5-2022

Translate-train embracing translationese artifacts

Sicheng YU
Singapore Management University, scyu.2018@phdcs.smu.edu.sg

Qianru SUN
Singapore Management University, qianrusun@smu.edu.sg

Hao ZHANG

Jing JIANG
Singapore Management University, jingjiang@smu.edu.sg

Citation

YU, Sicheng; SUN, Qianru; ZHANG, Hao; and JIANG, Jing. Translate-train embracing translationese artifacts. (2022). *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Dublin, Ireland, 2022 May 22-27*. 362-370.
Available at: https://ink.library.smu.edu.sg/sis_research/7475

Translate-Train Embracing Translationese Artifacts

Sicheng Yu[♠] **Qianru Sun**[♠] **Hao Zhang**[♣◇] **Jing Jiang**[♠]
[♠]Singapore Management University, Singapore
[♣]Nanyang Technological University, Singapore
[◇]Centre for Frontier AI Research, A*STAR, Singapore
scyu.2018@phdcs.smu.edu.sg, hzhang26@outlook.com
{qianrusun,jingjiang}@smu.edu.sg

Abstract

Translate-train is a general training approach to multilingual tasks. The key idea is to use the translator of the target language to generate training data to mitigate the gap between the source and target languages. However, its performance is often hampered by the artifacts in the translated texts (translationese). We discover that such artifacts have common patterns in different languages and can be modeled by deep learning, and subsequently propose an approach to conduct translate-train using *T*ranslationese *E*mbracing the effect of *A*rtifacts (TEA). TEA learns to mitigate such effect on the training data of a source language (whose original and translationese are both available), and applies the learned module to facilitate the inference on the target language. Extensive experiments on the multilingual QA dataset TyDiQA demonstrate that TEA outperforms strong baselines.

1 Introduction

Cross-lingual transfer has drawn wide attention in recent years (Hu et al., 2020; Liang et al., 2020). It has great potentials to be applied in advanced industries and real applications such as for improving dialog and advertisement systems in multilingual countries (Schuster et al., 2019; Yu et al., 2021). It aims to reuse NLP models trained on a *source* language for the task of a *target* language. The most intuitive method is transfer learning by leveraging pre-trained multilingual language models (LMs) such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020). These pre-trained LMs encode different languages into a joint space of multilingual representations (Wu and Dredze, 2019; Lauscher et al., 2020), and they perform well especially for zero-shot cross-lingual tasks (Wu and Dredze, 2019; Lauscher et al., 2020). Another method orthogonal to this is called translate-train (Hu et al., 2020; Fang et al., 2021). It translates training data from the source language into the target language and uses the translated texts for training. Our paper focuses on this second method.

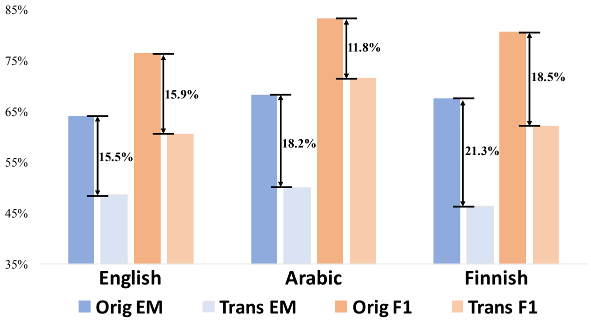


Figure 1: Monolingual QA performance comparison between (i) training using originals and (ii) training using translated texts on TyDiQA. Note that for each language, training data and test data are in the same language. "EM" stands for Exact Match.

Translate-train mitigates the language gap between the source and the target languages in multilingual tasks in a straightforward manner as it directly generates the needed target language training samples. However, the translation process introduces artifacts in the translated texts (*i.e.*, translationese[1]). It has been observed that translationese often exhibits features such as stylistic ones that are different from text written directly in the same language (which we call the *originals*) and thus can mislead model training (Selinker, 1972; Volansky et al., 2015; Bizzoni et al., 2020). In Figure 1, we show that even in a monolingual setting where training and test data are in the same language, when the test data are original texts, using translationese to train a QA model results in substantial performance drop compared with using originals for training.

To tackle the issue with translationese artifacts, inspired by domain mapping techniques (Zhu et al., 2017), we explore the idea of projecting originals

[1] We refer to texts directly written by humans in a certain language as *originals* of that language and translated texts (translated by either humans or machines) as *translationese*.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 362 - 370
May 22-27, 2022 ©2022 Association for Computational Linguistics

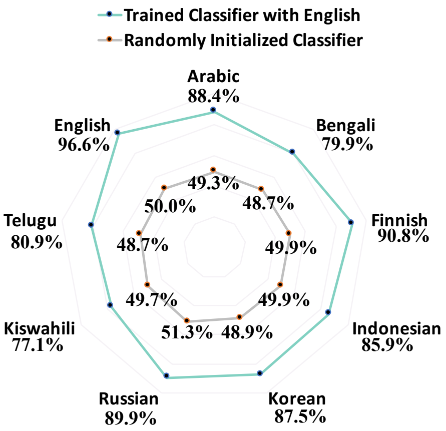


Figure 2: The XLM-R (Conneau et al., 2020) classification results of translationese for different languages on TyDiQA (Clark et al., 2020). We use the classifiers trained with only English pairs (originals and translationese), or randomly initialized (without any training). More details of the experiments are given in Appendix B.

and translationese into a common embedding space to close their gap. Since only the originals of the source language are available under the translate-train setting, whether this idea is feasible depends on whether the projection function is learnable from the source language and transferable to other languages. Therefore, we first conduct experiments to investigate if translationese artifacts, or patterns of differences between orginals and translationese, are recognizable and transferable across languages by deep learning models. Specifically, we train a binary classifier to distinguish English originals from English translationese. We then test the effectiveness of this binary classifier on other languages. Our intuition is that (1) if the model converges, it suggests that the patterns of translationese artifacts can be potentially learned to some extent, and 2) if the trained model recognizes the translationese of other languages, it means the model can likely transfer the learned patterns across different languages. Our results in Figure 2 validate both: 1) The model converges well and achieves 97% accuracy on English, the training language. (2) It also performs reasonably well on other languages (77% ~ 91%).

Based on the above intuition and validation, we propose a Translationese Embracing Artifacts (TEA) method that projects originals and translationese into a common space to mitigate the translationese artifacts. TEA explicitly learns a mapping function from originals to translationese using originals and translationese of the source language (English in our experiments), where learning is through minimizing the distance between the mapped representation of originals and of the corresponding translationese. TEA then applies this mapping function to the originals of the target language during the testing stage. For evaluation, we conduct experiments on multilingual QA using the TyDiQA dataset (Clark et al., 2020)[2]. Our results show that TEA outperforms translate-train baselines and SOTA translationese mitigation methods designed for machine translation (Marie et al., 2020; Wang et al., 2021).

2 Related Work

The effect of translationese has been widely studied in translation tasks (Lembersky et al., 2012; Zhang and Toral, 2019; Edunov et al., 2020; Graham et al., 2020; Freitag et al., 2020). Some works focus on mitigating or controlling the effect of translationese, *e.g.*, tagged training (Marie et al., 2020; Riley et al., 2020; Wang et al., 2021), which are adopted as baselines in our paper. In the field of cross-lingual transfer, there are very few works about translationese. Artetxe et al. (2020) is the only attempt for translate-test and zero-shot learning. In contrast, we focus on translate-train and aim to mitigate the artifacts in translationese.

Our research is also related to domain adaptation (DA) that aims to transfer the knowledge from a source domain to target domains. Our original-to-translationese projection function can be seen as something similar to projecting source domain and target domain data into a common space, which has been used before for domain adaptation (Zhu et al., 2017; Shen et al., 2017).

3 Our Approach (TEA)

Let $\mathbf{x}$ represent the input text and $\mathbf{y}$ represent the output label. $\mathcal{X}$ denotes the domain (*i.e.*, all possible values) of $\mathbf{x}$ and $\mathcal{Y}$ is the set of labels. The input $\mathbf{x}$ comes from different languages, and it can be either originals or translationese during training. Specifically, we use $\mathcal{X}_{\text{src, orig}}$ to denote the domain of *source* language *originals*, and define $\mathcal{X}_{\text{trgt, orig}}$ and $\mathcal{X}_{\text{trgt, trans}}$ in a similar way (where $_{\text{trgt}}$ refers to the target language and $_{\text{trans}}$ refers to translationese). We further use back-translation (Sennrich et al., 2016) to generate *source* language *trans-*

[2] To the best of our knowledge, TyDiQA is the only large-scale multilingual benchmark dataset where test data is original text written by humans.

lationese (*i.e.*, the source language originals are first translated into a pivot language and then translated back into the source language), denoted as $\mathcal{X}_{\text{src, trans}}$, for the purpose of learning a mapping function to project originals and translationese into the same space.

We now present our TEA method. Our ultimate goal is to learn a mapping function $f : \mathcal{X}_{\text{trgt, orig}} \rightarrow \mathcal{Y}$, which takes target language originals as input. However, we only have $\mathcal{D}_{\text{src, orig}} \in \mathcal{X}_{\text{src, orig}} \times \mathcal{Y}$ and $\mathcal{D}_{\text{trgt, trans}} \in \mathcal{X}_{\text{trgt, trans}} \times \mathcal{Y}$ during training. The challenge is that an f learned from either $\mathcal{D}_{\text{src, orig}}$ or $\mathcal{D}_{\text{trgt, trans}}$ may not work effectively on $\mathcal{X}_{\text{trgt, orig}}$ because of the differences between the source and the target languages and between originals and translationese. To mitigate the differences between the source and the target languages, we rely on pre-trained multilingual language models, as many existing works do. As for the differences between originals and translationese, based on the idea discussed in Section 1, we propose to mitigate the translationese artifacts of the target language using an original-to-translationese mapping function, and because of the lack of target originals, we propose to learn the original-to-translationese mapping function from the *source* language.

To concretely illustrate our idea, we break down the mapping from $\mathcal{X}$ to $\mathcal{Y}$ into the following steps[3]:

Multilingual Projection (MP): First, input $\mathbf{x}$ is projected into a language-agnostic multilingual space by using a pre-trained multilingual LM. We use $\mathcal{X}_{\text{ml}}$ to denote the projected multilingual space, and f_{MP} is a multilingual projection (*i.e.*, the multilingual LM) that maps an input $\mathbf{x}$ in any language into $\mathcal{X}_{\text{ml}}$.

Original-to-Translationese Projection (OTP): Suppose $\mathcal{X}_{\text{ml}}$ consists of two subspaces: $\mathcal{X}_{\text{ml}} = \mathcal{X}_{\text{ml, orig}} \bigcup \mathcal{X}_{\text{ml, trans}}$, where $\mathcal{X}_{\text{ml, orig}}$ and $\mathcal{X}_{\text{ml, trans}}$ denote the multilingual representations of any originals and translationese, respectively. To close the gap between originals and translationese, we define an original-to-translationese projection function $f_{\text{OTP}} : \mathcal{X}_{\text{ml, orig}} \rightarrow \mathcal{X}_{\text{ml, trans}}$ to convert the vector representation of a piece of originals to its corresponding representation of translationese.

Language-Agnostic QA (QA): The last step is a language-agnostic classifier for the QA task itself. We use $f_{\text{QA}} : \mathcal{X}_{\text{ml, trans}} \rightarrow \mathcal{Y}$ to denote it.

Given an input $\mathbf{x}$, depending on whether it is from originals or translationese, we use different compositions of the functions above to map $\mathbf{x}$ to $\mathbf{y}$:

$$\mathbf{y} = \begin{cases} f_{\text{QA}} \circ f_{\text{OTP}} \circ f_{\text{MP}}(\mathbf{x}) & \mathbf{x} \in \mathcal{X}_{*,\text{ orig}}, \\ f_{\text{QA}} \circ f_{\text{MP}}(\mathbf{x}) & \mathbf{x} \in \mathcal{X}_{*,\text{ trans}}. \end{cases}$$

Here $\circ$ represents the composition of two functions, *i.e.*, $f \circ g(x) = f(g(x))$, and $*$ denotes source language or target languages. More concretely, for $(\mathbf{x}, \mathbf{y}) \in \mathcal{D}_{\text{src, orig}}$, we use $\mathcal{X}_{\text{src, orig}} \xrightarrow{f_{\text{MP}}} \mathcal{X}_{\text{ml, orig}} \xrightarrow{f_{\text{OTP}}} \mathcal{X}_{\text{ml, trans}} \xrightarrow{f_{\text{QA}}} \mathcal{Y}$; for $(\mathbf{x}, \mathbf{y}) \in \mathcal{D}_{\text{trgt, trans}}$, we use $\mathcal{X}_{\text{trgt, trans}} \xrightarrow{f_{\text{MP}}} \mathcal{X}_{\text{ml, trans}} \xrightarrow{f_{\text{QA}}} \mathcal{Y}$.

As discussed in Section 1, we make use of the source language originals and translationese to learn f_{OTP}. Specifically, for $(\mathbf{x}, \mathbf{y}) \in \mathcal{D}_{\text{src, orig}}$, we use $\mathbf{x}' \in \mathcal{X}_{\text{src, trans}}$ to represent its corresponding translationese, *i.e.*, generated by back-translation (Sennrich et al., 2016) through a pivot language. Let $\{(\mathbf{x}, \mathbf{x}')\} \in \mathcal{D}_{\text{src, pairs}}$ denotes all the pairs of originals and translationese in the source language. Then, we minimize the distance between $f_{\text{OTP}}(f_{\text{MP}}(\mathbf{x}))$ and $f_{\text{MP}}(\mathbf{x}')$ to optimize f_{OTP}.

In summary, the loss function consists of the following three components:

$$\begin{aligned} L \quad = & \sum_{(\mathbf{x},\mathbf{y}) \in \mathcal{D}_{\text{src, orig}}} l(f_{\text{QA}} \circ f_{\text{OTP}} \circ f_{\text{MP}}(\mathbf{x}), \mathbf{y}) \\ & + \sum_{(\mathbf{x},\mathbf{y}) \in \mathcal{D}_{\text{trgt, trans}}} l(f_{\text{QA}} \circ f_{\text{MP}}(\mathbf{x}), \mathbf{y}) \\ & + \sum_{(\mathbf{x},\mathbf{x}') \in \mathcal{D}_{\text{src, pairs}}} (1 - g(\mathbf{x}, \mathbf{x}')), \end{aligned}$$

where $g(\mathbf{x}, \mathbf{x}') = \cos(f_{\text{OTP}}(f_{\text{MP}}(\mathbf{x})), f_{\text{MP}}(\mathbf{x}')$. $l(\cdot, \cdot)$ is standard cross entropy loss and $\cos(\cdot, \cdot)$ is the cosine similarity function.

Model Details. For f_{MP}, we use XLM-R (Conneau et al., 2020). For f_{OTP}, we utilize a transformer layer (Vaswani et al., 2017). f_{QA} is implemented by a linear layer.

4 Experiments

Dataset. We conduct experiments on TyDiQA (Clark et al., 2020). Specifically, we evaluate our approach on the gold-passage subtask of TyDiQA, which includes 9 languages. We set English as source language and others as target languages, and report the results on target languages. During training, we utilize translated training data in *all* target languages for joint training. We use Exact Match (EM) and F1 scores as evaluation metrics.

Implementation. Translations of English training data for target languages are from XTREME (Hu

[3] A diagram showing the pipeline is in Appendix A.

Method	**D**	**ar**	**bn**	**fi**	**id**	**ko**	**ru**	**sw**	**te**	***med***	***all-in-one***	***avg***
STT	✘	40.4/67.6	47.8/64.0	53.2/70.5	61.9/77.4	10.9/31.9	42.1/67.0	48.1/66.1	43.6/70.1	45.7/67.3	45.2/67.2	43.5/64.3
FILTER	✘	50.8/72.8	56.6/70.5	57.2/73.3	59.8/76.8	12.3/33.1	46.6/68.9	65.7/77.4	50.4/69.9	53.7/71.7	51.6/70.3	49.9/67.8
STT*	✘	58.0/76.6	54.6/70.2	59.0/74.8	64.7/80.2	48.0/61.6	49.5/71.2	58.7/74.6	57.0/76.2	57.5/74.7	56.8/74.4	56.2/73.2
TAG*	✔	56.9/76.4	55.5/70.0	59.4/75.2	64.4/79.6	48.6/61.7	49.1/70.4	60.7/76.0	57.8/76.4	57.4/75.5	56.9/74.5	56.5/73.2
TST*	✔	58.4/75.5	60.2/72.2	58.3/74.4	65.5/78.9	49.3/62.6	49.0/69.7	63.5/76.7	56.2/76.1	58.3/75.0	57.3/74.1	57.6/73.3
GRL*	✔	57.6/75.6	58.4/72.6	59.7/74.8	65.3/79.9	49.6/62.2	49.1/70.4	62.9/76.9	58.2/77.0	58.3/75.2	57.6/74.6	57.6/73.7
TEA*	✔	56.5/76.1	60.2/74.9	60.9/76.5	63.6/79.3	48.6/61.4	51.5/72.0	66.7/78.9	60.7/78.7	**60.5/76.3**	**58.6/75.6**	**58.6/74.7**

Table 1: Main results (Exact Match / F1 scores) on TyDiQA. All methods use XLM-R as backbone. The "D" column indicates whether the model design considers translationese artifacts. The columns "ar" to "te" are different target languages. The "med" and "avg" columns denote median and average performance across the 8 target languages. The "all-in-one" column is the result by combining all data as one dataset. * indicates our implementation.

et al., 2020) and translationese English is translated by Google Cloud Translation. German (de) is selected as the default pivot language in back-translation. More details are in the Appendix B.

Baselines. We compare our model with the following baselines: (1) Standard Translate-Train (STT) (Devlin et al., 2019). (2) FILTER (Fang et al., 2021) is an advanced translate-train method fully utilizing the parallel data. (3) Tagging (TAG) (Marie et al., 2020), which distinguishes originals and translationese by adding a tag for machine translation. (4) Two-Stage Training (TST) (Wang et al., 2021), which is another approach to address the gap between translationese and originals for machine translation. It first uses the combination of them for training followed by another round of training only on originals. (5) Gradient Reversal Layer (GRL) (Ganin and Lempitsky, 2015), which is a general DA method.

Main results. Table 1 summarizes the comparison between our TEA and the baselines. We make the following observations: (1) TEA outperforms all baselines. For instance, TEA surpasses STT by 2.4% (EM) and 1.5% (F1) on average, which demonstrates the effectiveness of our method. (2) Methods considering translationese artifacts generally perform better than methods without such design, which reinforces the importance of mitigating translationese artifacts. (3) Compared to the baselines for translationese artifacts, TEA still shows its superiority. We highlight that our OTP module with explicit projection is better than implicit DA approaches. E.g., TAG only uses different tags to distinguish the translationese from originals. (4) The improvements from TEA across different languages are different. For high-resource[4] target languages, TEA brings more gains on the languages in Indo-European family, *e.g.*, ru, and marginal gains on others, *e.g.*, ar. For low-resource target languages, the performance improvements are obvious, *e.g.*, sw. It is because both language model and machine translation model are of lower quality on low-resource languages, and thus mitigating the gap between translationese and originals shows more effectiveness in such scenario. For high-resource languages, TEA prefers Indo-European languages, which are closer to English.

Settings	**EM**	**F1**
STT	56.2	73.2
(1) STT+$\mathcal{X}_{\text{src, trans}}$	56.6	73.2
(2) STT+params	56.3	73.5
(3) TOP	57.9	74.1
(4) MLP in OTP	56.7	73.3
(5) MSE loss	58.0	73.9
Full method	**58.6**	**74.7**

Table 2: Ablation study on TyDiQA. We report the average EM and F1 performance on the 8 target languages.

Ablation studies. We conduct in-depth ablation studies to analyze TEA. Specifically, we explore the following settings: (1) Since we use 11% more data in TEA (unlabeled $\mathcal{X}_{\text{src, trans}}$) compared to STT, here we add labeled $\mathcal{X}_{\text{src, trans}}$ in STT. (2) Since we use additional 0.38% parameters (OTP) in our method compared to STT, here we add the same OTP module in STT. (3) We replace the Original-to-Translationese Projection (OTP) by Translationese-to-Original Projection (TOP). (4) We replace the self-attention layer in OTP with a multi-layer perceptron (MLP). (5) We replace the cosine distance function in loss with mean square function. The results are summarized in Table 2. Compared to the variants, our full method performs best over all settings. (1)/(2) incorporate additional data/parameters, which demonstrates the improvement of our method is not caused by the two factors.

[4] Here we distinguish high-resource and low-resource according to XLM-R (Conneau et al., 2020).

Settings	Language Family	EM	F1
Scottish (gd)	Indo-European	58.8	74.0
Korean (ko)	Koreanic	57.8	74.0
Chinese (zh)	Sino-Tibetan	57.6	73.8
German (de)	Indo-European	58.6	74.7

Table 3: Experiment results of utilizing different language as pivot language for generating $\mathcal{X}_{\text{src, trans}}$.

(3) proves that TOP still mitigates the artifacts, but OTP obtaining better performance. We argue that it is because most of the training data is translationese. (4) and (5) demonstrate the effectiveness of our loss function and architecture.

Pivot Languages Analysis. Here we study the effect of pivot language used in generating $\mathcal{X}_{\text{src, trans}}$. Specifically, we select four pivot languages, *i.e.*, German (de), Scottish (gd), Korean (ko) and Chinese (zh), for evaluation. We fix our approach and only replace the $\mathcal{X}_{\text{src, trans}}$ used in OTP. The results are reported in Table 3. We observe that pivot languages from Indo-European family are superior to that from other language families. We believe it is because the training data of other target languages in translate-train is translated from English, while English belongs to the Indo-European family.

5 Conclusions

We aim to mitigate the translationese artifacts when training translate-train models. After varifying the transferability of the translationese patterns across languages, we propose the TEA that mitigates artifacts using a source language and to facilitate the prediction on unseen target languages. Our approach is simple and generic. Moreover, our results on multilingual QA show its effectiveness.

Ethical Considerations

Although our method requires fine-tuning of the pre-trained multilingual language model, the computational cost of our experiments is not high. We utilize two pieces of NVIDIA V100 and it takes around 1 hour for the fine-tuning process. This is partly due to the relatively small QA training dataset used for fine-tuning. It is possible that if our method is applied to either a much larger training dataset for fine-tuning or a much larger pre-trained language model, the computational cost and power consumption will go up. To reduce such costs, one way is to fine-tune only part of the pre-trained language model. Another way is to apply the recently proposed Adapter method (Houlsby et al., 2019) to fine-tune the language model.

Our method relies on machine translation systems. It has been found in a previous study that industrial MT systems as well as SOTA academic MT systems may suffer from gender bias (Stanovsky et al., 2019), and it would not be surprising if other types of societal biases and stereotypes are also found in machine translated texts. If our method uses translationese containing societal biases, our learned original-to-translationese projection function will likely also contain such biases, which may affect the fairness of the final trained system. However, this is not due to our method but rather the translated text we use. Nevertheless, this is something we need to keep in mind if our method is adopted for real applications.

Acknowledgments

This research is supported by the National Research Foundation, Singapore under its Strategic Capabilities Research Centres Funding Initiative, and partially supported by the Agency for Science, Technology and Research (A*STAR) under its AME YIRG Grant (Project No. A20E6c0101) and the A*STAR Career Development Fund (Project No. C210812033). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684.

Yuri Bizzoni, Tom S Juzek, Cristina Espana-Bonet, Koel Dutta Chowdhury, Josef van Genabith, and Elke Teich. 2020. How human is machine translationese? comparing human and machine translations of text and speech. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 280–290.

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, Édouard 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.

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)*, pages 4171–4186.

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.

Yuwei Fang, Shuohang Wang, Zhe Gan, Siqi Sun, and Jingjing Liu. 2021. Filter: An enhanced fusion method for cross-lingual language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 14, pages 12776–12784.

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.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2020. Translationese in machine translation evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 72–81.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multitask benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. 2012. Adapting translation models to translationese improves smt. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 255–265.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, et al. 2020. Xglue: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018.

Fuli Luo, Wei Wang, Jiahao Liu, Yijia Liu, Bin Bi, Songfang Huang, Fei Huang, and Luo Si. 2021. VECO: Variable and flexible cross-lingual pre-training for language understanding and generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 3980–3994.

Benjamin Marie, Raphael Rubino, and Atsushi Fujita. 2020. Tagged back-translation revisited: Why does it really work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5990–5997.

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.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. 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)*, pages 3795–3805.

Larry Selinker. 1972. Interlanguage. *International Review of Applied Linguistics in Language Teaching*, pages 209–231.

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.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. *Advances in Neural Information Processing Systems*, 30.

Gabriel Stanovsky, Noah A Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

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

Shuo Wang, Zhaopeng Tu, Zhixing Tan, Shuming Shi, Maosong Sun, and Yang Liu. 2021. On the language coverage bias for neural machine translation. In *Findings of the 59th Annual Meeting of Association for Computational Linguistics*.

Xiangpeng Wei, Rongxiang Weng, Yue Hu, Luxi Xing, Heng Yu, and Weihua Luo. 2020. On learning universal representations across languages. In *International Conference on Learning Representations*.

Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. 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 833–844.

Puxuan Yu, Hongliang Fei, and Ping Li. 2021. Cross-lingual language model pretraining for retrieval. In *Proceedings of the Web Conference 2021*, pages 1029–1039.

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.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232.

A Training Pipeline

Figure 3 illustrates the training pipelines our method. The goal is to map the originals and translationese domains into the same embedding space for prediction. Specifically, the module on the top of the figure is to train the $\mathcal{D}_{\text{src, orig}}$ and $\mathcal{D}_{\text{trgt, trans}}$, *i.e.*, the first two terms of loss function, while the module at the bottom aims to map the originals-translationese pairs in $\mathcal{D}_{\text{src, pairs}}$ into same space, *i.e.*, the last term of loss function.

B Implementation

General Implementation. We adopt the HuggingFace Transformers (Wolf et al., 2020) toolkit to implement the pre-trained language model, *i.e.*, XLM-R. The maximal input length, *i.e.*, concatenation of question and passage tokens, is set as 384. We also utilize a document sliding window with stride length of 128 to tackle the long passage issue. The learning rate and batch size are set as $2e-5$ and 32, respectively. We use back-translate (Sennrich et al., 2016) to generate the translationese. Back-translate means to translate the source language to a pivot language and then translate back to the source language. By doing this, we are able to obtain the translationese of source language.

Implementation of Experiment in Figure 1. TyDiQA (Clark et al., 2020) provides both training and testing datasets of originals for all languages. Here we adopt the originals training data to generate the corresponding translationese training data through back-translation [5]. The results in Figure 1 are obtained by training originals and generated translationese data for en, ar and fi, respectively. Note all the translationese is generated by the Google Cloud Translation[6] service, where the English translationese is generated by back-translationese with de as pivot language, the translationeses of ar and fi use en as pivot language. The The test set is originals of each language.

Implementation of Experiment in Figure 2. Similarly, we generate the translationese of the originals for each language using Google Cloud Translation service, where en is set as pivot language for non-English languages, and German for English language. We split the originals-translationese pairs of English into two groups, where 80% samples are used for training, and the rest 20% samples together with all pairs of other languages are used for evaluation. As the originals and translationese are paired, a random guess could achieve 50% accuracy for all languages ideally.

Implementation of TEA. It is worth noting that we can only access the originals of English and the translationese of other target languages during training. We use the translationese data, *i.e.*, the target language data translated from English, from

[5] We emphasize that non-English originals data is only utilized in Figure 1 and Figure 2 for analysis purpose. In addition, we only utilize the originals data of English in experiment, which follows the same settings as previous works, for translate-train.

[6] https://cloud.google.com/translate

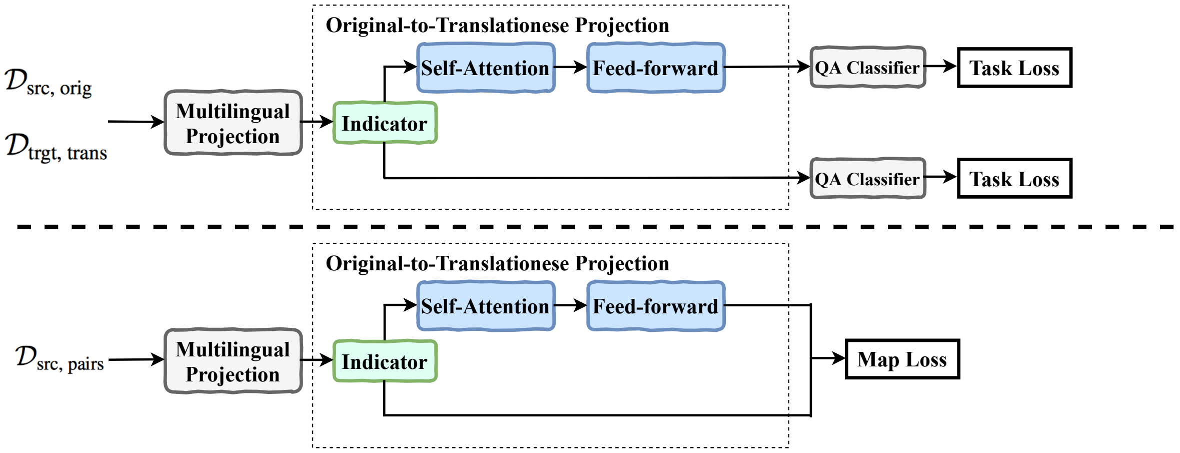

Original-to-Translationese Projection
$\mathcal{D}_{\text{src, orig}}$
$\mathcal{D}_{\text{trgt, trans}}$
Multilingual Projection
Indicator
Self-Attention
Feed-forward
QA Classifier
Task Loss
QA Classifier
Task Loss
Original-to-Translationese Projection
$\mathcal{D}_{\text{src, pairs}}$
Multilingual Projection
Indicator
Self-Attention
Feed-forward
Map Loss

Figure 3: Overall training pipeline of TEA. All modules are shared. Indicator is used to forward different kind of data into upper path or lower path, *i.e.*, originals data for upper path and translationese data for lower path. The task losses are standard cross entropy loss and the map loss is computed by cosine distance function.

XTREME official website[7], while the translated data from XTREME is utilized in translate-train for all previous works (Fang et al., 2021). Besides, we also augment translationese of English, which is generated by back-translation, in our TEA. Again, we resort to the Google Cloud Translation service to generate the translationese for all experiments in Section 4, where the German is set as pivot language by default.

C Data and Parameters

The sample sizes of the data sets in all 9 languages are equal, since they are all translated from English originals training data. The standard translate-train (STT) directly adopt the data samples of 9 languages for training. In addition to the data samples of 9 languages, we also incorporate the English translationese, leading to 11% more samples used compared to SST. Besides, our Original-to-Translationese Projection (OTP) module also introduce additional parameters compared to SST.

D Additional Experiments

Main Results. In this part, we replenish the TyDiQA results of two advanced multilingual language models, *i.e.*, VECO (Luo et al., 2021) and HICTL (Wei et al., 2020) in Table 4.

Originals-Translationese Pair Sample. In Figure 4, we list examples of originals-translationese pair in English used for TEA training.

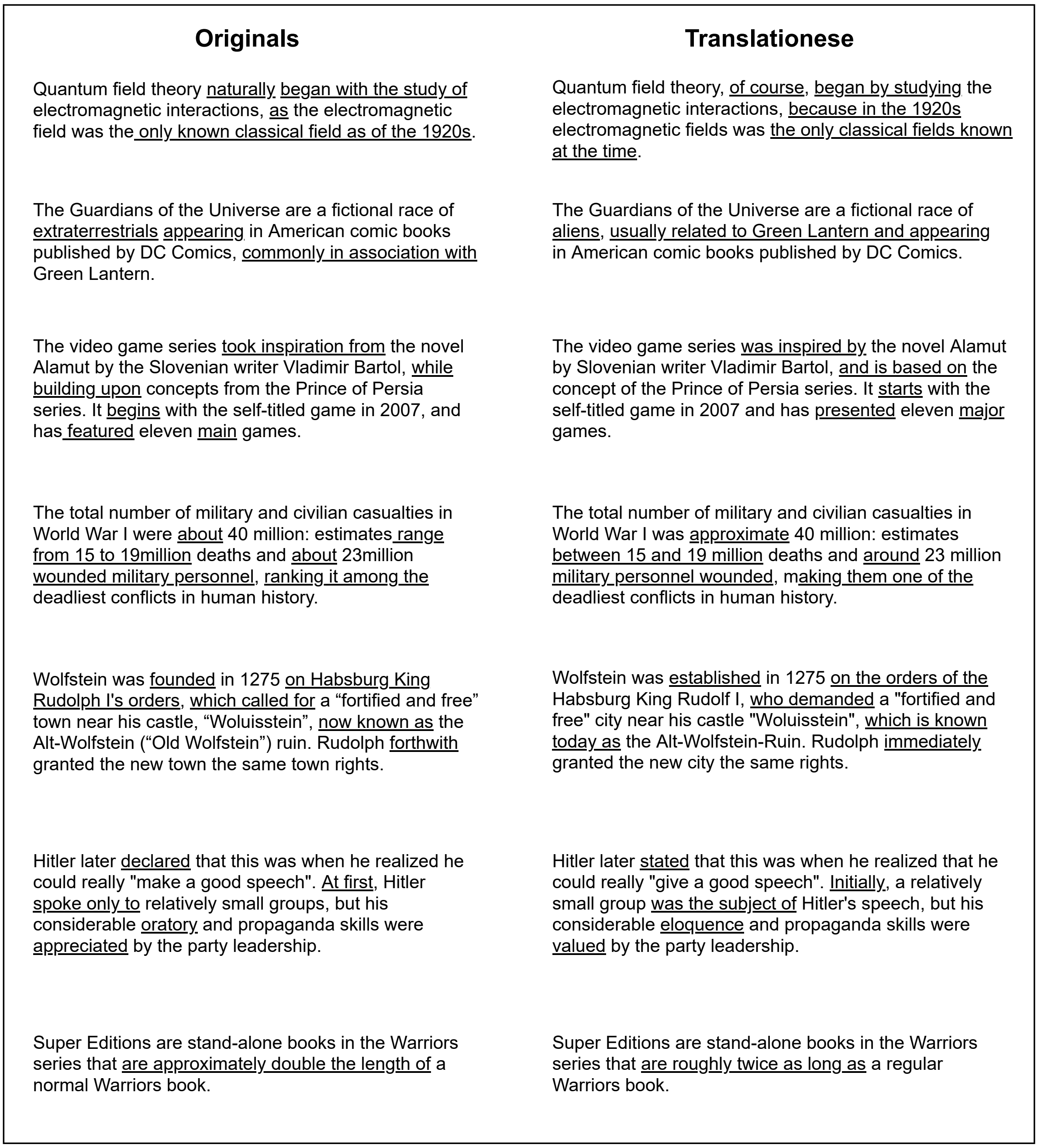
Effect of Translation Quality on Translationese English. Here we conduct an ablation study about the effect of translation quality on translationese English used in cosine distance loss. Due to the limited resource, we are unable to train a machine translation model from scratch by ourselves. Instead, we select the free Google Translate toolkit[8] (compared to the paid Google Cloud service) as the proxy of low-quality translator. We fix all the implementation settings and change the translationese English data only. Consequently, we obtain the average performance of $EM/F1$ = 58.0/73.9. The result indicates that a better translator is more effective for the translationese English generation. It is because that the low-quality translator may create more translation errors, then those errors are propagated during training, which hinders the learning of the originals to translationese mapping.

[7] https://console.cloud.google.com/storage/browser/xtreme_translations

[8] https://translate.google.com

LM	Method	ar	bn	fi	id	ko	ru	sw	te	avg
XLM-R	STT	40.4/67.6	47.8/64.0	53.2/70.5	61.9/77.4	10.9/31.9	42.1/67.0	48.1/66.1	43.6/70.1	43.5/64.3
	FILTER	50.8/72.8	56.6/70.5	57.2/73.3	59.8/76.8	12.3/33.1	46.6/68.9	65.7/77.4	50.4/69.9	49.9/67.8
	STT*	58.0/76.6	54.6/70.2	59.0/74.8	64.7/80.2	48.0/61.6	49.5/ 71.2	58.7/74.6	57.0/76.2	56.2/73.2
	TAG*	56.9/76.4	55.5/70.0	59.4/75.2	64.4/79.6	48.6/61.7	49.1/70.4	60.7/76.0	57.8/76.4	56.5/73.2
	TST.*	58.4/75.5	60.2/72.2	58.3/74.4	65.5/78.9	49.3/62.6	49.0/69.7	63.5/76.7	56.2/76.1	57.6/73.3
	GRL*	57.6/75.6	58.4/72.6	59.7/74.8	65.3/79.9	49.6/62.2	49.1/70.4	62.9/76.9	58.2/77.0	57.6/73.7
	TEA*	56.5/76.1	60.2/74.9	60.9/76.5	63.6/79.3	48.6/61.4	51.5/72.0	66.7/78.9	60.7/78.7	**58.6/74.7**
HICTL	STT	52.1/72.7	45.3/64.6	61.8/79.1	61.7/79.6	37.1/53.8	51.6/71.3	56.9/71.5	51.7/68.3	52.3/70.1
VECO	STT	57.5/77.0	56.6/72.2	59.3/76.6	64.4/80.0	52.2/63.4	50.5/72.8	67.1/79.4	58.0/76.0	58.2/74.7

Table 4: Main results on TyDiQA dataset. "LM": language models; "avg" denotes average performance across 8 languages; "∗": our implementation.



Originals	Translationese
Quantum field theory naturally began with the study of electromagnetic interactions, as the electromagnetic field was the only known classical field as of the 1920s.	Quantum field theory, of course, began by studying the electromagnetic interactions, because in the 1920s electromagnetic fields was the only classical fields known at the time.
The Guardians of the Universe are a fictional race of extraterrestrials appearing in American comic books published by DC Comics, commonly in association with Green Lantern.	The Guardians of the Universe are a fictional race of aliens, usually related to Green Lantern and appearing in American comic books published by DC Comics.
The video game series took inspiration from the novel Alamut by the Slovenian writer Vladimir Bartol, while building upon concepts from the Prince of Persia series. It begins with the self-titled game in 2007, and has featured eleven main games.	The video game series was inspired by the novel Alamut by Slovenian writer Vladimir Bartol, and is based on the concept of the Prince of Persia series. It starts with the self-titled game in 2007 and has presented eleven major games.
The total number of military and civilian casualties in World War I were about 40 million: estimates range from 15 to 19million deaths and about 23million wounded military personnel, ranking it among the deadliest conflicts in human history.	The total number of military and civilian casualties in World War I was approximate 40 million: estimates between 15 and 19 million deaths and around 23 million military personnel wounded, making them one of the deadliest conflicts in human history.
Wolfstein was founded in 1275 on Habsburg King Rudolph I's orders, which called for a "fortified and free" town near his castle, "Woluisstein", now known as the Alt-Wolfstein ("Old Wolfstein") ruin. Rudolph forthwith granted the new town the same town rights.	Wolfstein was established in 1275 on the orders of the Habsburg King Rudolf I, who demanded a "fortified and free" city near his castle "Woluisstein", which is known today as the Alt-Wolfstein-Ruin. Rudolph immediately granted the new city the same rights.
Hitler later declared that this was when he realized he could really "make a good speech". At first, Hitler spoke only to relatively small groups, but his considerable oratory and propaganda skills were appreciated by the party leadership.	Hitler later stated that this was when he realized that he could really "give a good speech". Initially, a relatively small group was the subject of Hitler's speech, but his considerable eloquence and propaganda skills were valued by the party leadership.
Super Editions are stand-alone books in the Warriors series that are approximately double the length of a normal Warriors book.	Super Editions are stand-alone books in the Warriors series that are roughly twice as long as a regular Warriors book.

Figure 4: Examples of originals-translationese pair in English from TyDiQA (Clark et al., 2020). The main differences are underlined.