translation performs worse.

Improving Data Augmentation Our approach (Row 2) achieves comparable improvements of translation performance with the monolingual data augmentation approaches (e.g. averaged BLEU: 31.2 vs. 30.7, and 37.6 vs. 37.9), while we do not use additional monolingual data to train the models.[3] Combining them can further improve performance (Rows 5-6), indicating that the two types of approaches are complementary to each other. This is straightforward, since our approach better exploits the bilingual data, while data augmentation introduces new knowledge from additional monolingual data. In addition, our approach consistently improves performance over both vanilla and tagged augmentation approaches, making it more robust in practical application across datasets.

5 Related Work

Our work is inspired by three lines of research in the NMT community.

[3]The monolingual data is only used to detect the original languages of training data and is invisible in model training.

#	Monolingual		Bilingual	X⇒En				En⇒X			
	Data	*Tagging*	*Fine-Tune*	*Zh*	*Ja*	*De*	*Ave.*	*Zh*	*Ja*	*De*	*Ave.*
1	×	n/a	×	28.4	20.9	41.4	30.2	42.3	35.8	32.3	36.8
2			✓	*29.7*	*21.8*	*42.2*	*31.2*	*43.3*	*36.6*	*32.9*	*37.6*
3	✓	×	×	28.9	21.2	38.1	29.4	*44.2*	*36.8*	*32.8*	*37.9*
4		✓		*29.4*	*21.2*	*41.5*	*30.7*	43.1	36.4	32.3	37.3
5		×	✓	30.4	22.1	42.3	31.6	**45.1**	**37.6**	**33.4**	**38.7**
6		✓		**30.6**	**22.2**	**42.7**	**31.8**	44.7	36.9	33.3	38.3

Table 10: Translation performance of augmenting English monolingual data with different strategies: back-translation for X⇒En tasks (blue cells), and forward-translation for En⇒X tasks (red cells). "Tagging" denotes adding a special tag to each synthetic sentence pair (Caswell et al., 2019). "Fine-Tune" denotes fine-tuning the pre-trained NMT models on the source-original bilingual data, as described in Section 4.1.

5.1 Translationese

Recently, the effect of translationese in NMT evaluation has attracted increasing attention (Zhang and Toral, 2019; Bogoychev and Sennrich, 2019; Edunov et al., 2020; Graham et al., 2020). Graham et al. (2020) show that the source-side translationese texts can potentially lead to distortions in automatic and human evaluations. Accordingly, the WMT competition starts to use only source-original test sets for most translation directions since 2019. Our study reconfirms the necessity of distinguishing the source- and target-original examples and takes one step further to distinguish examples in training data. Complementary to previous works, we investigate the effect of language coverage bias on machine translation, which is related to the content bias rather than the language style difference. Shen et al. (2021) also reveal the context mismatch between texts from different original languages. To alleviate this problem, they proposed to combine back- and forward-translation by introducing additional monolingual data, while we focus on better exploiting bilingual data by distinguishing the original languages, which is also helpful for back- and forward-translation.

Lembersky et al. (2011, 2012) propose to adapt machine translation systems to generate texts that are more similar to human-translations, while Riley et al. (2020) propose to model human-translated texts and original texts as separate languages in a multilingual model and perform zero-shot translation between original texts. Riley et al. (2020) and our work both aim to better utilize the bilingual training data. They aim to guide NMT models to produce original text, while we focus on improving translation adequacy by alleviating the language coverage bias problem.

5.2 Data Augmentation

Concerning model training, recent works find that back-translation can harm the translation of source-original test set, and attribute the quality drop to the stylistic and content-independent differences between translationese and original texts (Edunov et al., 2020; Marie et al., 2020). In this work, we empirically show that language coverage bias is another reason for the performance drop of back-translation, as well as the different performances between tagged forward-translation and tagged back-translation (Caswell et al., 2019). In addition, we show that our approach is also beneficial for data augmentation approaches, which can further improve the translation performance over both back-translation and forward-translation.

5.3 Domain Adaptation

Since high-quality and domain-specific parallel data is usually scarce or even unavailable, domain adaptation approaches are generally employed for translation in low-resource domains by leveraging out-of-domain data (Chu and Wang, 2018). Languages can be also regarded as different domains, since articles in different languages cover different topics (Bogoychev and Sennrich, 2019). Starting from this intuition, we distinguish examples with different original languages with tagging (Aharoni et al., 2019) and fine-tuning (Luong and Manning, 2015), which are commonly-used in domain adaptation and multi-lingual translation tasks.

Our work also benefits domain adaptation: distinguishing original languages in general domain

data consistently improves translation performance of NMT models in several specific domains (Table 16 in Appendix), making these models better start points for further domain adaptation.

6 Conclusion and Future Work

In this work, we first systematically examine why the language coverage bias problem is important for NMT models. We conducted extensive experiments on six WMT20 translation benchmarks. Empirically, we find that source-original data and target-original data differ significantly at the text content, and using target-original data together without discrimination is sub-optimal. Based on these observations, we propose two simple and effective approaches to distinguish the source- and target-original training data, which obtain consistent improvements in all benchmarks.

Furthermore, we link language coverage bias to two well-known problems in monolingual data augmentation, namely the performance drop of back-translation, and the different behaviors between tagged back-translation and tagged forward-translation. We show that language coverage bias can be considered as another reason for these problems, and fine-tuning on the source-original bilingual training data can further improve performance over both back- and forward-translation.

Future directions include exploring advanced methods to better alleviate the language coverage bias problem, as well as validating on other language pairs. It is also interesting to investigate the language coverage bias problem in multilingual translation, where we can better understand this problem by considering language family.

Acknowledgments

We thank all anonymous reviewers for their insightful comments and suggestions for this work. We also sincerely thank Cunxiao Du and Wenxiang Jiao for their valuable help and advice. This work was supported by the National Key R&D Program of China (No. 2017YFB0202204), National Natural Science Foundation of China (No. 61925601, No. 61772302), and the Tencent AI Lab Rhino-Bird Focused Research Program.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *NAACL*.

Mona Baker, Gill Francis, and Elena Tognini-Bonelli. 1993. Corpus linguistics and translation studies: Implications and applications. *Text and technology: In honour of John Sinclair*, 233:250.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, et al. 2020. Findings of the 2020 conference on machine translation (wmt20). In *WMT*.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *WMT*.

Nikolay Bogoychev and Rico Sennrich. 2019. Domain, translationese and noise in synthetic data for neural machine translation. *arXiv preprint arXiv:1911.03362*.

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. In *WMT*.

Chenhui Chu and Rui Wang. 2018. A survey of domain adaptation for neural machine translation. In *COLING*.

Dave D'Alessio and Mike Allen. 2000. Media bias in presidential elections: A meta-analysis. *Journal of communication*, 50(4):133–156.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *EMNLP*.

Sergey Edunov, Myle Ott, Marc'Aurelio Ranzato, and Michael Auli. 2020. On the evaluation of machine translation systems trained with back-translation. In *ACL*.

Marzieh Fadaee and Christof Monz. 2018. Back-translation sampling by targeting difficult words in neural machine translation. In *EMNLP*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. BLEU might be guilty but references are not innocent. In *EMNLP*.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2020. Statistical power and translationese in machine translation evaluation. In *EMNLP*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL*.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. 2011. Language models for machine translation: Original vs. translated texts. In *EMNLP*.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. 2012. Adapting translation models to translationese improves smt. In *EACL*.

Jianhua Lin. 1991. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151.

Minh-Thang Luong and Christopher D Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *IWSLT*.

Benjamin Marie, Raphael Rubino, and Atsushi Fujita. 2020. Tagged back-translation revisited: Why does it really work? In *ACL*.

Alexander Molchanov. 2020. Promt systems for wmt 2020 shared news translation task. In *WMT*.

Graham Neubig, Zi-Yi Dou, Junjie Hu, Paul Michel, Danish Pruthi, and Xinyi Wang. 2019. compare-mt: A tool for holistic comparison of language generation systems. In *NAACL*.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook fair's wmt19 news translation task submission. In *WMT*.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. In *WMT*.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *WMT*.

Vikas Raunak, Siddharth Dalmia, Vivek Gupta, and Florian Metze. 2020. On long-tailed phenomena in neural machine translation. In *Findings of EMNLP*.

Parker Riley, Isaac Caswell, Markus Freitag, and David Grangier. 2020. Translationese as a language in "multilingual" nmt. In *ACL*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *ACL*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *ACL*.

Jiajun Shen, Peng-Jen Chen, Matthew Le, Junxian He, Jiatao Gu, Myle Ott, Michael Auli, and Marc'Aurelio Ranzato. 2021. The source-target domain mismatch problem in machine translation. In *EACL*.

Tingxun Shi, Shiyu Zhao, Xiaopu Li, Xiaoxue Wang, Qian Zhang, Di Ai, Dawei Dang, Zhengshan Xue, and Jie Hao. 2020. Oppo's machine translation systems for wmt20. In *WMT*.

Antonio Toral. 2019. Post-editese: an exacerbated translationese. In *MT Summit*.

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 *NeurIPS*.

Vered Volansky, Noam Ordan, and Shuly Wintner. 2015. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118.

Felix Wu, Angela Fan, Alexei Baevski, Yann Dauphin, and Michael Auli. 2018. Pay less attention with lightweight and dynamic convolutions. In *ICLR*.

Mike Zhang and Antonio Toral. 2019. The effect of translationese in machine translation test sets. In *WMT*.

Yuhao Zhang, Ziyang Wang, Runzhe Cao, Binghao Wei, Weiqiao Shan, Shuhan Zhou, Abudurexiti Reheman, Tao Zhou, Xin Zeng, Laohu Wang, Xiaoqian Liu, Xunjuan Zhou, Yongyu Mu, Jingnan Zhang, Yinqiao Li, Bei Li, Tong Xiao, and Jingbo Zhu. 2020. The niutrans machine translation systems for wmt20. In *WMT*.

A Appendices

A.1 Data Preprocessing

We used all the parallel corpora provided by WMT20 and filtered sentences that are longer than 250 words. We tokenized English and German sentences with Moses (Koehn et al., 2007), and segmented Chinese and Japanese sentences with Jieba[4] and Mecab[5] respectively. We employed Byte pair encoding (BPE) (Sennrich et al., 2016b) with 32K merge operations for all language pairs. Specifically, we jointly trained the BPE code on both sides in En⇔De and independently learned the BPE code on each side in En⇔Zh and En⇔Ja.

As for the monolingual data, we combined the newscrawl data from 2017 to 2019 for English and German. Since the newscrawl corpora for Chinese and Japanese are significantly smaller, we augmented these two languages with the commoncrawl corpus. We preprocessed the monolingual data with the same rules as parallel data. Finally, we randomly selected 41.0M sentences for each language (i.e., En, De, Zh, Ja), which were used to train the language detection models. For data augmentation, to rule out the effect of the ratio between synthetic and authentic data, we downsampled the monolingual data to the same amount as the bilingual data for each language pair.

We used spaCy[6] to perform the Part-Of-Speech (POS) tagging for each language. Nouns, verbs, and adjectives belong to content words and the others belong to function words.

A.2 More Details of Model Training

In this work, we generally followed the default hyper-parameters used in Vaswani et al. (2017) except the batch size. Recent studies showed that training on large batches can further boost model performance (Ott et al., 2018; Wu et al., 2018). Accordingly, we followed them to train models with batches of approximately 460k tokens, using Adam (Kingma and Ba, 2015) with $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\epsilon = 10^{-8}$. We used the same cosine learning rate schedule as Wu et al. (2018), where the learning rate was warmed up linearly in the first 10K steps, and then decayed following a cosine rate within a single cycle. By default, NMT models and language models were both trained for 30k steps with the aforementioned batch size. Each model was trained using 8 NVIDIA V100 GPUs for about 20 hours.

[4]https://github.com/fxsjy/jieba
[5]https://taku910.github.io/mecab
[6]https://spacy.io

A.3 Effect of Detection Methods on Translation Performance

To further compare our proposed original language detection method and the FT classifier (Riley et al., 2020), we fine-tune the NMT model pre-trained on the whole training set using the source-original data detected by the two methods. Note that the two detection methods are developed using the same monolingual data sets. For fair comparison, the fine-tuning sets are of the same amount (50% of the whole training set) between the two methods in this experiment. Table 11 lists the results, indicating that our method performs better in detecting original languages in large-scale parallel data.

Fine-Tune Data	BLEU
×	27.5
FT	27.8
Ours	**28.4**

Table 11: Effect of original language detection methods. The results are reported on the validation set of the Zh⇒En translation task.

A.4 Divergence of Vocabulary Distributions

In this section, we report the JS divergence of the vocabulary distributions in more cases. Table 12 lists the results for different ratios $R\%$ on En⇔Zh, and Table 13 shows the results on all language pairs. The results show that the divergence of vocabulary distributions between the source- and target-original data is substantially larger than that between randomly split data, which reconfirms the existence of language coverage bias.

A.5 Effect of Language Coverage Bias for Other Language Pairs

Table 14 lists the translation adequacy of NMT models trained on only the source- or target-original data and on both of them. The results are reported on En⇔De and En⇔Ja, which exhibit the same trend as that on En⇔Zh (Table 4 in the main paper), indicating that the target-original data performs poorly in translating content words.

Data	10%			30%			50%		
	All	*Content*	*Function*	*All*	*Content*	*Function*	*All*	*Content*	*Function*
Random	20	51	0	7	18	0	4	10	0
S vs T	2660	5096	961	1371	2731	486	745	1503	261

Table 12: JS divergence ($\times 10^{-5}$) of the vocabulary distributions between the source- and target-original training data ("S vs T") for different labeled ratios on En⇔Zh. For reference, we also report the JS divergence between two sets of randomly selected examples ("Random", non-overlap).

Data	En-Zh			En-Ja			En-De		
	All	*Content*	*Function*	*All*	*Content*	*Function*	*All*	*Content*	*Function*
Random	4	10	0	8	17	0	2	4	0
S vs T	745	1503	261	1687	2910	666	870	1622	250

Table 13: JS divergence ($\times 10^{-5}$) of the vocabulary distributions between the source- and target-original training data for different language pairs. 50% examples are treated as source-original and the others are treated as target-original. For reference, we also report the JS divergence between randomly selected 50% examples and the others ("Random", non-overlap).

Data	En⇒Ja			En⇐Ja			En⇒De			En⇐De		
Origin	*noun*	*verb*	*adj*	*noun*	*verb*	*adj*	*noun*	*verb*	*adj*	*noun*	*verb*	*adj*
Target	60.9	47.7	62.1	44.5	29.8	46.2	70.5	54.3	58.4	70.8	53.6	67.0
Source	**61.4**	**51.8**	**63.5**	49.5	31.8	50.1	72.1	55.0	60.3	**75.3**	55.3	70.2
Both	61.3	51.7	63.2	**50.7**	**32.1**	**50.4**	**72.7**	**56.6**	**60.4**	74.9	**55.7**	**71.0**

Table 14: Translation adequacy of different types of content words measured by F-measure (Neubig et al., 2019). The results are reported on the validation sets.

Method	En⇒Ja			En⇐Ja			En⇒De			En⇐De		
	noun	*verb*	*adj*	*noun*	*verb*	*adj*	*noun*	*verb*	*adj*	*noun*	*verb*	*adj*
Baseline	62.0	53.0	59.1	54.0	35.5	50.4	66.7	48.1	53.6	75.0	54.0	70.2
Tag	62.5	53.3	61.1	**55.7**	36.5	**52.4**	**67.1**	48.6	**54.0**	**76.1**	54.4	70.7
Tune	**62.8**	**53.7**	**61.7**	55.3	**36.9**	51.8	**67.1**	**48.7**	**54.0**	75.7	**54.8**	**70.8**

Table 15: Translation adequacy of different types of content words measured by F-measure (Neubig et al., 2019). The results are reported on the test sets.

Domain	Baseline	Ours
Business	40.4	**40.8**
Crime	34.8	**35.5**
Entertainment	28.8	**30.0**
Politics	39.5	**40.3**
Sci-Tech	38.2	**39.9**
Sport	**31.5**	**31.5**
World	38.8	**38.9**
Overall	36.6	**37.2**

Table 16: Transformer performance on the validation set of the En⇒Zh task. We split the whole validation set into several parts by the domain tag. "Ours" denotes the "Bias-Tagging" approach as described in Section 4.1. The results indicate that distinguishing data with different original languages in the general domain training data can improve the performance of NMT models in many specific domains, making the models better start points for further domain adaptation.

A.6 Translation Adequacy on Test Sets for Other Language Pairs

We report the translation adequacy on test sets for En⇔De and En⇔Ja in Table 15, corresponding to Table 8 in the main paper. The results show that explicitly distinguishing the source- and target-original training data can consistently improve the translation adequacy for content words on all the six translation tasks.

A.7 Translation Performance in Specific Domains

We evaluate NMT models trained with and without explicit distinguishing between the source- and target-original data in several specific domains. The results are shown in Table 16, suggesting that our method can improve the translation performance of NMT models in several specific domains, which can be combined with further domain adaptation approaches.