Translation Artifacts in Cross-lingual Transfer Learning

Mikel Artetxe, Gorka Labaka, Eneko Agirre
HiTZ Center
University of the Basque Country (UPV/EHU)
{mikel.artetxe,gorka.labaka,e.agirre}@ehu.eus

Abstract

Both human and machine translation play a central role in cross-lingual transfer learning: many multilingual datasets have been created through professional translation services, and using machine translation to translate either the test set or the training set is a widely used transfer technique. In this paper, we show that such translation process can introduce subtle artifacts that have a notable impact in existing cross-lingual models. For instance, in natural language inference, translating the premise and the hypothesis independently can reduce the lexical overlap between them, which current models are highly sensitive to. We show that some previous findings in cross-lingual transfer learning need to be reconsidered in the light of this phenomenon. Based on the gained insights, we also improve the state-of-the-art in XNLI for the *translate-test* and *zero-shot* approaches by 4.3 and 2.8 points, respectively.

1 Introduction

While most NLP resources are English-specific, there have been several recent efforts to build **multilingual benchmarks**. One possibility is to collect and annotate data in multiple languages separately (Clark et al., 2020), but most existing datasets have been created through translation (Conneau et al., 2018; Artetxe et al., 2020). This approach has two desirable properties: it relies on existing professional translation services rather than requiring expertise in multiple languages, and it results in parallel evaluation sets that offer a meaningful measure of the cross-lingual transfer gap of different models. The resulting multilingual datasets are generally used for evaluation only, relying on existing English datasets for training.

Closely related to that, **cross-lingual transfer learning** aims to leverage large datasets available in one language—typically English—to build multilingual models that can generalize to other languages. Previous work has explored 3 main approaches to that end: machine translating the test set into English and using a monolingual English model (TRANSLATE-TEST), machine translating the training set into each target language and training the models on their respective languages (TRANSLATE-TRAIN), or using English data to fine-tune a multilingual model that is then transferred to the rest of languages (ZERO-SHOT).

The dataset creation and transfer procedures described above result in a **mixture of original,[1] human translated and machine translated data** when dealing with cross-lingual models. In fact, the type of text a system is trained on does not typically match the type of text it is exposed to at test time: TRANSLATE-TEST systems are trained on original data and evaluated on machine translated test sets, ZERO-SHOT systems are trained on original data and evaluated on human translated test sets, and TRANSLATE-TRAIN systems are trained on machine translated data and evaluated on human translated test sets.

Despite overlooked to date, we show that **such mismatch has a notable impact** in the performance of existing cross-lingual models. By using back-translation (Sennrich et al., 2016) to paraphrase each training instance, we obtain another English version of the training set that better resembles the test set, obtaining substantial improvements for the TRANSLATE-TEST and ZERO-SHOT approaches in cross-lingual Natural Language Inference (NLI). While improvements brought by machine translation have previously been attributed to data augmentation (Singh et al., 2019), we reject this hypothesis and show that the phenomenon is only present in translated test sets, but not in original ones. Instead, our analysis reveals that

[1] We use the term *original* to refer to non-translated text.

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 7674–7684,
November 16–20, 2020.

this behavior is caused by subtle **artifacts arising from the translation** process itself. In particular, we show that translating different parts of each instance separately (e.g., the premise and the hypothesis in NLI) can alter superficial patterns in the data (e.g., the degree of lexical overlap between them), which severely affects the generalization ability of current models. Based on the gained insights, we improve the state-of-the-art in XNLI, and show that some previous findings need to be reconsidered in the light of this phenomenon.

2 Related work

Cross-lingual transfer learning. Current cross-lingual models work by pre-training multilingual representations using some form of language modeling, which are then fine-tuned on the relevant task and transferred to different languages. Some authors leverage parallel data to that end (Conneau and Lample, 2019; Huang et al., 2019), but training a model akin to BERT (Devlin et al., 2019) on the combination of monolingual corpora in multiple languages is also effective (Conneau et al., 2020). Closely related to our work, Singh et al. (2019) showed that replacing segments of the training data with their translation during fine-tuning is helpful. However, they attribute this behavior to a data augmentation effect, which we believe should be reconsidered given the new evidence we provide.

Multilingual benchmarks. Most benchmarks covering a wide set of languages have been created through translation, as it is the case of XNLI (Conneau et al., 2018) for NLI, PAWS-X (Yang et al., 2019) for adversarial paraphrase identification, and XQuAD (Artetxe et al., 2020) and MLQA (Lewis et al., 2020) for Question Answering (QA). A notable exception is TyDi QA (Clark et al., 2020), a contemporaneous QA dataset that was separately annotated in 11 languages. Other cross-lingual datasets leverage existing multilingual resources, as it is the case of MLDoc (Schwenk and Li, 2018) for document classification and Wikiann (Pan et al., 2017) for named entity recognition. Concurrent to our work, Hu et al. (2020) combine some of these datasets into a single multilingual benchmark, and evaluate some well-known methods on it.

Annotation artifacts. Several studies have shown that NLI datasets like SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018) contain spurious patterns that can be exploited to obtain strong results without making real inferential decisions. For instance, Gururangan et al. (2018) and Poliak et al. (2018) showed that a hypothesis-only baseline performs better than chance due to cues on their lexical choice and sentence length. Similarly, McCoy et al. (2019) showed that NLI models tend to predict *entailment* for sentence pairs with a high lexical overlap. Several authors have worked on adversarial datasets to diagnose these issues and provide a more challenging benchmark (Naik et al., 2018; Glockner et al., 2018; Nie et al., 2020). Besides NLI, other tasks like QA have also been found to be susceptible to annotation artifacts (Jia and Liang, 2017; Kaushik and Lipton, 2018). While previous work has focused on the monolingual scenario, we show that translation can interfere with these artifacts in multilingual settings.

Translationese. Translated texts are known to have unique features like simplification, explicitation, normalization and interference, which are refer to as *translationese* (Volansky et al., 2013). This phenomenon has been reported to have a notable impact in machine translation evaluation (Zhang and Toral, 2019; Graham et al., 2019). For instance, back-translation brings large BLEU gains for reversed test sets (i.e., when translationese is on the source side and original text is used as reference), but its effect diminishes in the natural direction (Edunov et al., 2020). While connected, the phenomenon we analyze is different in that it arises from translation inconsistencies due to the lack of context, and affects cross-lingual transfer learning rather than machine translation.

3 Experimental design

Our goal is to analyze the effect of both human and machine translation in cross-lingual models. For that purpose, the core idea of our work is to (i) use machine translation to either translate the training set into other languages, or generate English paraphrases of it through back-translation, and (ii) evaluate the resulting systems on original, human translated and machine translated test sets in comparison with systems trained on original data. We next describe the models used in our experiments (§3.1), the specific training variants explored (§3.2), and the evaluation procedure followed (§3.3).

3.1 Models and transfer methods

We experiment with two models that are representative of the state-of-the-art in monolingual and

cross-lingual pre-training: (i) ROBERTA (Liu et al., 2019), which is an improved version of BERT that uses masked language modeling to pre-train an English Transformer model, and (ii) XLM-R (Conneau et al., 2020), which is a multilingual extension of the former pre-trained on 100 languages. In both cases, we use the large models released by the authors under the fairseq repository.[2] As discussed next, we explore different variants of the training set to fine-tune each model on different tasks. At test time, we try both machine translating the test set into English (TRANSLATE-TEST) and, in the case of XLM-R, using the actual test set in the target language (ZERO-SHOT).

3.2 Training variants

We try 3 variants of each training set to fine-tune our models: (i) the original one in English (ORIG), (ii) an English paraphrase of it generated through back-translation using Spanish or Finnish as pivot (BT-ES and BT-FI), and (iii) a machine translated version in Spanish or Finnish (MT-ES and MT-FI). For sentences occurring multiple times in the training set (e.g., premises repeated for multiple hypotheses), we use the exact same translation for all occurrences, as our goal is to understand the inherent effect of translation rather than its potential application as a data augmentation method.

In order to train the machine translation systems for MT-XX and BT-XX, we use the big Transformer model (Vaswani et al., 2017) with the same settings as Ott et al. (2018) and SentencePiece tokenization (Kudo and Richardson, 2018) with a joint vocabulary of 32k subwords. For English-Spanish, we train for 10 epochs on all parallel data from WMT 2013 (Bojar et al., 2013) and ParaCrawl v5.0 (Esplà et al., 2019). For English-Finnish, we train for 40 epochs on Europarl and Wiki Titles from WMT 2019 (Barrault et al., 2019), ParaCrawl v5.0, and DGT, EUbookshop and TildeMODEL from OPUS (Tiedemann, 2012). In both cases, we remove sentences longer than 250 tokens, with a source/target ratio exceeding 1.5, or for which `langid.py` (Lui and Baldwin, 2012) predicts a different language, resulting in a final corpus size of 48M and 7M sentence pairs, respectively. We use sampling decoding with a temperature of 0.5 for inference, which produces more diverse translations than beam search (Edunov et al., 2018) and performed better in our preliminary experiments.

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

3.3 Tasks and evaluation procedure

We use the following tasks for our experiments:

Natural Language Inference (NLI). Given a premise and a hypothesis, the task is to determine whether there is an *entailment*, *neutral* or *contradiction* relation between them. We fine-tune our models on MultiNLI (Williams et al., 2018) for 10 epochs using the same settings as Liu et al. (2019). In most of our experiments, we evaluate on XNLI (Conneau et al., 2018), which comprises 2490 development and 5010 test instances in 15 languages. These were originally annotated in English, and the resulting premises and hypotheses were independently translated into the rest of the languages by professional translators. For the TRANSLATE-TEST approach, we use the machine translated versions from the authors. Following Conneau et al. (2020), we select the best epoch checkpoint according to the average accuracy in the development set.

Question Answering (QA). Given a context paragraph and a question, the task is to identify the span answering the question in the context. We fine-tune our models on SQuAD v1.1 (Rajpurkar et al., 2016) for 2 epochs using the same settings as Liu et al. (2019), and report test results for the last epoch. We use two datasets for evaluation: XQuAD (Artetxe et al., 2020), a subset of the SQuAD development set translated into 10 other languages, and MLQA (Lewis et al., 2020) a dataset consisting of parallel context paragraphs plus the corresponding questions annotated in English and translated into 6 other languages. In both cases, the translation was done by professional translators at the document level (i.e., when translating a question, the text answering it was also shown). For our BT-XX and MT-XX variants, we translate the context paragraph and the questions independently, and map the answer spans using the same procedure as Carrino et al. (2020).[3] For the TRANSLATE-TEST approach, we use the official machine translated versions of MLQA, run inference over them, and map the predicted answer spans back to the target language.[4]

[3] We use FastAlign (Dyer et al., 2013) for word alignment, and discard the few questions for which the mapping method fails (when none of the tokens in the answer span are aligned).

[4] We use the same procedure as for the training set except that (i) given the small size of the test set, we combine it with WikiMatrix (Schwenk et al., 2019) to aid word alignment, (ii) we use Jieba for Chinese segmentation instead of the Moses tokenizer, and (iii) for the few unaligned spans, we return the English answer.

Model	Train	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	avg
Test set machine translated into English (TRANSLATE-TEST)																	
	ORIG	91.2	82.2	84.6	82.4	82.1	82.1	79.2	76.5	77.4	73.8	73.4	76.7	70.5	67.2	66.8	77.7 ±0.6
ROBERTA	BT-ES	**91.6**	85.7	**87.4**	85.4	85.1	85.1	83.6	81.3	81.5	78.7	78.2	81.1	76.3	72.7	71.5	81.7 ±0.2
	BT-FI	91.4	**86.0**	**87.4**	85.7	**85.7**	85.4	**84.4**	82.3	82.1	79.0	79.3	81.8	77.6	73.5	73.6	82.3 ±0.2
	ORIG	90.3	82.2	84.2	82.6	81.9	82.0	79.3	76.7	77.5	75.0	73.7	77.5	70.9	67.8	67.2	77.9 ±0.3
	BT-ES	90.2	84.1	86.3	84.5	84.5	84.1	82.2	79.6	80.7	78.5	77.3	80.8	75.2	72.5	71.2	80.8 ±0.3
XLM-R	BT-FI	89.5	84.9	85.5	84.5	84.5	84.6	82.9	80.6	81.4	78.9	78.1	81.5	76.3	73.3	72.5	81.3 ±0.2
	MT-ES	89.8	83.2	85.6	84.2	84.0	83.6	81.6	78.4	79.3	77.6	76.7	80.0	74.3	71.3	70.1	80.0 ±0.6
	MT-FI	89.8	84.4	85.3	84.7	84.1	84.0	82.0	79.8	80.3	77.4	77.7	80.6	74.7	71.8	71.3	80.5 ±0.3
Test set in target language (ZERO-SHOT)																	
	ORIG	90.4	84.4	85.5	84.3	81.9	83.6	80.1	80.1	79.8	81.8	78.3	80.3	77.7	72.8	74.5	81.0 ±0.2
	BT-ES	90.2	**86.0**	86.9	**86.5**	84.0	85.3	83.2	82.5	82.7	83.7	80.7	83.0	79.7	75.6	77.1	83.1 ±0.2
XLM-R	BT-FI	89.5	**86.0**	86.2	86.2	83.9	85.1	83.4	82.2	**83.0**	**83.9**	**81.2**	**83.9**	80.1	75.2	**78.1**	83.2 ±0.1
	MT-ES	89.9	85.7	87.3	85.6	83.9	85.4	82.9	82.0	82.3	83.6	80.0	82.6	79.9	75.5	76.8	82.9 ±0.4
	MT-FI	90.2	85.9	86.9	**86.5**	84.4	**85.5**	83.4	**83.0**	82.4	83.6	80.5	83.6	**80.4**	**76.5**	77.9	**83.4** ±0.2

Table 1: **XNLI dev results (acc).** BT-XX and MT-XX consistently outperform ORIG in all cases.

Both for NLI and QA, we run each system 5 times with different random seeds and report the average results. Space permitting, we also report the standard deviation across the 5 runs. In our result tables, we use an underline to highlight the best result within each block, and boldface to highlight the best overall result.

4 NLI experiments

We next discuss our main results in the XNLI development set (§4.1, §4.2), run additional experiments to better understand the behavior of our different variants (§4.3, §4.4, §4.5), and compare our results to previous work in the XNLI test set (§4.6).

4.1 TRANSLATE-TEST results

We start by analyzing XNLI development results for TRANSLATE-TEST. Recall that, in this approach, the test set is machine translated into English, but training is typically done on original English data. Our BT-ES and BT-FI variants close this gap by training on a machine translated English version of the training set generated through back-translation. As shown in Table 1, this brings substantial gains for both ROBERTA and XLM-R, with an average improvement of 4.6 points in the best case. Quite remarkably, MT-ES and MT-FI also outperform ORIG by a substantial margin, and are only 0.8 points below their BT-ES and BT-FI counterparts. Recall that, for these two systems, training is done in machine translated Spanish or Finnish, while inference is done in machine translated English. This shows that the loss of performance when generalizing from original data to machine translated data is substantially larger than the loss of performance when generalizing from one language to another.

4.2 ZERO-SHOT results

We next analyze the results for the ZERO-SHOT approach. In this case, inference is done in the test set in each target language which, in the case of XNLI, was human translated from English. As such, different from the TRANSLATE-TEST approach, neither training on original data (ORIG) nor training on machine translated data (BT-XX and MT-XX) makes use of the exact same type of text that the system is exposed to at test time. However, as shown in Table 1, both BT-XX and MT-XX outperform ORIG by approximately 2 points, which suggests that our (back-)translated versions of the training set are more similar to the human translated test sets than the original one. This also provides a new perspective on the TRANSLATE-TRAIN approach, which was reported to outperform ORIG in previous work (Conneau and Lample, 2019): while the original motivation was to train the model on the same language that it is tested on, our results show that machine translating the training set is beneficial even when the target language is different.

4.3 Original vs. translated test sets

So as to understand whether the improvements observed so far are limited to translated test sets or apply more generally, we conduct additional experiments comparing translated test sets to original ones. However, to the best of our knowledge, all

Model	Train	XNLI dev OR (en)	XNLI dev HT (es)	Our dataset OR (es)	Our dataset HT (en)	Our dataset MT (en)
ROBERTA	ORIG	**92.1**	-	-	78.7	79.0
	BT-ES	91.9	-	-	80.3	**80.5**
	BT-FI	91.4	-	-	**80.5**	**80.5**
XLM-R	ORIG	90.5	85.5	81.0	77.5	78.5
	BT-ES	90.3	87.1	**81.4**	78.6	79.4
	BT-FI	89.7	86.5	80.8	78.8	79.2
	MT-ES	90.2	**87.5**	81.3	78.4	78.9
	MT-FI	90.4	87.1	81.1	78.3	78.9

Table 2: **NLI results on original (OR), human translated (HT) and machine translated (MT) sets (acc).** BT-XX and MT-XX outperform ORIG in translated sets, but do not get any clear improvement in original ones.

Model	Train	Competence AT	Competence NR	Distraction WO	Distraction NG	Distraction LN	Noise SE
ROBERTA	ORIG	72.9	**65.7**	64.9	59.1	**88.4**	86.5
	BT-FI	56.6	57.2	**80.6**	67.8	87.7	**86.6**
XLM-R	ORIG	**78.4**	56.8	67.3	61.2	86.8	85.3
	BT-FI	60.6	51.7	76.7	64.6	86.2	85.4
	MT-FI	64.3	50.3	77.8	**68.5**	86.4	85.3

Table 3: **NLI Stress Test results (combined matched & mismatched acc).** AT = antonymy, NR = numerical reasoning, WO = word overlap, NG = negation, LN = length mismatch, SE = spelling error. BT-FI and MT-FI are considerably weaker than ORIG in the competence test, but substantially stronger in the distraction test.

existing non-English NLI benchmarks were created through translation. For that reason, we build a new test set that mimics XNLI, but is annotated in Spanish rather than English. We first collect the premises from a filtered version of CommonCrawl (Buck et al., 2014), taking a subset of 5 websites that represent a diverse set of genres: a newspaper, an economy forum, a celebrity magazine, a literature blog, and a consumer magazine. We then ask native Spanish annotators to generate an *entailment*, a *neutral* and a *contradiction* hypothesis for each premise.[5] We collect a total of 2490 examples using this procedure, which is the same size as the XNLI development set. Finally, we create a human translated and a machine translated English version of the dataset using professional translators from Gengo and our machine translation system described in §3.2,[6] respectively. We report results for the best epoch checkpoint on each set.

As shown in Table 2, both BT-XX and MT-XX clearly outperform ORIG in all test sets created through translation, which is consistent with our previous results. In contrast, the best results on the original English set are obtained by ORIG, and neither BT-XX nor MT-XX obtain any clear improvement on the one in Spanish either.[7] This confirms that the underlying phenomenon is limited to translated test sets. In addition, it is worth mentioning that the results for the machine translated test set in English are slightly better than those for the human translated one, which suggests that the difficulty of the task does not only depend on the translation quality. Finally, it is also interesting that MT-ES is only marginally better than MT-FI in both Spanish test sets, even if it corresponds to the TRANSLATE-TRAIN approach, whereas MT-FI needs to ZERO-SHOT transfer from Finnish into Spanish. This reinforces the idea that it is training on translated data rather than training on the target language that is key in TRANSLATE-TRAIN.

4.4 Stress tests

In order to better understand how systems trained on original and translated data differ, we run additional experiments on the NLI Stress Tests (Naik et al., 2018), which were designed to test the robustness of NLI models to specific linguistic phenomena in English. The benchmark consists of a competence test, which evaluates the ability to understand antonymy relation and perform numerical reasoning, a distraction test, which evaluates the robustness to shallow patterns like lexical overlap and the presence of negation words, and a noise test, which evaluates robustness to spelling errors. Just as with previous experiments, we report results for the best epoch checkpoint in each test set.

As shown in Table 3, ORIG outperforms BT-FI and MT-FI on the competence test by a large margin, but the opposite is true on the distraction test.[8] In particular, our results show that BT-FI and MT-FI are less reliant on lexical overlap and the presence of negative words. This feels intuitive, as translating the premise and hypothesis independently—as BT-FI and MT-FI do—is likely to reduce the lexical overlap between them. More generally, the trans-

[5] Unlike XNLI, we do not collect 4 additional labels for each example. Note, however, that XNLI kept the original label as the gold standard, so the additional labels are irrelevant for the actual evaluation. This is not entirely clear in Conneau et al. (2018), but can be verified by inspecting the dataset.

[6] We use beam search instead of sampling decoding.

[7] Note that the standard deviations are around 0.3.

[8] We observe similar trends for BT-ES and MT-ES, but omit these results for conciseness.

lation process can alter similar superficial patterns in the data, which NLI models are sensitive to (§2). This would explain why the resulting models have a different behavior on different stress tests.

4.5 Output class distribution

With the aim to understand the effect of the previous phenomenon in cross-lingual settings, we look at the output class distribution of our different models in the XNLI development set. As shown in Table 4, the predictions of all systems are close to the true class distribution in the case of English. Nevertheless, ORIG is strongly biased for the rest of languages, and tends to underpredict *entailment* and overpredict *neutral*. This can again be attributed to the fact that the English test set is original, whereas the rest are human translated. In particular, it is well-known that NLI models tend to predict *entailment* when there is a high lexical overlap between the premise and the hypothesis (§2). However, the degree of overlap will be smaller in the human translated test sets given that the premise and the hypothesis were translated independently, which explains why *entailment* is underpredicted. In contrast, BT-FI and MT-FI are exposed to the exact same phenomenon during training, which explains why they are not that heavily affected.

So as to measure the impact of this phenomenon, we explore a simple approach to correct this bias: having fine-tuned each model, we adjust the bias term added to the logit of each class so the model predictions match the true class distribution for each language.[9] As shown in Table 5, this brings large improvements for ORIG, but is less effective for BT-FI and MT-FI.[10] This shows that the performance of ORIG was considerably hindered by this bias, which BT-FI and MT-FI effectively mitigate.

4.6 Comparison with the state-of-the-art

So as to put our results into perspective, we compare our best variant to previous work on the XNLI test set. As shown in Table 6, our method improves the state-of-the-art for both the TRANSLATE-TEST and the ZERO-SHOT approaches by 4.3 and 2.8 points, respectively. It also obtains the best overall results published to date, with the additional advantage that the previous state-of-the-art required a machine translation system between English and each of the 14 target languages, whereas our method uses a single machine translation system between English and Finnish (which is not one of the target languages). While the main goal of our work is not to design better cross-lingual models, but to analyze their behavior in connection to translation, this shows that the phenomenon under study is highly relevant, to the extent that it can be exploited to improve the state-of-the-art.

[9] We achieve this using an iterative procedure where, at each step, we select one class and set its bias term so the class is selected for the right percentage of examples.

[10] Note that we are adjusting the bias term in the evaluation set itself, which requires knowing its class distribution and is thus a form of cheating. While useful for analysis, a fair comparison would require adjusting the bias term in a separate validation set. This is what we do for our final results in §4.6, where we adjust the bias term in the XNLI development set and report results on the XNLI test set.

Model	Train	EN			EN → XX (avg)		
		ent	neu	con	ent	neu	con
ROBERTA (*translate-test*)	ORIG	33.4	32.8	33.8	23.2	40.7	36.1
	BT-FI	34.5	31.9	33.6	30.2	35.7	34.1
XLM-R (*zero-shot*)	ORIG	32.4	33.2	34.4	27.0	37.8	35.2
	BT-FI	34.3	31.6	34.1	33.1	32.9	34.0
	MT-FI	33.6	32.6	33.9	30.8	35.3	33.9
Gold Standard		33.3	33.3	33.3	33.3	33.3	33.3

Table 4: **Output class distribution on XNLI dev.** All systems are close to the true distribution in English, but ORIG is biased toward *neu* and *con* in the transfer languages. BT-FI and MT-FI alleviate this issue.

Model	Train	Base	Unbias	+Δ
ROBERTA (*translate-test*)	ORIG	77.7 ±0.6	80.6 ±0.2	2.9 ±0.5
	BT-FI	82.3 ±0.2	82.8 ±0.1	0.4 ±0.2
XLM-R (*zero-shot*)	ORIG	81.0 ±0.2	82.4 ±0.2	1.4 ±0.3
	BT-FI	83.2 ±0.1	83.3 ±0.1	0.1 ±0.1
	MT-FI	83.4 ±0.2	83.8 ±0.1	0.4 ±0.2

Table 5: **XNLI dev results with class distribution unbiasing (average acc across all languages).** Adjusting the bias term of the classifier to match the true class distribution brings large improvements for ORIG, but is less effective for BT-FI and MT-FI.

5 QA experiments

So as to understand whether our previous findings apply to other tasks besides NLI, we run additional experiments on QA. As shown in Table 7, BT-FI and BT-ES do indeed outperform ORIG for the TRANSLATE-TEST approach on MLQA. The improvement is modest, but very consistent across different languages, models and runs. The results for MT-ES and MT-FI are less conclusive, presumably because mapping the answer spans across languages might introduce some noise. In contrast, we do not ob-

Model	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	avg
Fine-tune an English model and machine translate the test set into English (TRANSLATE-TEST)																
BERT (Devlin et al., 2019)	88.8	81.4	82.3	80.1	80.3	80.9	76.2	76.0	75.4	72.0	71.9	75.6	70.0	65.8	65.8	76.2
Roberta (Liu et al., 2019)	**91.3**	82.9	84.3	81.2	81.7	83.1	78.3	76.8	76.6	74.2	74.1	77.5	70.9	66.7	66.8	77.8
Proposed (ROBERTA – BT-FI)	90.6	85.4	86.3	84.3	85.2	85.7	82.3	80.6	81.5	77.8	78.6	81.2	77.1	73.5	72.3	81.5
+ Unbiasing (tuned in dev)	90.5	**85.8**	**86.6**	84.6	**85.5**	85.8	82.9	81.2	82.3	78.7	79.7	82.3	77.6	74.4	72.9	82.1
Fine-tune a multilingual model on all machine translated training sets (TRANSLATE-TRAIN-ALL)																
Unicoder (Huang et al., 2019)	85.6	81.1	82.3	80.9	79.5	81.4	79.7	76.8	78.2	77.9	77.1	80.5	73.4	73.8	69.6	78.5
XLM-R (Conneau et al., 2020)	88.7	85.2	85.6	84.6	83.6	85.5	82.4	81.6	80.9	**83.4**	**80.9**	**83.3**	79.8	75.9	74.3	82.4
Fine-tune a multilingual model on the English training set (ZERO-SHOT)																
mBERT (Devlin et al., 2019)	82.1	73.8	74.3	71.1	66.4	68.9	69.0	61.6	64.9	69.5	55.8	69.3	60.0	50.4	58.0	66.3
XLM (Conneau and Lample, 2019)	85.0	78.7	78.9	77.8	76.6	77.4	75.3	72.5	73.1	76.1	73.2	76.5	69.6	68.4	67.3	75.1
Unicoder (Huang et al., 2019)	85.1	79.0	79.4	77.8	77.2	77.2	76.3	72.8	73.5	76.4	73.6	76.2	69.4	69.7	66.7	75.4
XLM-R (Conneau et al., 2020)	88.8	83.6	84.2	82.7	82.3	83.1	80.1	79.0	78.8	79.7	78.6	80.2	75.8	72.0	71.7	80.1
Proposed (XLM-R – MT-FI)	88.8	84.8	85.7	84.6	84.2	85.7	82.9	81.8	82.0	82.1	79.9	81.8	79.8	75.9	76.7	82.4
+ Unbiasing (tuned in dev)	88.7	85.0	86.1	**84.8**	84.8	**86.1**	**83.5**	**82.2**	**82.4**	83.0	80.8	82.6	**80.3**	**76.0**	**77.3**	**82.9**

Table 6: **XNLI test results (acc).** Results for other methods are taken from their respective papers or, if not provided, from Conneau et al. (2020). For those with multiple variants, we select the one with the best results.

serve any clear improvement for the ZERO-SHOT approach on this dataset. Our XQuAD results in Table 8 are more positive, but still inconclusive.

These results can partly be explained by the translation procedure used to create the different benchmarks: the premises and hypotheses of XNLI were translated independently, whereas the questions and context paragraphs of XQuAD were translated together. Similarly, MLQA made use of parallel contexts, and translators were shown the sentence containing each answer when translating the corresponding question. As a result, one can expect both QA benchmarks to have more consistent translations than XNLI, which would in turn diminish this phenomenon. In contrast, the questions and context paragraphs are independently translated when using machine translation, which explains why BT-ES and BT-FI outperform ORIG for the TRANSLATE-TEST approach. We conclude that the translation artifacts revealed by our analysis are not exclusive to NLI, as they also show up on QA for the TRANSLATE-TEST approach, but their actual impact can be highly dependent on the translation procedure used and the nature of the task.

6 Discussion

Our analysis prompts to reconsider previous findings in cross-lingual transfer learning as follows:

The cross-lingual transfer gap on XNLI was overestimated. Given the parallel nature of XNLI, accuracy differences across languages are commonly interpreted as the loss of performance when generalizing from English to the rest of languages. However, our work shows that there is another factor that can have a much larger impact: the loss of performance when generalizing from original to translated data. Our results suggest that the real cross-lingual generalization ability of XLM-R is considerably better than what the accuracy numbers in XNLI reflect.

Overcoming the cross-lingual gap is not what makes TRANSLATE-TRAIN work. The original motivation for TRANSLATE-TRAIN was to train the model on the same language it is tested on. However, we show that it is training on translated data, rather than training on the target language, that is key for this approach to outperform ZERO-SHOT as reported by previous authors.

Improvements previously attributed to data augmentation should be reconsidered. The method by Singh et al. (2019) combines machine translated premises and hypotheses in different languages (§2), resulting in an effect similar to BT-XX and MT-XX. As such, we believe that this method should be analyzed from the point of view of dataset artifacts rather than data augmentation, as the authors do.[11] From this perspective, having the premise and the hypotheses in different languages can reduce the superficial patterns between them, which would explain why this approach is better than using examples in a single language.

[11] Recall that our experimental design prevents a data augmentation effect, in that the number of unique sentences and examples used for training is always the same (§3.2).

Model	Train	en	es	de	ar	vi	zh	hi	avg
Test set machine translated into English (TRANSLATE-TEST)									
ROBERTA	ORIG	**84.7** / **71.4**	70.1 / 49.7	60.5 / 41.2	55.7 / 32.5	65.6 / 40.8	53.5 / 26.0	42.7 / 20.7	61.8 ±0.1 / 40.3 ±0.2
	BT-ES	84.4 / 71.2	70.9 / 50.7	61.0 / 41.6	56.5 / 33.3	66.7 / 41.8	54.4 / 27.1	43.0 / 21.1	62.4 ±0.1 / 41.0 ±0.2
	BT-FI	83.8 / 70.4	70.3 / 50.1	61.1 / 41.9	56.5 / 33.4	66.8 / 42.1	54.9 / 27.5	42.8 / 21.3	62.3 ±0.1 / 40.9 ±0.2
XLM-R	ORIG	84.1 / 71.0	69.9 / 49.2	60.8 / 42.5	55.2 / 31.8	65.4 / 40.6	54.3 / 27.8	43.6 / 21.3	61.9 ±0.1 / 40.6 ±0.1
	BT-ES	83.8 / 70.8	70.5 / 50.0	61.4 / 43.5	56.1 / 33.1	66.5 / 41.6	55.4 / 29.0	44.0 / 22.2	62.5 ±0.2 / 41.5 ±0.2
	BT-FI	82.7 / 69.6	70.0 / 49.7	61.1 / 43.3	56.0 / 33.1	66.2 / 41.5	55.6 / 29.2	43.7 / 22.0	62.2 ±0.1 / 41.2 ±0.2
	MT-ES	83.4 / 69.7	70.0 / 49.1	61.0 / 42.7	55.6 / 32.2	65.9 / 40.9	54.9 / 28.1	43.9 / 21.6	62.1 ±0.3 / 40.6 ±0.2
	MT-FI	82.6 / 69.0	69.7 / 48.6	61.0 / 42.8	55.7 / 32.3	65.8 / 40.9	54.8 / 27.9	43.9 / 21.6	61.9 ±0.3 / 40.4 ±0.2
Test set in target language (ZERO-SHOT)									
XLM-R	ORIG	84.1 / 71.0	74.5 / 56.3	70.3 / 55.1	66.5 / 45.9	74.3 / 53.1	67.8 / 43.4	71.6 / 53.4	72.7 ±0.1 / 54.0 ±0.1
	BT-ES	83.8 / 70.8	74.7 / 56.8	70.3 / **55.2**	66.9 / **46.5**	74.3 / 53.0	**68.2** / **43.8**	71.4 / **53.6**	72.8 ±0.2 / **54.3** ±0.2
	BT-FI	82.7 / 69.6	74.1 / 56.3	69.8 / 54.5	66.6 / 46.0	73.3 / 52.3	67.9 / 43.4	71.0 / 53.2	72.2 ±0.2 / 53.6 ±0.2
	MT-ES	83.4 / 69.7	**75.2** / **57.3**	**70.5** / 55.1	**67.5** / **46.5**	**74.5** / **53.2**	67.5 / 42.5	71.7 / 52.7	**72.9** ±0.3 / 53.9 ±0.4
	MT-FI	82.6 / 69.0	74.1 / 56.0	70.2 / 54.6	66.9 / 46.0	73.7 / 52.6	67.2 / 41.5	**71.9** / 53.4	72.4 ±0.2 / 53.3 ±0.4

Table 7: **MLQA test results (F1 / exact match).**

Model	Train	en	es	de	el	ru	tr	ar	vi	th	zh	hi	avg
XLM-R (*zero-shot*)	ORIG	**88.2**	82.7	**80.8**	80.9	80.1	76.1	76.0	80.1	75.4	71.9	76.4	79.0 ±0.2
	BT-ES	87.9	83.5	80.5	**81.2**	**80.7**	**76.8**	**77.4**	80.2	76.4	73.0	76.9	**79.5** ±0.3
	BT-FI	87.1	82.5	80.2	80.7	79.8	75.7	76.6	79.4	75.7	71.5	76.8	78.7 ±0.3
	MT-ES	87.1	**84.1**	80.3	**81.2**	80.1	76.0	**77.4**	**80.9**	76.7	72.7	77.1	79.4 ±0.3
	MT-FI	86.3	81.4	80.2	80.5	80.2	76.6	77.0	80.3	**77.6**	**74.5**	**77.8**	79.3 ±0.2

Table 8: **XQuAD results (F1).** Results for the exact match metric are similar.

The potential of TRANSLATE-TEST was underestimated. The previous best results for TRANSLATE-TEST on XNLI lagged behind the state-of-the-art by 4.6 points. Our work reduces this gap to only 0.8 points by addressing the underlying translation artifacts. The reason why TRANSLATE-TEST is more severely affected by this phenomenon is twofold: (i) the effect is doubled by first using human translation to create the test set and then machine translation to translate it back to English, and (ii) TRANSLATE-TRAIN was inadvertently mitigating this issue (see above), but equivalent techniques were never applied to TRANSLATE-TEST.

Future evaluation should better account for translation artifacts. The evaluation issues raised by our analysis do not have a simple solution. In fact, while we use the term *translation artifacts* to highlight that they are an unintended effect of translation that impacts final evaluation, one could also argue that it is the original datasets that contain the artifacts, which translation simply alters or even mitigates.[12] In any case, this is a more general issue that falls beyond the scope of cross-lingual transfer learning, so we argue that it should be carefully controlled when evaluating cross-lingual models. In the absence of more robust datasets, we recommend that future multilingual benchmarks should at least provide consistent test sets for English and the rest of languages. This can be achieved by (i) using original annotations in all languages, (ii) using original annotations in a non-English language and translating them into English and other languages, or (iii) if translating from English, doing so at the document level to minimize translation inconsistencies.

7 Conclusions

In this paper, we have shown that both human and machine translation can alter superficial patterns in data, which requires reconsidering previous findings in cross-lingual transfer learning. Based on the gained insights, we have improved the state-of-the-art in XNLI for the TRANSLATE-TEST and ZERO-SHOT approaches by a substantial margin. Finally, we have shown that the phenomenon is not specific to NLI but also affects QA, although it is less pronounced there thanks to the translation procedure used in the corresponding benchmarks. So as to facilitate similar studies in the future, we release

[12] For instance, the high lexical overlap observed for the *entailment* class is usually regarded a spurious pattern, so reducing it could be considered a positive effect of translation.

our NLI dataset,[13] which, unlike previous benchmarks, was annotated in a non-English language and human translated into English.

Acknowledgments

We thank Nora Aranberri and Uxoa Iñurrieta for helpful discussion during the development of this work, as well as the rest of our colleagues from the IXA group that worked as annotators for our NLI dataset.

This research was partially funded by a Facebook Fellowship, the Basque Government excellence research group (IT1343-19), the Spanish MINECO (UnsupMT TIN2017-91692-EXP MCIU/AEI/FEDER, UE), Project BigKnowledge (Ayudas Fundación BBVA a equipos de investigación científica 2018), and the NVIDIA GPU grant program.

This research is supported via the BETTER Program contract #2019-19051600006 (ODNI, IARPA activity). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637. Association for Computational Linguistics.

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.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Christian Buck, Kenneth Heafield, and Bas van Ooyen. 2014. N-gram counts and language models from the Common Crawl. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3579–3584, Reykjavik, Iceland. European Language Resources Association (ELRA).

Casimiro Pio Carrino, Marta R. Costa-jussà, and José A. R. Fonollosa. 2020. Automatic Spanish translation of the SQuAD dataset for multi-lingual question answering. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5515–5523, Marseille, France. European Language Resources Association.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32*, pages 7059–7069.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*,

[13] https://github.com/artetxem/esxnli

pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Miquel Esplà, Mikel Forcada, Gema Ramírez-Sánchez, and Hieu Hoang. 2019. ParaCrawl: Web-scale parallel corpora for the languages of the EU. In *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*, pages 118–119, Dublin, Ireland. European Association for Machine Translation.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia. Association for Computational Linguistics.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2019. Translationese in machine translation evaluation. *arXiv preprint arXiv:1906.09833*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *arXiv preprint arXiv:2003.11080*.

Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. 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 2485–2494, Hong Kong, China. Association for Computational Linguistics.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Divyansh Kaushik and Zachary C. Lipton. 2018. How much reading does reading comprehension require? A critical investigation of popular benchmarks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5010–5015, Brussels, Belgium. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea. Association for Computational Linguistics.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 1–9, Brussels, Belgium. Association for Computational Linguistics.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. WikiMatrix: Mining 135m parallel sentences in 1620 language pairs from Wikipedia. *arXiv preprint arXiv:1907.05791*.

Holger Schwenk and Xian Li. 2018. A corpus for multilingual document classification in eight languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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.

Jasdeep Singh, Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2019. XLDA: Cross-lingual data augmentation for natural language inference and question answering. *arXiv preprint arXiv:1905.11471*.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008.

Vered Volansky, Noam Ordan, and Shuly Wintner. 2013. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. 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 3687–3692, Hong Kong, China. 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.