Original or *Translated*? A Causal Analysis of the Impact of Translationese on Machine Translation Performance

Anonymous ACL submission

Abstract

Human-translated text displays distinct features from naturally written text in the same language. This phenomena, known as *translationese*, has been argued to confound the machine translation (MT) evaluation. Yet, we find that existing work on translationese neglects some important factors and the conclusions are mostly correlational but not causal. In this work, we collect CAUSALMT, a dataset where the MT training data are also labeled with the human translation directions. We inspect two critical factors, the train-test alignment (whether the human translation directions in the training and test sets are aligned), and data-model alignment (whether the model learns in the same direction as the human translation direction in the dataset). We show that these two factors have a large causal effect on the MT performance, in addition to the test-model misalignment highlighted by existing work on the impact of translationese in the test set. In light of our findings, we provide a set of suggestions for MT training and evaluation.[1]

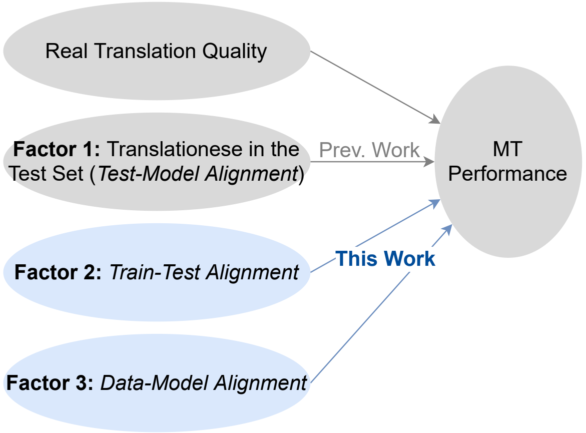

Real Translation Quality
Factor 1: Translationese in the Test Set (*Test-Model Alignment*)
Prev. Work
MT Performance
Factor 2: *Train-Test Alignment*
This Work
Factor 3: *Data-Model Alignment*

Figure 1: Three different factors illustrate the impact of translationese on MT performance. Previous work focuses on how translationese in the test set (Factor 1) inflates BLEU score and makes it favor some translation systems over others. Our work investigates the causal effect of the other two key factors, the train-test alignment (Factor 2; i.e., whether the training set and the test set are collected with the same human translation direction), and data-model alignment (Factor 3; i.e., whether the dataset collection direction and model translation direction are the same).

1 Introduction

MT has long been concerned with the artifacts introduced by *translationese*, the human-translated text that is systematically different from naturally written text in the same language, or original text (Toury, 1980; Gellerstam, 1986; Toury, 1995; Baker, 1993; Baroni and Bernardini, 2006). For a translation system translating from language X to language Y, there can be two types of test data: sentences that originated in language X and are human-translated into language Y (denoted as $X\overset{\text{H}}{\rightarrow}Y$), and sentences that originated in language Y and human-translated into language X (denoted as $Y\overset{\text{H}}{\rightarrow}X$). The main concern raised by this distinction of the two sets is whether the reported performance on a mixed test set truly reflects the actual translation quality. Previous work in MT has shown that translationese is a confounder in evaluating translation quality (Lembersky et al., 2012; Toral et al., 2018; Läubli et al., 2018; Freitag et al., 2020).

Recent studies on causality have also brought to attention the importance of distinguishing the *data-model alignment*, namely whether the data collection direction is the same as or opposite to the model direction, also known as *causal* or *anticausal* learning (Jin et al., 2021; Veitch et al., 2021; Schölkopf et al., 2012). If the dataset is collected by human annotators who see the input X and produce an output Y, then learning an X-to-Y model is causal learning, and learning a Y-to-X model is anticausal learning.

In this work, we study the artifacts in MT

[1] Our code and data will be open-sourced after acceptance.

brought by translationese from the viewpoint of causality, specifically, the interaction between the data and model direction. We study two factors of variation in MT: *human translation direction* (in both the training and the test set) and *model translation direction*. Then, we study the effect of translationese in the test set as the *test-model alignment* problem, and causal/anticausal learning as the *data-model alignment* problem. Further, we identify another important factor, the *train-test alignment* problem, namely, whether the training set and the test set are collected with the same human translation direction. Given these three factors that influence MT performance, we study the interaction in Figure 1. While previous work has mainly studied the impact of test-model alignment on MT performance (Toral et al., 2018; Graham et al., 2020; Edunov et al., 2020), we show that train-test alignment and data-model alignment can also have a large causal impact on the MT performance. This impact can sometimes even overshadow the effect of test-model alignment analyzed in previous work.

We use causal inference (Pearl, 2009; Peters et al., 2017) to analyze the causal effects of these key factors on MT performance, beyond previous work which is mainly based on correlations (Graham et al., 2020). Specifically, our causal analysis isolates and controls for other key causal factors in translation performance, such as sentence length and content.

We build CausalMT, a new dataset on five language pairs labeled with the human translation directions, and statistically verify that translationese tend to be simpler and more verbose, corroborating previous observations on translationese (Toury, 1980; Gellerstam, 1986; Toury, 1995; Baker, 1993). Then, we rigorously analyze CausalMT, leading to the following new insights and contributions:

C1. Previous work claims that translationese in the test set inflates MT model performance and thus suggests removing the translationese-to-original half of the test set (Toral et al., 2018; Zhang and Toral, 2019; Graham et al., 2020; Barrault et al., 2019). Our work shows that the translationese-to-original half of the test set does not necessarily inflate MT performance in all cases. In some cases, it can even be more challenging than the other half, depending on the human translation direction in the training corpus. Hence, we suggest still reporting performance on both test sets, but also reporting the training data direction if available.

C2. Previous work (Burlot and Yvon, 2018) claims that back translation (BT) (Sennrich et al., 2016) is usually more effective than supervised training (ST) (He et al., 2019). Our work shows that BT is not necessarily better than ST in all cases. This result too depends on how the pseudo-parallel corpus aligns with the human translation direction in the test set. We suggest choosing BT or ST depending on this alignment.

C3. Previous work claims that BT's performance improvement is largely reflected on the translationese-to-original half of the test set, but the improvement is very small on the other half (Toral et al., 2018; Freitag et al., 2019). Our work shows that the improvement of BT can be larger on the other half of the test set as well, as long as the pseudo-parallel corpus aligns with the human translation direction in the test set.

C4. Our work shows that data-model alignment also has a large causal effect on the MT performance, with up to 12.25 BLEU scores after adjusting for other covariates using backdoor adjustment (Pearl, 1995).

2 CausalMT Dataset

To investigate the effect of train-test alignment and data-model alignment, we need to collect translation data in different human translation directions.[2]

2.1 Data Collection

To construct our CausalMT dataset consisting of a large number of translation pairs labeled with the human translation direction, we use the EuroparlExtract toolkit (Ustaszewski, 2019) to filter translation pairs by meta-information (e.g., the tag specifying the original language of the speaker). Specifically, in the EuroParl corpus (Koehn, 2005), we iterate over each transcript that has an origination label and mark a sentence as original text if

[2] Most existing datasets do not distinguish the human translation direction for the training set (Kolias et al., 2014; Barrault et al., 2019). Some works train a classifier to identify the human translation direction (Kurokawa et al., 2009; Riley et al., 2020), but they are not our ideal choice since this classification may interact with the domain difference of the two directions (Rabinovich and Wintner, 2015). Our dataset can be considered as an extended version of the dataset collected in Jin et al. (2021), but ours is significantly larger to enable the various analyses in our study.

	De$\xrightarrow{\text{H}}$En	En$\xrightarrow{\text{H}}$De	De$\xrightarrow{\text{H}}$Fr	Fr$\xrightarrow{\text{H}}$De	En$\xrightarrow{\text{H}}$Fr	Fr$\xrightarrow{\text{H}}$En	En$\xrightarrow{\text{H}}$Es	Es$\xrightarrow{\text{H}}$En	Es$\xrightarrow{\text{H}}$Fr	Fr$\xrightarrow{\text{H}}$Es
Training Size	248K	248K	220K	220K	203K	203K	93K	93K	92K	92K
# Words/Sample	22.4/25.5	23.9/22.9	22.6/28.7	30.4/25.4	24.5/28.9	30.5/27.5	24.0/25.7	31.9/31.6	32.4/36.5	30.5/27.9
# Sents/Sample	1.05/1.04	1.02/1.04	1.05/1.86	1.94/1.07	1.03/1.89	1.95/1.05	1.03/1.05	1.08/1.08	1.09/2.18	1.95/1.07
Passive Voice (%)	-/12.90	11.48/-	-/-	-/-	11.70/-	-/13.45	11.49/-	-/14.94	-/-	-/-
Vocab Size	119K/37K	40K/113K	108K/49K	55K/106K	40K/53K	56K/38K	29K/46K	47K/26K	47K/38K	42K/48K
Expansion Factor	en:de=1.13	en:de=1.04	fr:de=1.26	fr:de=1.19	fr:en=1.18	fr:en=1.10	es:en=1.06	es:en=1.01	fr:es=1.12	fr:es=1.09

Table 1: Detailed characteristics of the CausalMT dataset, including the number of words per sample, number of sentences per sample, percentage of samples with passive voice, vocabulary size, and the expansion factor. The expansion factor from language X to language Y $(X{:}Y)$ is calculated by the average word count per sample in language X divided by the average word count per sample in language Y.

the original language of the speaker is the same as the language this sentence is in, or otherwise mark it as the translated text. After extracting the direction-labeled language pairs, we remove all duplicates in the entire dataset. Since our study needs to compare training on parallel corpora of the same language pair but with two different human translation directions, e.g., De$\xrightarrow{\text{H}}$En and En$\xrightarrow{\text{H}}$De, we control the size of the two corpora to be the same by downsampling the larger set.

Among all language pairs we can obtain, we keep five language pairs with the largest number of data samples. As shown in Table 1, the CausalMT dataset contains over 200K training data for three language pairs and over 90K data for the other two language pairs. The development set and test set contain 1K and 2K data samples for all language pairs in each direction.

2.2 Dataset Characteristics

We analyze the characteristics of the CausalMT dataset in light of how translated text differs from naturally written text in the same language.

Our findings echo with the observations by previous work on the distinct features of translationese (Toury, 1980; Gellerstam, 1986; Toury, 1995; Baker, 1993; Baroni and Bernardini, 2006; Volansky et al., 2015). For example, translationese tends to be simpler and more standardized (Baker, 1993; Toury, 1995; Laviosa-Braithwaite, 1998), such as having a smaller vocabulary and using certain discourse markers more often (Baker, 1993, 1995, 1996). Translationese also tends to be influenced by the source language in terms of its lexical and word order choice (Gellerstam, 1986).

In the CausalMT data, we observe three properties. (1) Within each language pair (e.g., German and English), the same language's *translationese always has a smaller vocabulary* than its naturally written text corpus. For example, the translationese German in En$\xrightarrow{\text{H}}$De has only 113K vocabulary, which is 5K smaller than the vocabulary of the German corpus in De$\xrightarrow{\text{H}}$En. (2) *Translationese tends to be more verbose*. For each language pair, we calculate the expansion factor from language X to language Y $(X{:}Y)$ as the average word count per sample in language X divided by the average word count per sample in language Y. For example, for each (English, German) translation pair, the number of English words is 1.13 times that of German words when English is the translationese (i.e., en:de expansion factor=1.13). On the other hand, the en:de expansion factor is only 1.04 when English is the naturally written text. (3) We use a syntax-based parser to detect the percentage of samples with passive voice in English.[3] There is a clear distinction that *translationese English tends to use more passive voice than original English*, e.g., 14.94% translationese samples in passive voice in the Es$\xrightarrow{\text{H}}$En corpus in contrast with 11.49% original English samples in the reverse direction.

3 The Overshadowing Effect of Train-Test Alignment

The first analysis of this paper aims to expand the existing understanding of the relationship between translationese and MT performance by consideringthe effect of the train-test alignment.

Previous work observes that the translationese-to-original test set inflates the score. To evaluate a model with the X-to-Y translation direction, traditionally, the test set is a mixture of two halves, one with the human translation direction $X\xrightarrow{\text{H}}Y$ (aligned) and the other $Y\xrightarrow{\text{H}}X$ (unaligned, or translationese-to-original) (Bojar et al., 2018).

Previous studies propose that the unaligned, translationese-to-original test set is easier to translate than the other aligned test set because transla-

[3] We use this passive voice checker (only available in English).

tionese inputs are easy for the MT model to handle (Toral et al., 2018; Zhang and Toral, 2019; Graham et al., 2020). The inflated test performance caused by translationese has long been speculated (Lembersky et al., 2012; Toral et al., 2018; Läubli et al., 2018), and, recent work has statistically verified the correlation (Graham et al., 2020).

With the previous understanding, some works suggest removing the unaligned half of the test set (Toral et al., 2018; Zhang and Toral, 2019; Graham et al., 2020), which was adopted by the 2019 WMT shared task (Barrault et al., 2019), whereas others suggest keeping both but report the performance separately (Freitag et al., 2019; Edunov et al., 2020). The motivations from the two sides are that in the unaligned half, although its the source text being translationese is an easy input to the model, its target text being naturally written text makes the evaluation more natural.

This "inflation" depends on train-test alignment. We take a step back from the argument on whether the unaligned test set positively or negatively affects the MT performance evaluation. Instead, we call attention to the fact that, beyond the test-model alignment, there can be other factors also playing a critical in the MT performance evaluation, i.e., the train-test alignment.

For a given machine translation task to learn the X-to-Y translation, there can be two questions: the question by previous work is whether we should use the test set aligned with the model translation direction (T1) or the test set unaligned with the model translation direction (T2) to evaluate the model fairly, whereas the question answered by our work is *which training data should be used to achieve the best performance*.

Our analysis aims to obtain causal conclusions on how intervening on the train-test alignment affects the MT performance. Therefore, we control all other possible confounders. For each language pair, we control the total training data size to be the same[4] when varying the portion of data in two directions. We also enumerate all other possible interventions, such as varying the model in two model translation directions and reporting performance on two different halves of the test set with two human translation directions. We also control that all translation models use the same Transformer architecture (Vaswani et al., 2017) by fairseq (Ott et al., 2019), with experimental details in Appendix C.

De-to-En Translation			**En-to-De Translation**		
$\alpha\%$	T1 (de, en*)	T2 (de*, en)	$\alpha\%$	T1 (en, de*)	T2 (en*, de)
0%	24.68	35.86	0%	21.24	26.27
25%	28.98	35.40	25%	25.60	25.44
50%	30.86	34.53	50%	27.29	24.70
75%	31.52	31.92	75%	27.82	23.23
100%	31.33	27.07	100%	28.94	20.32
De-to-Fr Translation			**Fr-to-De Translation**		
$\alpha\%$	T1 (de, fr*)	T2 (de*, fr)	$\alpha\%$	T1 (fr, de*)	T2 (fr*, de)
0%	24.37	36.44	0%	18.85	22.62
25%	28.60	36.21	25%	24.30	22.88
50%	28.87	34.06	50%	25.91	22.10
75%	30.11	32.42	75%	27.41	20.94
100%	30.45	27.65	100%	27.79	18.68
En-to-Fr Translation			**Fr-to-En Translation**		
$\alpha\%$	T1 (en, fr*)	T2 (en*, fr)	$\alpha\%$	T1 (fr, en*)	T2 (fr*, en)
0%	31.74	38.09	0%	31.91	40.74
25%	36.64	37.84	25%	35.94	38.69
50%	38.00	36.83	50%	37.36	37.51
75%	39.00	36.10	75%	39.11	36.61
100%	39.74	33.88	100%	40.27	33.01
En-to-Es Translation			**Es-to-En Translation**		
$\alpha\%$	T1 (en, es*)	T2 (en*, es)	$\alpha\%$	T1 (es, en*)	T2 (es*, en)
0%	31.74	38.09	0%	31.91	40.74
25%	36.64	37.84	25%	35.94	38.69
50%	38.00	36.83	50%	37.36	37.51
75%	39.00	36.10	75%	39.11	36.61
100%	39.74	33.88	100%	40.27	33.01
Es-to-Fr Translation			**Fr-to-Es Translation**		
$\alpha\%$	T1 (es, fr*)	T2 (es*, fr)	$\alpha\%$	T1 (fr, es*)	T2 (fr*, es)
0%	37.32	46.25	0%	39.16	41.60
25%	40.60	46.43	25%	41.81	40.64
50%	41.94	45.57	50%	43.48	39.66
75%	42.39	43.88	75%	45.13	39.03
100%	42.46	40.00	100%	45.42	37.56

Table 2: BLEU scores of all five language pairs on training sets mixed by $\alpha\%$ $X\xrightarrow{\text{H}}Y$ and $(1-\alpha\%)$ $Y\xrightarrow{\text{H}}X$ data, where the mixture rate $\alpha = 0, 25, 50, 75, 100$. We always use T1 to denote the test set aligned with the model direction, and T2 to denote the unaligned one. For readability, we use * to denote the translationese language. For example, "(de, en*)" means original German and translated English pairs.

We report the experiment results of how intervening the train-test alignment affects the MT performance in BLEU scores (Papineni et al., 2002) in Table 2. The main takeaways are as follows:

(1) It is not always the case that, for the same model, the unaligned test set T2 yields higher/more inflated results than the aligned test set T1. When the training data has 75–100% aligned training samples, performance reported on T2 is, in most cases, no longer larger than that on the other half. With

[4] A side benefit of controlling the training data size is that our experiments can help answer what the best nature (i.e., human translation direction) of the training data given a fixed annotation or computation budget is. We leave the space for future work to increase the total training set size with all available training data in both directions.

such training data, usually, T1 inflates the BLEU score more.

(2) The train-test alignment can have an overshadowing effect over the artifacts introduced by the translationese-to-original test set, since no matter which test set we use, the more train-test alignment, the higher the performance reported on T1 than T2 is. Specifically, as we vary the portion of the aligned training data from 0 to 100%, the performance on T1 keeps increasing, the performance on T2 keeps decreasing. Additionally, if the training data is an equal mix or has about 0–50% samples aligned with the model translation direction, then, in many cases, T2 is higher than T1, which might explain the previous observations that T2 inflates the BLEU score (Toral et al., 2018; Graham et al., 2020). To account for another possible interpretation such as the domain shift between the training and test sets, we also conduct an additional evaluation using the newstest2014 test sets, which do not share any domain similarity with our training sets, but still support our observation (in Appendix Table 5).

Hence, the two constructive suggestions for future work is to (1) still report on both test sets, and also the training data direction if available, and (2) if the model will be evaluated only on one type of the test sets, then try to train on as many training data in the same direction as possible.

We should use monolingual data in the original language of the test set. With the intuition that the train-test alignment is a crucial factor for MT performance, we also look into its implications on semi-supervised learning.

Given additional monolingual data, a common question in MT is what type of monolingual data to use, and the accompanying question, whether to use self-training (ST) for the source language monolingual corpus (He et al., 2019; Yarowsky, 1995) or back-translation (BT) for the target language monolingual corpus (Bojar and Tamchyna, 2011; Sennrich et al., 2016; Poncelas et al., 2018). We reframe the question as "*with unlimited monolingual data from both languages, but limited computation resources, which data (together with the corresponding semi-supervised learning method) should we choose?*"

In previous work, BT is the most widely used technique (Bojar et al., 2018; Edunov et al., 2018; Ng et al., 2019; Barrault et al., 2019, p. 15), and is reported to outperform ST (Burlot and Yvon, 2018).

English-to-French (en-to-fr) Translation		
	Test 1 (en, fr*)	Test 2 (en*, fr)
Sup. on Equal Mix	16.16	16.65
+ ST (en, fr**)	**+2.04 (Aligned)**	+1.74
+ BT (en**, fr)	+1.91	**+2.45 (Aligned)**
French-to-English (fr-to-en) Translation		
	Test 1 (fr, en*)	Test 2 (fr*, en)
Sup. on Equal Mix	18.39	15.09
+ ST (fr, en**)	**+2.64 (Aligned)**	+2.24
+ BT (fr**, en)	+2.17	**+3.26 (Aligned)**
English-to-German (en-to-de) Translation		
	Test 1 (en, de*)	Test 2 (en*, de)
Sup. on Equal Mix	10.59	8.80
+ ST (en, de**)	**+1.92 (Aligned)**	+1.60
+ BT (en**, de)	+1.86	**+2.25 (Aligned)**
German-to-English (de-to-en) Translation		
	Test 1 (de, en*)	Test 2 (de*, en)
Sup. on Equal Mix	11.99	13.46
+ ST (de, en**)	**+2.28 (Aligned)**	+1.25
+ BT (de**, en)	+1.99	**+3.72 (Aligned)**

Table 3: Performance on the en-fr and en-de test sets of *newstest2014*. There are two test sets for each task, where * marks the translated language. We use an equal mixture of supervised data in two human translation directions ("*Sup. on Equal Mix*"). Both ST and BT generate pseudo-parallel data (marked by **), with which we find that **aligned** directions between the test set and the pseudo-parallel data lead to larger performance gain.

Another line of previous work inspects the performance gain by BT. Some argue that BT is helpful mostly on the test set aligned with the model (Toral et al., 2018; Freitag et al., 2019; Edunov et al., 2020, Appendix A Table 7) but not the unaligned test set, while others show that BT improves performance on both test sets (Edunov et al., 2020).

We re-inspect the two previous lines of work, and find (1) BT does *not always* outperform ST, especially when ST can make use of the monolingual data in the original language of the test set (to produce pseudo-aligned training data), and (2) the performance gain by BT is *not always* larger on the unaligned test set, but depends on the model direction, especially when BT generates pseudo-aligned training data with the test set.

We implement BT by Edunov et al. (2020), and ST by He et al. (2019). To fairly compare the performance of ST vs. BT, for each language pair X and Y, we split half both training corpora into $X\xrightarrow{\text{H}}Y$-Half1, $X\xrightarrow{\text{H}}Y$-Half2, $Y\xrightarrow{\text{H}}X$-Half1, and $Y\xrightarrow{\text{H}}X$-Half2. We construct the supervised training data as an equal mix (i.e., α=50) combining $X\xrightarrow{\text{H}}Y$-Half1 and $Y\xrightarrow{\text{H}}X$-Half1. The development data is the combination of both development sets, which is also an equal mix.

To train ST or BT, we use the second halves of the training data only as the monolingual corpora. For example, if the translation task is English-to-German translation, ST generates a pseudo-parallel corpus with original English paired with machine-translated pseudo-German, which we denote as (en, de**). For readability, we mark the machine-translation direction with ST and BT by ** and the human translation direction by *.

Our hypothesis is that the machine-translated text pairs (en, de**) will also show similar properties as the human-translated training data (en, de*). Specifically, the more the pseudo-training data is aligned with the test set, the higher performance the semi-supervised learning method will achieve. This is confirmed by the experiment results in Table 3, where, across all settings, no matter which semi-supervised learning method is used, when the pseudo-training data has the same translation direction as the test set, the resulting performance is generally higher. The experiments conducted on CAUSALMT test sets also generally show the same trend, and, due to the space limit, we include the results in the Appendix Table 6.

4 Causal Effect of Data-Model Alignment

The second contribution of this work is to inspect how much another factor, the data-model alignment, causally affects the MT performance. Formally, our research question is that, for a given translation task X-to-Y, considering an equal mix of the test set, does the human translation direction of the training data still matter? If so, how large is the effect, and is it language-/task-dependent?

In this section, we will use causal inference to isolate the effect of data-model alignment from other possible confounders and discuss its effect in different languages and translation tasks.

Our previous experiments show that data-model alignment correlates with MT performance. Our first step is to verify whether data-model alignment is a cause for MT performance. One motivation is that in our previous experiment results in Table 2, for each translation task, there is a clear difference between the causal learning and anticausal learning model. We present the difference in the correlation ("Corr") column of Table 4, referring to the fact that this observation is about how the data-model alignment *correlates* with MT performance on the given CAUSALMT dataset.

We denote this correlation as $P(\mathrm{perf}|\mathrm{aligned})$ between the performance perf and the data-model alignment $\mathrm{aligned}$, which is distinct from the causal relationship $P(\mathrm{perf}|\mathrm{do}(\mathrm{aligned}))$ of how the performance will change when intervening on the data-model alignment, where the do-operator formulates the intervention on a variable by do-calculus (Pearl, 1995) in causal inference.

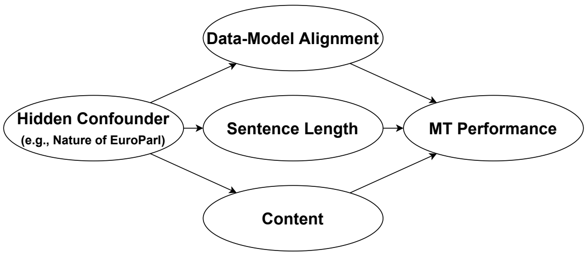


Figure 2: Causal graph about how the data-model alignment affects MT performance.

Formulating the causal effect. Since the true causal effect we want is $P(\mathrm{perf}|\mathrm{do}(\mathrm{aligned}))$ instead of just the correlation, we first need to consider what might interfere with the relationship between data-model alignment and MT performance. The main additional factors we need to control for are shown in the causal graph in Figure 2. We make the assumption that it is very likely that the two corpora of different human translation directions also vary by sentence lengths and the distribution of content (Bogoychev and Sennrich, 2019) due to a hidden confounder (i.e., a common cause) such as the nature of EuroParl. Note that since our research question is about which training data to use given a translation task, the data-model alignment is equivalent to the human translation direction of the training data, as the model translation direction is fixed.

We aim to estimate the causal effect of the data-model alignment (i.e., causal vs. anticausal learning) $\mathrm{aligned}$ on the translation performance perf, while adjusting for other important factors others (sentence lengths and topics).[5] We formulate the average treatment effect (ATE) as follows:

$$\begin{aligned}\mathrm{ATE} = P(&\mathrm{perf}|\mathrm{do}(\mathrm{aligned}=1)) \\ &- P(\mathrm{perf}|\mathrm{do}(\mathrm{aligned}=0))\ ,\end{aligned} \tag{1}$$

where the operator $\mathrm{do}(\mathrm{aligned} = 0 \text{ or } 1)$ means to intervene on the data-model alignment to be 0

[5] Note that there are two notions of causality here, one is the treatment we are interested in, namely the data-model direction alignment, known as causal vs. anticausal learning, and the other is the meta-level causality we are interested in, namely how much the data-model direction alignment (as a binary variable) causally affect the translation performance.

English-to-German (en-to-de) Translation					German-to-English (de-to-en) Translation				
	T1 (en, de*)	T2 (en*, de)	**Cau. – Ant.**	Corr		T1 (de, en*)	T2 (de*, en)	**Cau. – Ant.**	Corr
Cau. (en, de*)	21.88	28.77	**+3.13**	+1.75	Cau. (de, en*)	31.70	28.68	**-1.89**	-2.14
Ant. (en*, de)	25.33	22.19			Ant. (de*, en)	26.35	35.92		
French-to-German (fr-to-de) Translation					**German-to-French (de-to-fr) Translation**				
	T1 (fr, de*)	T2 (fr*, de)	**Cau. – Ant.**	Corr		T1 (de, fr*)	T2 (de*, fr)	**Cau. – Ant.**	Corr
Cau. (fr, de*)	18.36	25.45	**+5.57**	+5.0	Cau. (de, fr*)	32.25	29.98	**-3.58**	-2.71
Ant. (fr*, de)	20.46	17.78			Ant. (de*, fr)	28.07	37.74		
French-to-English (fr-to-en) Translation					**English-to-French (en-to-fr) Translation**				
	T1 (fr, en*)	T2 (fr*, en)	**Cau. – Ant.**	Corr		T1 (en, fr*)	T2 (en*, fr)	**Cau. – Ant.**	Corr
Cau. (fr, en*)	33.77	37.42	**+1.43**	+2.53	Cau. (en, fr*)	42.60	37.34	**-0.14**	-0.65
Ant. (fr*, en)	37.67	32.09			Ant. (en*, fr)	38.05	42.03		
Spanish-to-English (es-to-en) Translation					**English-to-Spanish (en-to-es) Translation**				
	T1 (es, en*)	T2 (es*, en)	**Cau. – Ant.**	Corr		T1 (en, es*)	T2 (en*, es)	**Cau. – Ant.**	Corr
Cau. (es, en*)	37.79	33.64	**+12.25**	+0.63	Cau. (en, es*)	39.04	33.68	**+3.50**	+3.79
Ant. (es*, en)	21.69	25.24			Ant. (en*, es)	30.76	34.96		
French-to-Spanish (fr-to-es) Translation					**Spanish-to-French (es-to-fr) Translation**				
	T1 (fr, es*)	T2 (fr*, es)	**Cau. – Ant.**	Corr		T1 (es, fr*)	T2 (es*, fr)	**Cau. – Ant.**	Corr
Cau. (fr, es*)	37.09	43.40	**+5.84**	+2.22	Cau. (es, fr*)	41.67	41.57	**-2.74**	-1.11
Ant. (fr*, es)	38.45	36.20			Ant. (es*, fr)	39.36	46.62		

Table 4: BLEU scores of causal learning (Cau.) vs. anticausal (Ant.) directions after topic control. We calculate the ATE by taking each model's average performance on T1 and T2, and comparing how much causal models outperform anticausal models. In comparison, we show the correlation (Corr), which is the difference by directly comparing the results of causal and anticausal model without topic control in Table 2.

(i.e., anticausal learning) or 1 (i.e., causal learning). This formulation of ATE is about how much the model performance perf will differ if intervening the data-model alignment to be 0 or 1.

Given the causal graph in Figure 2, the ATE in Eq. (1) can be calculated by conditioning on the set of variables others which blocks the backdoor paths (Pearl, 1995) between $\mathrm{aligned}$ and perf. (others fits the backdoor criterion (Pearl, 1993) in that the sentence lengths and content block all non-directed paths from $\mathrm{aligned}$ to perf, and neither is a descendant of any node on the directed path from $\mathrm{aligned}$ to perf.) An intuitive interpretation can be that when we directly look at the correlation between the data-model alignment and MT performance, it might also be due to that different corpora have different distributions of sentence lengths and content. Therefore, we need to control the sentence lengths and content so that the performance difference will be solely due to the data-model alignment.

Formally, the ATE using the do-notation can be calculated by conditioning on the others. Specifically, we integrate over the distribution of $P(\mathrm{others})$, and calculate the difference in the conditional probability distribution $P(\mathrm{perf}|\mathrm{aligned} = 1, \mathrm{others} = Z) - P(\mathrm{perf}|\mathrm{aligned} = 0, \mathrm{others} = Z)$ of perf given the data-model alignment value $\mathrm{aligned}$ conditioned on the other key variables others for each of its possible value Z, as shown in Eq. (2):

$$\mathrm{ATE} = \int_Z [(P(\mathrm{perf}|\mathrm{aligned} = 1, \mathrm{others} = Z) - P(\mathrm{perf}|\mathrm{aligned} = 0, \mathrm{others} = Z))P(Z)] \tag{2}$$

$$= \mathbb{E}_Z[\mathrm{perf}|\mathrm{aligned} = 1, \mathrm{others} = Z] - \mathbb{E}_Z[\mathrm{perf}|\mathrm{aligned} = 0, \mathrm{others} = Z]\,. \tag{3}$$

Finally, we estimate it by comparing the expected values of the model performance perf given $\mathrm{aligned} = 0$ or 1 over all possible values of others, as shown in Eq. (3).

Causal effect estimation by matching. To estimate the ATE in Eq. (3), the intuition is that we need to take care of the covariates in others so that the aligned setting and the unaligned setting are comparable. We follow the covariate matching method in causal inference (Rosenbaum and Rubin, 1983; Iacus et al., 2012) and the adjustment in the high-dimensional setting of text (Roberts et al., 2020; Veitch et al., 2020). Specifically, matching is a method in causal inference to subsample the treated (i.e., the aligned corpus with the model direction) and control samples (i.e., the unaligned corpus with the model direction) so that the covariates of interest are matched.

We aim to match subsets of the causal and anticausal datasets so that the two sets have similar distributions of sentence lengths and content. In our implementation, we match pairs of samples, one from the causal corpus and the other from the anticausal corpus, where we constrain them to share similar contents and similar sentence lengths. We include our experimental details of the matching process, and quality check of the matched distributions in Appendix E.1.

Based on the matched datasets that control for the sentence lengths and contents, we calculate ATE as the differences of MT performance of models trained on the two directions of the new datasets.

Causal effect results. As shown in the results in Table 4, we can have three observations: (1) The data-model alignment is a clear cause for MT performance. The causal effect (ATE) of data-model alignment on MT performance can be up to 12.25 BLEU scores, for example, in the Spanish-to-English translation task. (2) The ATE varies by language and translation tasks. For the English-Spanish language pair, both translation directions get higher BLEU scores if the models are trained in the causal learning direction. For other language pairs, the data-model alignment can sometimes have a distinct positive impact and can also sometimes have a negative impact. (3) The results of correlation (Corr) analysis are, in most cases, smaller than that of the causal analysis by ATE. This indicates that the correlation analysis neglects other important factors such as the sentence length and content, which might also be reflected in the overall correlation. The causal analysis is a more appropriate method to isolate the influence of the data-model alignment.

5 Limitations and Future Work

We list the limitations of the study and corresponding future work directions: (1) The current study mainly looks into clear cases of causal or anticausal learning, but there can potentially be a third case where both languages are translated from a third language, as pointed out in Riley et al. (2020, Figure 1), which is worth exploring for future work. (2) Due to financial budgets, we did not use human evaluation in addition to the BLEU scores, which is reported to be more reflective of the real translation quality (Edunov et al., 2020). We could also potentially add perplexity scores, although that could vary language to language and also not necessarily fair across the original and translationese language. (3) The experiments on the train-test alignment could be extended since real-world MT systems do not need to be limited to trade-offs between training data in two directions, so there could be future work exploring what the best way to make use of the unaligned training data is.

6 Related Work

Linguistic studies have long observed the distinct properties of translationese from text originally authored in the same language (Toury, 1980; Gellerstam, 1986; Baker, 1993; Toury, 1995). Recent work in MT identifies that the translationese-to-original portion of the test sets (i.e., test sets unaligned with the model direction) being statistically significantly easier (Graham et al., 2020), echoing with many previous observations (Toral et al., 2018; Lembersky et al., 2012; Läubli et al., 2018) and thus some suggest to exclude this portion from future test sets (Toral et al., 2018; Zhang and Toral, 2019; Graham et al., 2020; Barrault et al., 2019).

Based on this speculated inflation of MT performance due to the translationese in the test set, further work inspects what previous conclusions about the effectiveness of MT models should be recalibrated. Some discover that models with BT mostly improve on the inflated test set but not the other more challenging portion (Toral et al., 2018; Freitag et al., 2019; Edunov et al., 2020, Appendix A Table 7) and raises concerns that BT is not as effective as expected. Others argue that BT can still improve on both test sets (Edunov et al., 2020).

Our work differs from all previous work in that we bring in two new important factors when considering how translationese affects MT performance, namely the train-test alignment, and data-model alignment. Moreover, beyond the correlation-based analysis in previous papers (Graham et al., 2020; Edunov et al., 2020), we conduct causal inference (Pearl, 2009; Peters et al., 2017) to contribute causal insights on how translationese affects MT.

7 Conclusion

In conclusion, this work proposed two critical factors for MT performance the train-test alignment and data-model alignment. With strict controls for other confounders, we estimated the causal effect size of each factor on MT performance, and provided suggestions for future study in MT, such as using more training data in the aligned direction and paying attention to whether the nature of the translation task is causal or anticausal.

Ethical Considerations

This research mainly focuses on translation using the EuroParl (Koehn, 2005) corpus, which is widely adopted in the community. There is no data privacy issues or bias against certain demographics with regard to this dataset. The potential use of this study is to improve future MT practice in terms of both evaluation and training. Most conclusions in this study are language-agnostic and potentially help MT in all language pairs, although due to the limitation of available data, the study mainly uses the relatively rich-resource languages, English, German, French, and Spanish. There is a possibility that the findings of the study will need to be further adjusted for low-resource or languages with a very different nature than the studied ones, which we strongly encourage future work to explore.

References

Mona Baker. 1993. Corpus linguistics and translation studies: Implications and applications. In *Text and Technology*. John Benjamins.

Mona Baker. 1995. Corpora in translation studies: An overview and some suggestions for future research. *Target. International Journal of Translation Studies*, 7(2):223–243.

Mona Baker. 1996. Corpus-based translation studies: The challenges that lie ahead. *Benjamins Translation Library*, 18:175–186.

Marco Baroni and Silvia Bernardini. 2006. A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, 21(3):259–274.

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 *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2001. Latent dirichlet allocation. In *Advances in Neural Information Processing Systems 14 [Neural Information Processing Systems: Natural and Synthetic, NIPS 2001, December 3-8, 2001, Vancouver, British Columbia, Canada]*, pages 601–608. MIT Press.

Nikolay Bogoychev and Rico Sennrich. 2019. Domain, translationese and noise in synthetic data for neural machine translation. *CoRR*, abs/1911.03362.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Ondřej Bojar and Aleš Tamchyna. 2011. Improving translation model by monolingual data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 330–336, Edinburgh, Scotland. Association for Computational Linguistics.

Franck Burlot and François Yvon. 2018. Using monolingual data in neural machine translation: a systematic study. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 144–155, Brussels, Belgium. Association for Computational Linguistics.

Efim A Dinic. 1970. Algorithm for solution of a problem of maximum flow in networks with power estimation. In *Soviet Math. Doklady*, volume 11, pages 1277–1280.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Marc'Aurelio Ranzato, and Michael Auli. 2020. On the evaluation of machine translation systems trained with back-translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2836–2846, Online. Association for Computational Linguistics.

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 (Volume 1: Research Papers)*, pages 34–44, Florence, Italy. Association for Computational Linguistics.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. Bleu might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71.

Martin Gellerstam. 1986. Translationese in swedish novels translated from english. *Translation studies in Scandinavia*, 1:88–95.

Clark Glymour, Kun Zhang, and Peter Spirtes. 2019. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10:524.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2020. Statistical power and translationese in machine translation evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 72–81, Online. Association for Computational Linguistics.

Junxian He, Jiatao Gu, Jiajun Shen, and Marc'Aurelio Ranzato. 2019. Revisiting self-training for neural sequence generation. *CoRR*, abs/1909.13788.

David Heckerman, Christopher Meek, and Gregory Cooper. 1999. A bayesian approach to causal discovery. *Computation, causation, and discovery*, 19:141–166.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Biwei Huang, Kun Zhang, Yizhu Lin, Bernhard Schölkopf, and Clark Glymour. 2018. Generalized score functions for causal discovery. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, pages 1551–1560. ACM.

Stefano M Iacus, Gary King, and Giuseppe Porro. 2012. Causal inference without balance checking: Coarsened exact matching. *Political analysis*, 20(1):1–24.

Zhijing Jin, Di Jin, Jonas Mueller, Nicholas Matthews, and Enrico Santus. 2019. IMaT: Unsupervised text attribute transfer via iterative matching and translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3097–3109, Hong Kong, China. Association for Computational Linguistics.

Zhijing Jin, Julius von Kügelgen, Jingwei Ni, Tejas Vaidhya, Ayush Kaushal, Mrinmaya Sachan, and Bernhard Schoelkopf. 2021. Causal direction of data collection matters: Implications of causal and anticausal learning for NLP. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9499–9513, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Vassilis Kolias, Ioannis Anagnostopoulos, and Eleftherios Kayafas. 2014. Exploratory analysis of a terabyte scale web corpus. *CoRR*, abs/1409.5443.

David Kurokawa, Cyril Goutte, and Pierre Isabelle. 2009. Automatic detection of translated text and its impact on machine translation. In *Proceedings of Machine Translation Summit XII: Papers*, Ottawa, Canada.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? A case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Sara Laviosa-Braithwaite. 1998. Universals of translation. *Routledge encyclopedia of translation studies. London: Routledge*, pages 288–291.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. 2012. Language models for machine translation: Original vs. translated texts. *Computational Linguistics*, 38(4):799–825.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Judea Pearl. 1993. [bayesian analysis in expert systems]: Comment: Graphical models, causality and intervention. *Statistical Science*, 8(3):266–269.

Judea Pearl. 1995. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688.

Judea Pearl. 2009. *Causality*. Cambridge university press.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of causal inference: Foundations and learning algorithms*. The MIT Press.

Alberto Poncelas, Dimitar Sht. Shterionov, Andy Way, Gideon Maillette de Buy Wenniger, and Peyman Passban. 2018. Investigating backtranslation in neural machine translation. *CoRR*, abs/1804.06189.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Ella Rabinovich and Shuly Wintner. 2015. Unsupervised identification of translationese. *Transactions of the Association for Computational Linguistics*, 3:419–432.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Parker Riley, Isaac Caswell, Markus Freitag, and David Grangier. 2020. Translationese as a language in "multilingual" NMT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7737–7746, Online. Association for Computational Linguistics.

Margaret E Roberts, Brandon M Stewart, and Richard A Nielsen. 2020. Adjusting for confounding with text matching. *American Journal of Political Science*, 64(4):887–903.

Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris M. Mooij. 2012. On causal and anticausal learning. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*. icml.cc / Omnipress.

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

Pater Spirtes, Clark Glymour, and Richard Scheines. 2000a. Constructing bayesian network models of gene expression networks from microarray data. In *Proceedings of the Atlantic Symposium on Computational Biology*.

Peter Spirtes, Clark Glymour, and Richard Scheines. 2000b. *Causation, Prediction, and Search, Second Edition*. Adaptive computation and machine learning. MIT Press.

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, Brussels, Belgium. Association for Computational Linguistics.

Gideon Toury. 1980. *In search of a theory of translation*. Porter Institute for Poetics and Semiotics, Tel Aviv University.

Gideon Toury. 1995. *Descriptive translation studies and beyond*, volume 4. John Benjamins.

Michael Ustaszewski. 2019. Optimising the Europarl corpus for translation studies with the EuroparlExtract toolkit. *Perspectives*, 27(1):107–123.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. 2021. Counterfactual invariance to spurious correlations in text classification. *Advances in Neural Information Processing Systems*, 34.

Victor Veitch, Dhanya Sridhar, and David M. Blei. 2020. Adapting text embeddings for causal inference. In *Proceedings of the Thirty-Sixth Conference on Uncertainty in Artificial Intelligence, UAI 2020, virtual online, August 3-6, 2020*, volume 124 of *Proceedings of Machine Learning Research*, pages 919–928. AUAI Press.

Vered Volansky, Noam Ordan, and Shuly Wintner. 2015. On the features of translationese. *Digit. Scholarsh. Humanit.*, 30(1):98–118.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd Annual Meeting of the Association for Computational Linguistics*, pages 189–196, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

Mike Zhang and Antonio Toral. 2019. The effect of translationese in machine translation test sets. In *Proceedings of the Fourth Conference on Machine*

Translation (Volume 1: Research Papers), pages 73–81, Florence, Italy. Association for Computational Linguistics.

A Reproducibility, License, and Copyright

We open-source our codes and datasets, which are both uploaded to the submission system. In our data, we include all three variations: the full CausalMT dataset, the split used for the semi-supervised learning experiments, and the subset after matching the contents and sentence lengths. In our codes, we include all commands with hyperparameters to help future work to reproduce our results.

The codes and data are under MIT license. Note that the EuroParl dataset has no copyright restriction, according to its official website.[6]

B Linguistic Property Analysis

We also open-source the codes to calculate the linguistic properties of our dataset in Table 1. We use the Python library Stanza[7] (Qi et al., 2020) to tokenize the sentences when calculating the number of sentences per sample. For speed concerns, we use NLTK[8] (Bird et al., 2009) to tokenize the words and count the vocabulary. We use the Python library spaCy[9] (Honnibal and Montani, 2017) to calculate the passive voice ratio and punctuation per sample.

C Implementation Details

C.1 Preprocessing

To prepare the text for the models, we follow the preprocessing scripts of fairseq (Ott et al., 2019).[10] Specifically, we use the Moses tokenizer (Koehn et al., 2007),[11] the default byte pair encoding (BPE) size of 40K subwords, and remove sentence pairs that of larger than 1.5 length ratio from the training set.

C.2 Evaluation Script

We use the fairseq-generate script[12] to calculate the BLEU score (Papineni et al., 2002) of each translation model, with beam width of 5, BPE removed, detoknized by moses.

C.3 Model Details

We use the sequence-to-sequence Transformer model (Vaswani et al., 2017) implemented by the fairseq library (Ott et al., 2019). Specifically, we use a six-layer Transformer, a label smoothing of 0.1, a weight decay of 0.0001, a dropout of 0.3, 4000 warming updates, and a learning rate of 0.0005. All results are reported by a single run but a fixed random seed.

For the semi-supervised learning, we implement the BT model following Edunov et al. (2020) to use the Facebook-FAIR system of the WMT'19 news shared translation task.[13] All the hyperparameters are the same as the supervised system, with a learning rate of 0.0007 on both the supervised training data and the generated pseudo-parallel corpus. We implement the ST model by He et al. (2019) following their script,[14] and also keep the hyperparameters the same as the supervised model.

C.4 Training Details

We train the supervised learning model and each step in the semi-supervised learning scripts for 1000 epochs. We select the model with the best performance on the development set and report the final evaluation results on the test set.

All experiments are run on NVIDIA RTX2080 GPUs. Each supervised learning experiment takes around 32 GPU hours, and each semi-supervised learning experiment takes about 128 GPU hours.

D Additional Experimental Results

D.1 Effect of Train-Test Alignment on Supervised Learning

To inspect the influence of train-test alignment on the MT performance, we conduct all experiments on our CausalMT test sets and also the standard newstest2014 test sets. For the supervised learning performance, we list the performance on the CausalMT test sets in the main paper in Table 2, and list the additional performance on the newstest2014 test sets in Table 5.

For better visualization of the trends, we also provide line plots of the same experimental results in Table 2. Specifically, we plot the results of German-English translation in Figure 3a using our previous

[6] https://www.statmt.org/europarl/

[7] https://stanfordnlp.github.io/stanza/

[8] https://www.nltk.org/

[9] https://spacy.io/

[10] https://github.com/pytorch/fairseq/

[11] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/tokenizer.perl

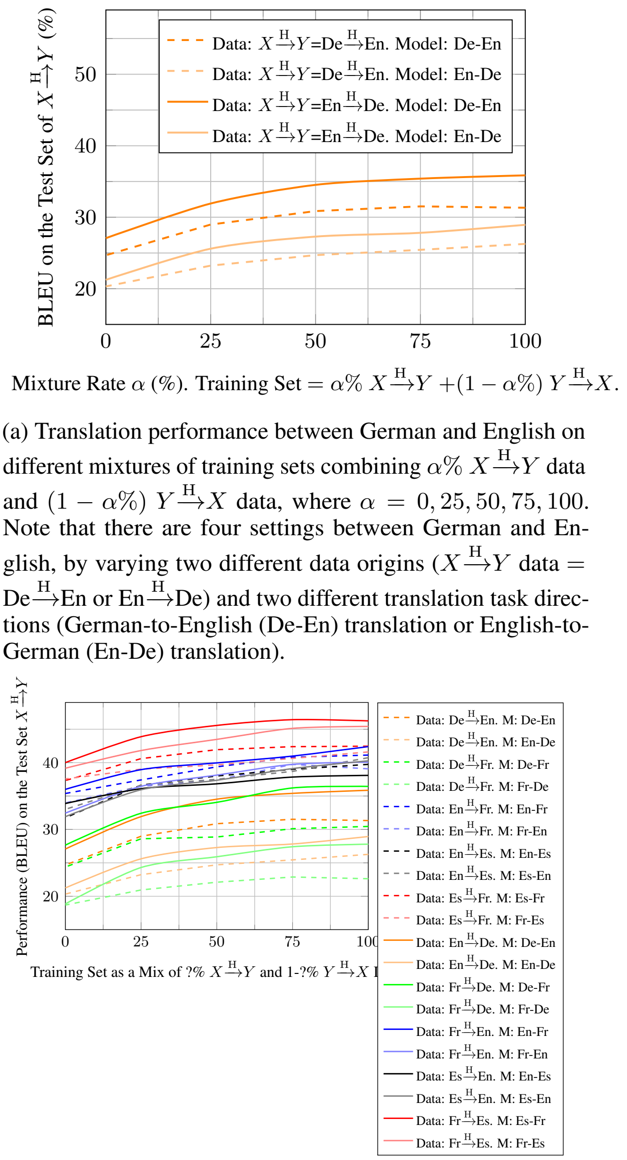
[12] https://github.com/pytorch/fairseq/blob/main/fairseq_cli/generate.py

[13] https://github.com/pytorch/fairseq/tree/main/examples/backtranslation

[14] https://github.com/jxhe/self-training-text-generation/blob/master/self_train.sh

de-to-en Translation			**en-to-de Translation**		
α%	T1 (de, en*)	T2 (de*, en)	α%	T1 (en, de*)	T2 (en*, de)
0%	14.21	19.10	0%	11.18	15.49
25%	15.71	18.69	25%	12.69	14.29
50%	16.77	18.17	50%	13.30	14.33
75%	16.91	16.27	75%	13.38	13.16
100%	16.02	12.91	100%	13.28	10.68
en-to-fr Translation			**fr-to-en Translation**		
α%	T1 (en, fr*)	T2 (en*, fr)	α%	T1 (fr, en*)	T2 (fr*, en)
0%	16.61	21.33	0%	16.34	23.26
25%	18.56	20.95	25%	18.81	23.31
50%	20.45	21.66	50%	19.75	23.20
75%	21.19	21.05	75%	21.09	22.01
100%	21.43	19.30	100%	20.02	19.78

Table 5: Effect of train-test alignment on the en-fr and en-de test sets of *newstest2014*.



(a) Translation performance between German and English on different mixtures of training sets combining $\alpha\%\ X \xrightarrow{H} Y$ data and $(1-\alpha\%)\ Y \xrightarrow{H} X$ data, where $\alpha = 0, 25, 50, 75, 100$. Note that there are four settings between German and English, by varying two different data origins ($X \xrightarrow{H} Y$ data = De$\xrightarrow{H}$En or En$\xrightarrow{H}$De) and two different translation task directions (German-to-English (De-En) translation or English-to-German (En-De) translation).

Performance (BLEU) on the Test Set $X \xrightarrow{H} Y$
20
30
40
0
25
50
75
100
Training Set as a Mix of ?% $X \xrightarrow{H} Y$ and 1-?% $Y \xrightarrow{H} X$ I
Data: De$\xrightarrow{H}$En. M: De-En
Data: De$\xrightarrow{H}$En. M: En-De
Data: De$\xrightarrow{H}$Fr. M: De-Fr
Data: De$\xrightarrow{H}$Fr. M: Fr-De
Data: En$\xrightarrow{H}$Fr. M: En-Fr
Data: En$\xrightarrow{H}$Fr. M: Fr-En
Data: En$\xrightarrow{H}$Es. M: En-Es
Data: En$\xrightarrow{H}$Es. M: Es-En
Data: Es$\xrightarrow{H}$Fr. M: Es-Fr
Data: Es$\xrightarrow{H}$Fr. M: Fr-Es
Data: En$\xrightarrow{H}$De. M: De-En
Data: En$\xrightarrow{H}$De. M: En-De
Data: Fr$\xrightarrow{H}$De. M: De-Fr
Data: Fr$\xrightarrow{H}$De. M: Fr-De
Data: Fr$\xrightarrow{H}$En. M: En-Fr
Data: Fr$\xrightarrow{H}$En. M: Fr-En
Data: Es$\xrightarrow{H}$En. M: En-Es
Data: Es$\xrightarrow{H}$En. M: Es-En
Data: Fr$\xrightarrow{H}$Es. M: Es-Fr
Data: Fr$\xrightarrow{H}$Es. M: Fr-Es

(b) Translation performance between all five language pairs on different mixtures of training sets combining $\alpha\%\ X \xrightarrow{H} Y$ data and $(1-\alpha\%)\ Y \xrightarrow{H} X$ data, where the mixture rate $\alpha = 0, 25, 50, 75, 100$.

experiment results in Table 2. We also include the diagram of all five language pairs in Figure 3b.

In Figure 3a, we use lines with the same darkness of color for the same model trained on different data directions. Results show that the data-model alignment matter significantly. Taking the German-to-English translation models (- - - and —), the two data directions can cause up to 4.53 difference in BLEU scores. In the current figures, we also see that the data direction with a smaller expansion factor is a better training corpus than the other one.

We use the same line type (dashed or solid) for models trained on the same data. Using the same data, the performance of the two different directions of models cannot be compared directly because the target language is different, causing the BLEU calculation to be different.

D.2 Effect of Train-Test Alignment on Semi-Supervised Learning

For the semi-supervised learning performance, we show the performance on the newstest2014 test sets in Table 3 in the main paper, and performance on the test sets of CAUSALMT in Table 6. Note that the decrease of ST performance on En-Es and Es-Fr pairs is possible because ST is more sensitive to the quality of the model learned on the supervised data, and these language pairs have a smaller training data size of 90K compared with 200K+ data for all the other language pairs.

E Implementation Details for Causal Inference

E.1 Matching for Causal effect estimation

Implementation Details of Matching For each sentence in the aligned corpus, we select its most similar match in the unaligned corpus. We want pairs of samples with similar content and sentence lengths. Empirically, we limit the sentence length ratio of each matched pair to be no larger than 1.1 and the content to have a cosine similarity larger than 0.7, following the threshold to match a content-similar pseudo-parallel corpus in Jin et al. (2019). To perform the matching, we use Dinic's maximal matching algorithm (Dinic, 1970).

To calculate the content-wise similarity of a pair of samples, we represent each sentence by the sentence BERT embedding (Reimers and Gurevych, 2019). In case of multiple languages as candidates to match the sentence embeddings in, we set a pri-

German-to-English (de-to-en) Translation		
	Test 1 (de, en*)	Test 2 (de*, en)
Sup. on Equal Mix	25.52	29.02
+ ST (de, en**)	**+3.25 (Aligned)**	+2.59
+ BT (de**, en)	+0.97	**+2.67 (Aligned)**
English-to-German (en-to-de) Translation		
	Test 1 (en, de*)	Test 2 (en*, de)
Sup. on Equal Mix	23.76	21.48
+ ST (en, de**)	**+1.39 (Aligned)**	+1.44
+ BT (en**, de)	-0.63	+0.51 (Aligned)
German-to-French (de-to-fr) Translation		
	Test 1 (de, fr*)	Test 2 (de*, fr)
Sup. on Equal Mix	25.42	30.10
+ ST (de, fr**)	**+1.79 (Aligned)**	+1.23
+ BT (de**, fr)	+0.35	**+1.69 (Aligned)**
French-to-German (fr-to-de) Translation		
	Test 1 (fr, de*)	Test 2 (fr*, de)
Sup. on Equal Mix	21.89	18.60
+ ST (fr, de**)	**+2.46 (Aligned)**	+2.09
+ BT (fr**, de)	+1.07	+0.87 (Aligned)
English-to-French (en-to-fr) Translation		
	Test 1 (en, fr*)	Test 2 (en*, fr)
Sup. on Equal Mix	35.64	36.19
+ ST (en, fr**)	**+2.04 (Aligned)**	+1.89
+ BT (en**, fr)	+0.13	+1.33 (Aligned)
French-to-Englih (fr-to-en) Translation		
	Test 1 (fr, en*)	Test 2 (fr*, en)
Sup. on Equal Mix	34.35	33.75
+ ST (fr, en**)	**+1.76 (Aligned)**	+2.28
+ BT (fr**, en)	+0.43	+1.89 (Aligned)
English-to-Spanish (en-to-es) Translation		
	Test 1 (en, es*)	Test 2 (en*, es)
Sup. on Equal Mix	33.65	34.01
+ ST (en, es**)	-0.10 (Aligned)	-0.75
+ BT (en**, es)	+0.36	**+1.04 (Aligned)**
Spanish-to-English (es-to-en) Translation		
	Test 1 (es, en*)	Test 2 (es*, en)
Sup. on Equal Mix	35.00	33.82
+ ST (es, en**)	-0.46 (Aligned)	-0.41
+ BT (es**, en)	+0.63	**+1.77 (Aligned)**
Spanish-to-French (es-to-fr) Translation		
	Test 1 (es, fr*)	Test 2 (es*, fr)
Sup. on Equal Mix	38.30	40.40
+ ST (es, fr**)	+0.58 (Aligned)	+0.83
+ BT (es**, fr)	+1.00	**+2.01 (Aligned)**
French-to-Spanish (fr-to-es) Translation		
	Test 1 (fr, es*)	Test 2 (fr*, es)
Sup. on Equal Mix	40.61	38.55
+ ST (fr, es**)	-0.84 (Aligned)	-1.09
+ BT (fr**, es)	-0.14	**+0.12 (Aligned)**

Table 6: Performance analogous to Table 3 but on our CAUSALMT test sets.

oritization order of "En>De>Fr>Es" for sentence embedding matching.

Note that since the set of factors to control is in a high-dimensional vector space, it is less realistic to use other common matching methods such as propensity score stratification and matching, as pointed out by Roberts et al. (2020).

Quality Check We check the quality of the matched corpora. First, we list the statistics of the new corpora in Table 8, and analyze its linguistic properties in Table 7.

More importantly, we check whether the covariates are well controlled. Taking the German-English language pair as an example, we plot the distributions of sentence lengths across the De$\xrightarrow{H}$En and En$\xrightarrow{H}$ De corpora in Figure 4 and the distributions of topics after learning an Latent Dirichlet Allocation (LDA) topic model (Blei et al., 2001) in Figure 5. We also list some example matched samples in English in Table 9.

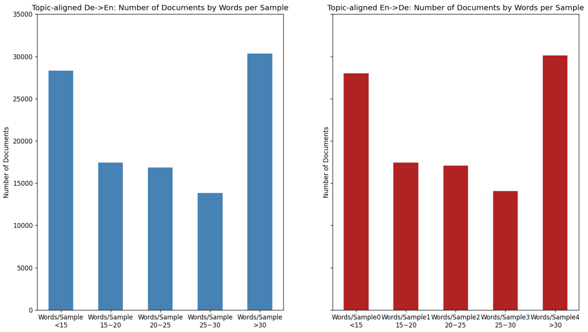


Figure 4: Distribution of sentence lengths after matching, using the German-English language pair as an example.

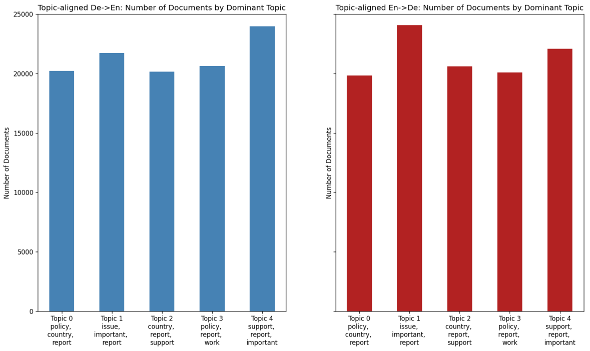


Figure 5: Distribution of different topics after matching, using the German-English language pair as an example.

E.2 Confirming the Causal Graph by Causal Discovery

To check our causal graph assumption, we first verify whether data-model alignment is a cause for MT performance using causal discovery.

We use the causal discovery algorithm, fast causal inference (FCI) (Spirtes et al., 2000a), to verify that the data-model alignment causally affects the translation performance, conditioned on other factors such as the sentence length and topics.

FCI is the most appropriate causal inference method for this analysis since there might exist hidden confounders that affect the MT performance, which normal causal discovery methods such as score-based methods (Heckerman et al., 1999; Huang et al., 2018) and other constraint-

	De$\xrightarrow{H}$En	En$\xrightarrow{H}$De	De$\xrightarrow{H}$Fr	Fr$\xrightarrow{H}$De	En$\xrightarrow{H}$Fr	Fr$\xrightarrow{H}$En	En$\xrightarrow{H}$Es	Es$\xrightarrow{H}$En	Es$\xrightarrow{H}$Fr	Fr$\xrightarrow{H}$Es
# Words/Sample	21.05/23.82	23.82/22.64	23.81/30.09	29.19/24.39	25.40/29.76	28.02/25.33	26.69/28.18	27.14/26.96	30.59/34.33	33.45/30.39
# Sents/Sample	1.032/1.031	1.020/1.041	1.041/1.925	1.902/1.056	1.033/1.949	1.872/1.040	1.028/1.053	1.070/1.057	1.076/2.100	2.088/1.068
Sent Expansion Factor	en:de=1.00	en:de=0.98	fr:de=1.85	fr:de=1.80	fr:en=1.88	fr:en=1.80	es:en=1.02	es:en=1.01	fr:es=1.95	fr:es=1.96
Passive Voice (%)	-/0.1128	0.1036/-	-/-	-/-	0.1073/-	-/0.1185	0.1155/-	-/0.1256	-/-	-/-
# Punctuation/Sample	3.04/2.83	2.63/3.04	3.45/6.04	6.43/3.35	2.82/5.89	6.16/3.11	2.93/2.71	3.07/3.12	3.42/7.02	7.44/3.50
# Syllables/Word	2.002/1.744	1.755/2.059	1.988/1.553	1.546/2.068	1.758/1.562	1.55/1.78	1.760/2.022	2.01/1.78	2.010/1.567	1.544/2.030
Flesch Reading Ease	31.90/35.22	33.78/29.30	35.25/46.30	46.1/28.0	31.93/45.80	49.91/31.09	30.55/50.22	51.94/30.05	48.82/43.26	42.04/46.87
MATTR	58.93/52.68	53.31/60.58	59.32/52.77	52.89/61.74	53.19/52.55	52.32/53.38	53.90/54.91	53.90/52.33	53.60/51.85	52.29/54.84
Lexical Density	49.15/49.24	49.91/50.75	48.86/55.30	55.21/50.82	49.99/55.18	55.16/50.28	50.24/49.76	48.88/49.56	48.72/55.02	55.14/50.01
Vocab Size	58K/22K	23K/56K	78K/37K	39K/71K	22K/31K	31K/21K	19K/31K	29K/16K	32K/26K	29K/34K

Table 7: Detailed characteristics of the matched dataset.

Human Trans. Dir.	**Train**	**Dev**	**Test**
De $\xrightarrow{H}$En	107K	1K	2K
En $\xrightarrow{H}$De	107K	1K	2K
De $\xrightarrow{H}$Fr	133K	1K	2K
Fr $\xrightarrow{H}$De	133K	1K	2K
En $\xrightarrow{H}$Fr	87K	1K	2K
Fr $\xrightarrow{H}$En	87K	1K	2K
En $\xrightarrow{H}$Es	47K	1K	2K
Es $\xrightarrow{H}$En	47K	1K	2K
Es $\xrightarrow{H}$Fr	50K	1K	2K
Fr $\xrightarrow{H}$Es	50K	1K	2K

Table 8: Dataset statistics for five language pairs after matching. Each language pair has data from two human translation directions (Human Trans. Dir.), e.g., De$\xrightarrow{H}$En and En$\xrightarrow{H}$De.

based algorithms like Peter-Clark (PC) algorithm (Spirtes et al., 2000b, §5.4.2, pp. 84–88) cannot handle (Glymour et al., 2019). FCI gives asymptotically correct results in the presence of confounders, and outputs Markov equivalence classes, i.e., a set of causal structures satisfying the same conditional independences.

Given a language pair X and Y, we generate eight sets of experiment results, by varying the two training directions, two test directions, and two model directions. We extract the test samples of all eight experiments, and since each test set is 2K, there are 16K samples in total. On the 16K samples, besides keeping the label of their data-model alignment, translation performance in BLEU, we also calculate the other factors such as the test-model alignment, train-test alignment, source sentence length, and the topic vector by topic modeling on all the training data of the language pair X and Y. We run the FCI algorithm using the causal-learn Python package[15] over all the variables of interest. The implementation details are in the Appendix.

Corpus	**Matched Sample**
En$\xrightarrow{H}$De	However, I have one or two points.
De$\xrightarrow{H}$En	Let me make some comments on specific points.
En$\xrightarrow{H}$De	That greater urgency has been recognised in the Council suggestion that we should have an intergovernmental conference beginning next year, something which we subscribe to.
De$\xrightarrow{H}$En	From our perspective, it is now urgently necessary that the Council also accepts this proposal, so that the negotiations can commence as soon as possible.
En$\xrightarrow{H}$De	I agree that the European Union needs an integrated, coherent and consistent European energy policy that maintains Europe's competitiveness, safeguards our environmental objectives and ensures our security of supply.
De$\xrightarrow{H}$En	We want a European Union that is strong, effective and democratic, and all those who want to make it no more than a free trade zone within Europe will have a fight on their hands.

Table 9: Examples of matched samples between the En$\xrightarrow{H}$De and De$\xrightarrow{H}$En corpora.

The resulting causal graph on the German-English language pair is in Figure 2. The results confirm our hypothesis that the data-model alignment (causal vs. anticausal direction) does have a causal effect on the BLEU score, together with other factors such as the sentence length and topics.

[15] https://github.com/cmu-phil/causal-learn