Why Does Zero-Shot Cross-Lingual Generation Fail? An Explanation and a Solution

Tianjian Li[1] and **Kenton Murray**[1,2]
[1]Center for Language and Speech Processing
[2]Human Language Technology Center of Excellence
Johns Hopkins University
{tli104, kenton}@jhu.edu

Abstract

Zero-shot cross-lingual transfer is when a multilingual model is trained to perform a task in one language and then is applied to another language. Although the zero-shot cross-lingual transfer approach has achieved success in various classification tasks (Wu and Dredze, 2019), its performance on natural language generation tasks falls short in quality (Rönnqvist et al., 2019; Vu et al., 2022) and sometimes outputs an incorrect language (Xue et al., 2021). In our study, we show that the fine-tuning process learns language invariant representations, which is beneficial for classification tasks but harmful for generation tasks. Motivated by this, we propose a simple method to regularize the model from learning language invariant representations and a method to select model checkpoints without a development set in the target language, both resulting in better generation quality. Experiments on three semantically diverse generation tasks show that our method reduces the accidental translation problem by 68% and improves the ROUGE-L score (Lin, 2004) by 1.5 on average.

1 Introduction

Language Models (LMs) pre-trained on multilingual corpora (Devlin et al., 2019a; Conneau et al., 2020a; Liu et al., 2020; Xue et al., 2021) exhibit zero-shot cross-lingual transfer ability (Wu and Dredze, 2019). Given only annotated data in one language for a task, multilingual LMs are able to perform this task in languages seen only during the pre-training stage. The cross-lingual transferability of multilingual LMs reduces the need for annotated data in low-resource languages, which is valuable for building practical multilingual NLP systems.

Existing studies on cross-lingual transfer select tasks such as word alignment (Artetxe et al., 2020b), POS tagging (Pires et al., 2019), dependency parsing and sentence classification (Wu and Dredze, 2019) to investigate cross-lingual transferability of multilingual LMs (Hu et al., 2020), and few works focus on cross-lingual transfer in generation tasks (Maurya et al., 2021; Maurya and Desarkar, 2022). Cross-lingual transfer approach in generation tasks are known to produce incoherent text (Rönnqvist et al., 2021), generate in a wrong language (Xue et al., 2021), and suffer from catastrophic forgetting (Vu et al., 2022). Table 1 illustrates a common problem where the multilingual LM generates text in an incorrect language. Moreover, such a problem becomes more severe when under a true zero-shot setting (Zhao et al., 2021; Schmidt et al., 2022), where we do not have annotated data in the target language to guide model selection.

We show that the reason why zero-shot cross-lingual transfer in text generation fails is because the **fine-tuning process learns language invariant representations**, which is beneficial for classification tasks, but detrimental to generation tasks. In our paper, we use the cosine similarity between parallel sentence representations in different languages to measure the Cross-Lingual Representation Similarity (**XLRS**). We use a range of tasks from classification to extractive question answering, then to abstractive generation to show that in the best performing model, the XLRS after fine-tuning decreases as we move from classification to generation.

The fact that language invariant representations causes the degradation in generation tasks challenges the common belief that invariant representations generally enhance cross-lingual transfer on all downstream tasks (Cao et al., 2020; Conneau et al., 2020b; Yang et al., 2022; Xian et al., 2022). To the best of our knowledge, our work is the first to provide an analysis of how XLRS affects cross-lingual transfer in language generation tasks.

Motivated by our findings, we propose to use an auxiliary source language that implicitly regularizes the XLRS being too large and results in better generation performance over three complex natural

Findings of the Association for Computational Linguistics: ACL 2023, pages 12461–12476
July 9-14, 2023 ©2023 Association for Computational Linguistics

Prediction	Speak to your doctor, understand the dangers of alcohol consumption..
Target	与您的医生交谈：改变您对戒酒的想法。
	(Speak to your doctor: change your thinking towards quitting alcohol)
Prediction	Review accounting books periodically.
Target	기간을 결정한다. 회계장부를 모두 검토한다. 누락된 정보를 취합한다.
	(Determine the period. Review all accounting books. Gather missing information.)

Table 1: Example predictions of mT5 model fine-tuned on English WikiHow instructions and evaluated on Chinese and Korean input. The model outputs relevant text in an incorrect language.

language generation tasks (Summarization, Story Completion, and Title Generation). Under a true zero-shot setting, choosing the model checkpoint with the lowest XLRS results in an average of 4.1 point increase in ROUGE-L over using a source development set in two generation datasets.

To sum up, our contributions are threefold:

- We show that fine-tuning on a single source language increases the cosine similarity between sentence representations of different languages (XLRS).

- We show that the increase in XLRS causes degradation of cross-lingual transfer in generation tasks, and argue that the prevalent understanding of the benefit of similar representations does not apply to generation tasks.

- We empirically show that using two gold-annotated source languages instead of one regularizes the XLRS, resulting in an average increase of 1.5 in ROUGE-L.

2 Related Works

Multilingual Language Models. One line of work is to train multilingual versions of modern Language Models. **mBERT** (Devlin et al., 2019b) is the multilingual version of BERT (Devlin et al., 2019a), which uses the same encoder-only model architecture but is only trained on multilingual corpora. **XLM-R** (Conneau et al., 2020a) is the multilingual version of RoBERTa (Liu et al., 2019), which implements multiple optimization tricks and is larger in scale, resulting in better performance than BERT. **mBART** (Liu et al., 2020) is the multilingual version of BART (Lewis et al., 2020), an encoder-decoder model trained to reconstruct the original text through various types of artificially introduced noises. **mT5** (Xue et al., 2021) is the multilingual version of T5 (Raffel et al., 2020), an encoder-decoder model trained on a span denoising objective.

Cross-lingual Transfer. Multilingual models are able to be fine-tuned on annotated data of a task in only one source language and transfer the knowledge to other target languages to perform the same task without any supervision. While Pires et al. (2019) states that sub-word overlap between source and target facilitates cross-lingual transfer, K et al. (2020) shows that cross-lingual transfer manifests in pairs of source and target with zero sub-word overlap and word order is instead the most crucial ingredient. The performance of cross-lingual transfer between languages with a different order severely drops. Although the importance of word order is echoed by later studies (Artetxe et al., 2020b; Dufter and Schütze, 2020), recent studies have also debated in favor of the importance of matching script also contributing to cross-lingual transfer (Lauscher et al., 2020; Fujinuma et al., 2022). Wu et al. (2022) points out that the optimal set of parameters that generalizes well to all languages is a subset of parameters that achieves good performance on the source language. Therefore it is hard to find the optimal zero-shot cross-lingual transfer parameters by only optimizing source language performance. Chen and Ritter (2021) train a scoring model with the input features being the model's hidden representations and the output score being how well it generalizes to a given target language. However, previous studies focus on lower-level NLP tasks, which include text classification, dependency parsing, and extractive question answering (Hu et al., 2020) and rarely touch on language generation.

Another line of work focuses on applying cross-lingual transfer to a wide range of multilingual

NLP applications, which include sequence tagging (Yang et al., 2016), Named Entity Recognition (Xie et al., 2018), dependency parsing (Ahmad et al., 2019), sentence classification (Conneau et al., 2018; Yang et al., 2019), and information retrieval (Izacard et al., 2022). Empirical studies also train ranking models (Lin et al., 2019), use meta-learning (Nooralahzadeh et al., 2020), or use Shapley Value (Parvez and Chang, 2021) to predict which sources perform the best for a given target language.

Natural Language Generation. Multilingual LMs are prone to produce text that is repetitive (Xu et al., 2022), contains hallucinations (Raunak et al., 2021), or is in the wrong language (Zhang et al., 2020; Xue et al., 2021; Vu et al., 2022). Vu et al., 2022 proposed to use parameter efficient fine-tuning methods (Lester et al., 2021; Qin and Eisner, 2021; Li and Liang, 2021) to regularize the model to generate in a desired language. Other ways to improve generation quality include using back translation (Gu et al., 2019; Zhang et al., 2020), and transliteration (Sun et al., 2022) as data augmentation techniques, mixing in the pretrain objective during fine-tuning (Xue et al., 2021) and using an auxiliary source language in machine translation (Xu et al., 2021). Two concurrent efforts are close to our work: Xu and Murray (2022) and Schmidt et al. (2022) both empirically show that using multiple languages during fine-tuning in few-shot cross-lingual transfer improves performance in text classification. Our work differs in that we evaluated **text generation** under a **true zero-shot setting**, where we have access to neither a few examples to train on nor an annotated development set to guide model checkpoint selection.

3 Setup

The consensus of the literature (Cao et al., 2020; Conneau et al., 2020b; Tiyajamorn et al., 2021; Yang et al., 2022; Xian et al., 2022) is that if a model can produce similar representations for parallel sentences, the model would be able to achieve good cross-lingual transfer performance. Intuitively, if a model maps parallel sentences in English and French into nearly identical representations, and is able to predict the sentiment of the English sentence, it will also be able to predict the sentiment of the French sentence.

We hypothesize that the fine-tuning process increases the similarity between sentence representations of different languages. We use the following setups and tasks to verify our hypothesis.

3.1 Models and Datasets

Models We select the state-of-the-art multilingual language model: mT5-base (Xue et al., 2021). We use the Huggingface (Wolf et al., 2020) implementation. We use a uniform learning rate of 7e-5, a batch size of 32 for 10 epochs for all tasks described below.

Name	Task	Metric
UDPOS	Part-of-speech tagging	Acc.
PAWS-X	Paraphrase Identification	F1
TyDiQA	Question Answering	F1/EM
WikiLingua	Summarization	ROUGE

Table 2: Summary of tasks used in §5.

Datasets Table 2 describes the tasks we used in the following section to show the transition from classification to generation[1]. We use the **UDPOS** (Nivre et al., 2018) dataset containing sentences and the part-of-speech tag of each word. For sentence-level classification, we use the **PAWS-X** (Yang et al., 2019) dataset containing pairs of sentences and a binary tag on whether the second sentence entails the first sentence. For extractive generation, we use the **TyDiQA-GoldP** (Clark et al., 2020) dataset which contains paragraphs and questions whose answers are spans extracted from the paragraphs. For abstractive generation, we use the **WikiLingua** (Ladhak et al., 2020) dataset, which contains WikiHow instructions and their summaries in 17 languages.

We use the story completion (**SG**) and title generation (**TG**) task in the MTG benchmark (Chen et al., 2022), a recently introduced benchmark to evaluate multilingual text generation. We follow Vu et al., 2022, which uses the WikiLingua dataset to construct *WikiLingua-0*, where the model is only fine-tuned on English and evaluated on other languages. We extend *WikiLingua-0* and use languages besides English as the source and evaluate the zero-shot directions.

In all of our experiments, we report the results averaged across three runs with different random seeds. For each source language, we only use the

[1] We follow (Vu et al., 2022) and report the SP-ROUGE score.

top 10k training examples to train our model to ablate the effect of training data size on cross-lingual transfer. Unless specified otherwise, we evaluate under a **true zero-shot** setting, where we select the model checkpoint based on its performance on a dev set of the source language.

3.2 Sequence to Sequence Learning

We cast sequence labeling and sentence classification tasks into a text-to-text format using templates described in Table 10 in the Appendix. We follow Raman et al. (2022) and cast sequence labeling tasks into the sentinel + tag format. We follow Schick and Schütze (2021) and cast the sentence entailment task into a cloze question, supervising the model to predict the word "yes" for entailment and the word "no" for non-entailment.

4 Learning Dynamics of Cross-lingual Transfer

We plot the average cosine similarity between representations of parallel sentences (XLRS)[2] for each training iteration in two classification tasks: POS tagging and paraphrase identification (PAWS-X) at Figure 1 and Figure 2, respectively.

In both tasks, the plot displays an increasing trend of XLRS between parallel sentences between English and all the other languages. Notably, languages that have the same script have a higher similarity. Our findings show that the fine-tuning process on classification tasks does make the sentence representations of different languages more similar.

We then plotted the XLRS between representations of parallel sentences of a model when fine-tuned on WikiLingua: a summarization dataset in Figure 3. The average similarity gradually increases as we progress further into the training iterations, confirming our hypothesis that **fine-tuning on a single source language increases the XLRS between the source and other languages**.

Based on our findings, we conjecture that the model jointly minimizes two metrics, resulting in cross-lingual transfer:

- The Cross-Entropy loss between the predicted labels and the ground-truth labels, given an input in the source language (the standard training objective).

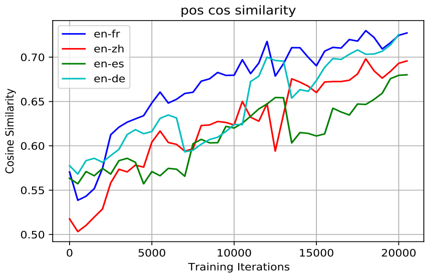

Figure 1: Average cosine similarity between parallel sentence representations (XLRS) in pretrained mT5-base model fine-tuned on English POS tagging data.

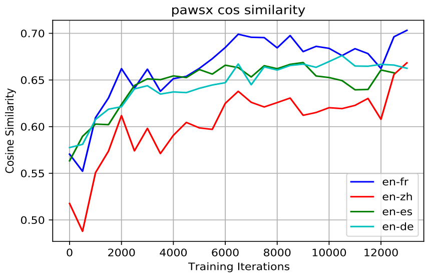

Figure 2: Average cosine similarity between parallel sentence representations (XLRS) in pretrained mT5-base model fine-tuned on English PAWS-X data.

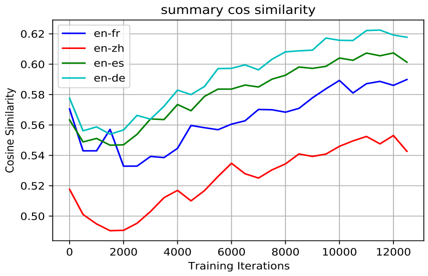

Figure 3: Average cosine similarity between parallel sentences representations (XLRS) of pretrained mT5-base model fine-tuned on English WikiLingua data.

- The distance between parallel sentences of the source and target languages (increase in XLRS).

And as a result, the cross-entropy loss between the predicted and ground-truth labels, given a context in the target language, is minimized, enabling cross-lingual transfer.

5 Unified View of Tasks

With the intuition of how the model does cross-lingual transfer in classification tasks, we note that language generation is a classification on a large *vocabulary set*, rather than a small label set. Thus, we point out that the reason why good performance

[2] We use the mean-pooled token encoder hidden states as the sentence representation. We randomly sample 512 sentences from the test set of the MTG story completion task.

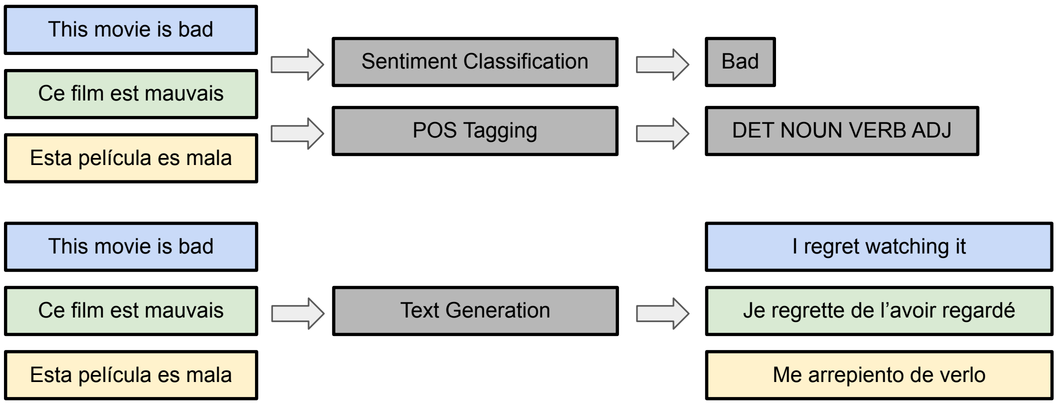

Figure 4: Illustration of the difference between classification tasks (top), where the model needs to map parallel sentences to the same label ($d = 100\%$), and generation tasks (bottom), where the model needs to map parallel sentences into different labels ($d = 0\%$). We use label overlap, d, to denote the fraction of parallel sentences that map to the same label.

with cross-lingual transfer on generation tasks is harder to achieve is actually caused by increasing XLRS.

Figure 4 illustrates our intuition. In classification tasks, the model needs to map parallel sentences to the same label. Ideally, the model produces identical representations for parallel sentences, resulting in the highest possible XLRS of 1. This is why model cross-lingual transfers better with a high XLRS. However, in generation tasks, if we view it as classifying over the entire vocabulary set, we are mapping parallel sentences to different labels. In an extreme case when XLRS is 1, the model fails to identify the source language, resulting in the common problem of the model producing an incorrect language (Xue et al., 2021). We introduce the notion of **label overlap**, d, to measure the percentage of examples in a dataset where the model needs to map parallel sentences into the same label.

We use $\mathcal{C}$ to denote the set of all discrete contexts and $\mathcal{C}_s$ to denote the set of all discrete contexts in language s. In classification tasks, the model learns to predict the ground-truth label $\hat{y} = l(c)$ over a set of candidate labels $\mathcal{Y}$ for context $c \in \mathcal{C}$. Similarly, in a simplified view of generation, the model learns to predict the next word v given context c. Therefore, we can essentially view language generation as classification where the label set is the entire vocabulary. In both cases, given context c, the model learns a probability distribution $p_{\cdot|c}$. The only difference is between their classification label set cardinality.

We define cross-lingual label overlap as an indicator of difficulty to cross-lingual transfer for a given task at §5.1. We then use a range of tasks: word-level classification (POS tagging §5.2) - Sentence level classification (Entailment classification §5.3) - Span Extraction (Extractive Generation §5.4) - Summarization (Abstractive generation §5.5) to show an increasing level of difficulty to perform cross-lingual transfer.

5.1 Cross-lingual Label Overlap

We denote a task's difficulty in transferring knowledge from one language to another by the percentage of overlap of their label set for parallel sentences. Given n parallel sentences $\{c_s^1, c_s^2, ..., c_s^n\}$ and $\{c_t^1, c_t^2, ..., c_t^n\}$ in source language s and target language t, the cross-lingual overlap d for task α is defined as:

$$d_\alpha(s,t) = \frac{\sum_{i=1}^{n} \mathbb{1}(l_\alpha(c_s^i) = l_\alpha(c_t^i))}{n}$$

In our analysis, we use English as the source language and evaluate the difficulty of performing cross-lingual transfer on other languages. The total label overlap for each task is the average label overlap for each target language.

$$d_\alpha = \sum_{j=1}^{m} d_\alpha(\text{english}, t_j)$$

Where t_j is the jth target language.

A higher d_α indicates an easier task to transfer knowledge from one language to another, whereas a lower d_α indicates a more difficult task to cross-lingual transfer.

5.2 POS tagging

The word-level label overlap for part-of-speech tagging should be close to 100%. With such a high percentage of label overlap, the model benefits from producing identical representations for parallel sentences to predict the same labels for different languages without supervision of the target language.

For example, if the model maps the English sentence *the dog ran* and the french sentence *le chien courir* into nearly identical representations and simultaneously learns a function to map the English words to their respective POS tags "DET NOUN VERB", the model would also be able to predict the correct label for french even without supervision. We denote the amount of "label overlap" as a metric defining the difficulty for a model to perform cross-lingual transfer on it.

5.3 Sentence Classification

The classification task discussed in this section (PAWS-X) includes sentiment classification of a single sentence and entailment classification between two sentences. For semantically equivalent parallel sentences, their sentiment or entailment labels are always the same. Therefore, $d = 100\%$.

Ideally, in sentence classification tasks, parallel sentences in different languages should map to the same probability distribution. For example, if the English sentence *I am happy* and the french sentence *Je suis content* maps to nearly identical representations and the model learns to predict the sentiment in English, the model would be able to cross-lingual transfer the ability to predict sentiment in English to French without any supervision.

5.4 Span Extraction

Span extraction requires a model to select a correct answer span from a passage given a question. Even though the data in TyDiQA is in different languages, and not parallel, 16% of the answer spans are pure numbers, and 50.6% of answer spans are mainly composed of numbers (Time and Dates).

This indicates that span extraction is a harder task than sentence classification, but with such a high amount of label overlap, the task is solvable through cross-lingual transfer.

5.5 Generation

The amount of label overlap in abstractive generation tasks (e.g. summarization, story completion, title generation) is close to zero as the model needs to predict words in completely different languages. The amount of label overlap for a subset of five languages (En, De, Fr, Es, Zh) of the WikiLingua (Ladhak et al., 2020) dataset is $d = 0.13\%$[3].

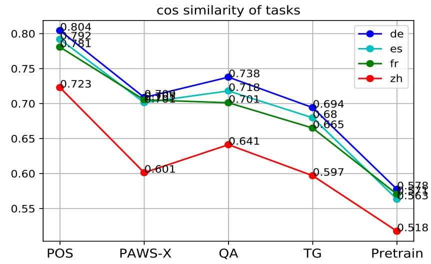

Figure 5: Average cosine similarity between representations of parallel sentences in English and 4 languages for the best performing model in 4 different tasks.

In a generation task, if the model maps the source and the target into identical representations, the model predicts the next word to be the same. Even if this is correct in semantics and possibly results in the code-switched results as shown in Figure 1, the model fails to generate in the correct language.

5.6 Analysis

We plotted the XLRS between English and four different languages in the best-performing English supervised models for the four tasks and the pretrained model at Figure 5.

The plot confirms our belief that for tasks (POS tagging, PAWS-X, TyDiQA) with large label overlap, the model cross-lingual transfers from increasing XLRS, whereas in generation tasks with label overlap close to zero (title generation), the best-performing model has a lower XLRS.

Following Yang et al. (2022), we calculate the Spearman's rank correlation score between (a) XLRS between English and 4 target languages (German, French, Chinese, Spanish), and (b) The averaged zero-shot cross-lingual transfer performance in each task. The results are reported at Table 4. In both classification tasks, XLRS positively correlates with cross-lingual performance. In contrast, in our three generation tasks, XLRS negatively correlates with cross-lingual performance[4],

[3] This small percentage represents mainly numbers and named entities (i.e. cities) that are the same across languages.

[4] We observed a stronger negative correlation between cosine similarity and ROUGE-2 in generation tasks but opted to report the ROUGE-L results to be consistent with our main results.

	AR	ZH	CS	NL	EN	FR	HI	ID	IT	JA	KO	PT	RU	ES	TH	TR	VI
EN*	17.4	15.1	17.8	20.1	39.6	22.4	9.1	23.0	20.3	14.6	17.3	23.8	15.3	23.3	17.9	17.5	21.9
EN	24.1	22.4	18.6	20.0	31.7	22.4	18.2	19.4	20.6	**21.0**	23.1	23.7	17.6	23.5	20.9	17.8	21.5
EN+ZH	24.3	27.5	20.4	22.6	33.2	23.8	18.8	21.6	22.1	18.4	21.2	**27.2**	20.2	24.9	**21.8**	18.2	24.3
DE	23.9	22.5	20.4	23.2	24.1	24.1	19.2	23.2	22.5	18.3	23.1	26.1	19.7	25.2	19.5	17.9	27.0
DE+ZH	**24.8**	27.9	21.2	**24.2**	**26.0**	**25.2**	19.5	**24.2**	**24.1**	20.3	**23.7**	26.9	**21.6**	**25.9**	**21.7**	19.1	27.1
EN+DE	**24.9**	**25.2**	**21.4**	24.0	22.3	25.3	**20.1**	23.7	22.8	19.3	24.1	**27.2**	20.3	**25.8**	20.0	**19.3**	**27.4**

Table 3: ROUGE-L results on the WikiLingua (Ladhak et al., 2020) dataset. Left: Source languages that we fine-tune on. Top: Target language that we evaluate on. The bolded numbers refer to the highest zero-shot performance. Note that some of the directions are not zero-shot. The amount of training instances used in each row is the same. * indicate results reported in Vu et al. (2022).

Task	POS	PAWS-X	TG	SG	WikiLingua
ρ	0.89*	0.91*	-0.37*	-0.39*	-0.33*

Table 4: Spearman's rank correlation ρ between the average cosine similarity between parallel sentences in source and 4 target languages (De, Es, Fr, Zh) and the average zero-shot cross-lingual transfer performance (F1 for POS tagging, Acc. for PAWS-X and ROUGE-L for generation) on two classification tasks and three generation tasks, * indicates that the p-value is less than 0.05.

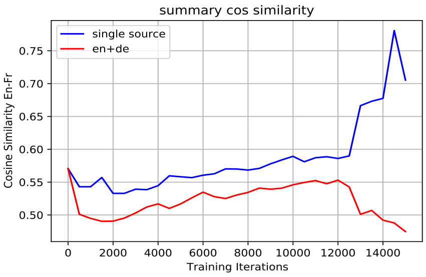

Figure 6: Average cosine similarity between representations of parallel sentences (XLRS) in English and French for model trained on one and two source languages.

indicating that **XLRS is strongly correlated to transfer performance in classification tasks but is detrimental to cross-lingual transfer in generation tasks.**

6 Text Generation Experiments

Now that we know XLRS is negatively correlated with cross-lingual transfer in text generation, since calculating XLRS during every iteration is computationally expensive, we wonder if we can **regularize XLRS implicitly**. Motivated by using auxiliary source languages improves machine translation (Xu et al., 2021) and few-shot cross-lingual transfer (Xu and Murray, 2022; Schmidt et al., 2022), we propose to use an additional source language to regularize XLRS.

To verify our hypothesis, we plot the XLRS between English and French during training on two different sources (En, De)in the story completion task, compared to training only on one source (En) in Figure 6. We observe that when the model is only given one language as the source, the XLRS increases, whereas using two source languages allows the model to learn to control the XLRS from being too high, resulting in fewer accidental translations and better quality.

To show that regularizing XLRS does result in better generation quality, we experiment with three semantically diverse generation tasks: Summarization (**WikiLingua**), Title Generation (**TG**), and Story Completion (**SG**).

6.1 Results

Table 3 shows the results of fine-tuning with multiple languages. We observe that adding Chinese to English data improves the performance in 13 out of 15 zero-shot directions[5] compared to only using English. We point out that our improvement is not due to an increase in the amount of training data since we used the same amount of training data for all experiments. We further observe that adding Chinese as an additional language to German also improves the performance in all 14 zero-shot directions, which often results in the best zero-shot performance.

Table 5 and 6 show the ROUGE-L results in the title generation (TG) and story completion (SG) in the MTG (Chen et al., 2022) benchmark, respectively. Again, we are able to observe that using two source languages almost always improves the ROUGE-L score. Notably, using two related

[5] The results when the target language are Chinese (ZH) or English (EN) is not zero-shot.

languages often results in degraded performance than using two unrelated languages with different scripts. We hypothesize that a language with a different script and order provides a more substantial regularization effect, preventing the cosine similarity between the source and target sentence representations from being too high.

	EN	ES	DE	FR	ZH	Avg.
EN	32.3	26.0	24.4	25.3	19.6	25.5
DE	**30.2**	24.7	22.5	23.9	18.7	24.0
ZH	25.1	24.5	21.4	23.7	26.0	24.1
DE+ZH	29.2	24.8	23.6	24.9	22.7	25.0
DE+EN	27.8	24.6	23.6	22.3	**20.8**	23.8
EN+ZH	33.3	**28.4**	**26.8**	**27.4**	22.4	**27.7**

Table 5: ROUGE-L results of title generation task in MTG benchmark. All experiments used same amount of data. The best zero-shot performance on each target language is bold.

To verify that our method helps against the accidental translation problem, we follow previous work (Vu et al., 2022) and calculate the language id confidence score on the source language and target language on the title generation task. The results are shown at Table 9 in Appendix A. Fine-tuning with multiple source languages helps the model learn which language it should produce.

6.2 Model Selection Using Parallel Sentences

Since XLRS negatively correlates with the performance of cross-lingual generation, we use it as a criterion for model selection in the absence of an annotated dev set. We report the performance on the WikiLingua dataset and the story completion task in MTG benchmark at Figure 7 and 8, when selecting the model using English dev set performance (**en-dev**), selecting the model with the lowest XLRS between English and the target language (**cos-sim**), and selecting the model using an annotated dev set on each target language (**tgt-dev**), which serves as an upper bound for true zero-shot cross-lingual transfer.

	EN	ES	DE	FR	ZH	Avg.
EN	29.1	28.9	27.8	28.9	20.3	27.0
DE	29.3	28.7	29.9	**32.0**	**20.4**	28.0
ZH	22.3	21.2	22.3	28.6	26.6	24.2
DE+ZH	**31.5**	**29.7**	28.6	31.8	22.5	28.8
DE+EN	30.4	27.3	26.4	28.2	19.8	26.4
EN+ZH	31.8	28.5	**29.3**	29.1	28.6	**29.5**

Table 6: ROUGE-L results of story completion task in MTG benchmark. All experiments used the same amount of data. The best zero-shot performance on each target language is bold.

In both tasks, selecting the model checkpoint with the lowest XLRS results in better performance than using an English development set on all target languages. The performance is on average less than one ROUGE-L point less on Spanish and German in both datasets, compared to using an annotated dev set. Our method results in an average increase of 5 ROUGE-L points in a distant language (Chinese).

	ES	DE	FR	ZH	Δ
en-dev	23.5	19.8	22.4	22.4	-3.23
cos-sim	25.3	21.2	24.6	26.5	-0.85
tgt-dev	**25.4**	**21.9**	**25.4**	**28.3**	0

Table 7: ROUGE-L results by selecting model based on English development set (**en-dev**), similarity of representations between English and target language (**cos-sim**) and using target language development set (**tgt-dev**) on WikiLingua (Ladhak et al., 2020).

	ES	DE	FR	ZH	Δ
en-dev	28.9	27.8	28.9	20.3	-4.88
cos-sim	30.8	30.3	**35.6**	26.4	-0.58
tgt-dev	**31.2**	**30.5**	**35.6**	**28.1**	0

Table 8: ROUGE-L results by selecting model checkpoints in Story Completion (SG) task in MTG benchmark (Chen et al., 2022).

7 Conclusion

We show that multilingual LMs transfer supervision from one language to another by increasing Cross-Lingual Representation Similarity (XLRS). Such a learning process results in decent zero-shot cross-lingual transfer performance in classification tasks but is harmful to text generation performance . We demonstrate that regularizing XLRS improves text generation quality and use parallel sentences to guide model selection without annotated data in the target languages. We believe that this is valuable under a practical setting (Artetxe et al., 2020c) where we have access to parallel data between the source and target languages, but not task-specific data in the target language.

Limitations

Our work sheds light on understanding the training dynamics of cross-lingual transfer learning of multilingual LMs. In our work, we selected to use English as the source of cross-lingual transfer following previous work (Vu et al., 2022). We acknowledge that using other languages as the source language can provide benefits depending on the task (Lin et al., 2019; Turc et al., 2021). Our work does not focus on choosing source language to maximize downstream performance but instead focuses on the difference between classification tasks and generation tasks in cross-lingual transfer.

Secondly, we acknowledge that some of the datasets (Yang et al., 2019; Chen et al., 2022) used in our work are created by machine translation and human annotation. Previous studies have pointed out that translationese in datasets affects cross-lingual transfer performance (Artetxe et al., 2020a; Artetxe et al., 2020c). We believe that translationese in datasets also have impact on XLRS. We leave the study of how dataset features (size, quality, translationese) affect cross-lingual transfer for future work.

Acknowledgements

We sincerely thank Haoran Xu, Kelly Marchisio and Daniel Khashabi for their helpful suggestions.

References

Wasi Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard Hovy, Kai-Wei Chang, and Nanyun Peng. 2019. On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing. 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 2440–2452, Minneapolis, Minnesota. Association for Computational Linguistics.

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

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

Mikel Artetxe, Sebastian Ruder, Dani Yogatama, Gorka Labaka, and Eneko Agirre. 2020c. A call for more rigor in unsupervised cross-lingual learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7375–7388, Online. Association for Computational Linguistics.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In *International Conference on Learning Representations*.

Yang Chen and Alan Ritter. 2021. Model selection for cross-lingual transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5675–5687, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yiran Chen, Zhenqiao Song, Xianze Wu, Danqing Wang, Jingjing Xu, Jiaze Chen, Hao Zhou, and Lei Li. 2022. MTG: A benchmark suite for multilingual text generation. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2508–2527, Seattle, United States. Association for Computational Linguistics.

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. In *Transactions of the Association of Computational Linguistics*.

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

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.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. 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, Minneapolis, Minnesota. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. mbert blog post. `https://github.com/google-research/bert/blob/master/multilingual.md`. Accessed: 2022-11-05.

Philipp Dufter and Hinrich Schütze. 2020. Identifying elements essential for BERT's multilinguality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4423–4437, Online. Association for Computational Linguistics.

Yoshinari Fujinuma, Jordan Boyd-Graber, and Katharina Kann. 2022. Match the script, adapt if multilingual: Analyzing the effect of multilingual pretraining on cross-lingual transferability. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1500–1512, Dublin, Ireland. Association for Computational Linguistics.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor O.K. Li. 2019. Improved zero-shot neural machine translation via ignoring spurious correlations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1258–1268, Florence, Italy. 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 generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048, Online. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

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*.

Kaushal Maurya and Maunendra Desarkar. 2022. Meta-X$_{NLG}$: A meta-learning approach based on language clustering for zero-shot cross-lingual transfer and generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 269–284, Dublin, Ireland. Association for Computational Linguistics.

Kaushal Kumar Maurya, Maunendra Sankar Desarkar, Yoshinobu Kano, and Kumari Deepshikha. 2021. ZmBART: An unsupervised cross-lingual transfer framework for language generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2804–2818, Online. Association for Computational Linguistics.

Joakim Nivre, Mitchell Abrams, Željko Agić, Lars Ahrenberg, Lene Antonsen, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, Mohammed Attia, et al. 2018. Universal dependencies 2.2.

Farhad Nooralahzadeh, Giannis Bekoulis, Johannes Bjerva, and Isabelle Augenstein. 2020. Zero-shot cross-lingual transfer with meta learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4547–4562, Online. Association for Computational Linguistics.

Md Rizwan Parvez and Kai-Wei Chang. 2021. Evaluating the values of sources in transfer learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5084–5116, Online. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Guanghui Qin and Jason Eisner. 2021. Learning how to ask: Querying LMs with mixtures of soft prompts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5203–5212, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Karthik Raman, Iftekhar Naim, Jiecao Chen, Kazuma Hashimoto, Kiran Yalasangi, and Krishna Srinivasan. 2022. Transforming sequence tagging into a seq2seq task. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183, Online. Association for Computational Linguistics.

Samuel Rönnqvist, Jenna Kanerva, Tapio Salakoski, and Filip Ginter. 2019. Is multilingual BERT fluent in language generation? In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, pages 29–36, Turku, Finland. Linköping University Electronic Press.

Samuel Rönnqvist, Valtteri Skantsi, Miika Oinonen, and Veronika Laippala. 2021. Multilingual and zero-shot is closing in on monolingual web register classification. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 157–165, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Fabian David Schmidt, Ivan Vulić, and Goran Glavaš. 2022. Don't stop fine-tuning: On training regimes for few-shot cross-lingual transfer with multilingual language models. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Simeng Sun, Angela Fan, James Cross, Vishrav Chaudhary, Chau Tran, Philipp Koehn, and Francisco Guzmán. 2022. Alternative input signals ease transfer in multilingual machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5291–5305, Dublin, Ireland. Association for Computational Linguistics.

Nattapong Tiyajamorn, Tomoyuki Kajiwara, Yuki Arase, and Makoto Onizuka. 2021. Language-agnostic representation from multilingual sentence encoders for cross-lingual similarity estimation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7764–7774, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Iulia Turc, Kenton Lee, Jacob Eisenstein, Ming-Wei Chang, and Kristina Toutanova. 2021. Revisiting the primacy of english in zero-shot cross-lingual transfer.

Tu Vu, Aditya Barua, Brian Lester, Daniel Cer, Mohit Iyyer, and Noah Constant. 2022. Overcoming catastrophic forgetting in zero-shot cross-lingual generation. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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, Hong Kong, China. Association for Computational Linguistics.

Shijie Wu, Benjamin Van Durme, and Mark Dredze. 2022. Zero-shot cross-lingual transfer is underspecified optimization. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 236–248, Dublin, Ireland. Association for Computational Linguistics.

Ruicheng Xian, Heng Ji, and Han Zhao. 2022. Cross-lingual transfer with class-weighted language-invariant representations. In *International Conference on Learning Representations*.

Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A. Smith, and Jaime Carbonell. 2018. Neural cross-lingual named entity recognition with minimal resources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 369–379, Brussels, Belgium. Association for Computational Linguistics.

Haoran Xu and Kenton Murray. 2022. Por qué não utiliser alla språk? mixed training with gradient optimization in few-shot cross-lingual transfer. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2043–2059, Seattle, United States. Association for Computational Linguistics.

Jin Xu, Xiaojiang Liu, Jianhao Yan, Deng Cai, Huayang Li, and Jian Li. 2022. Learning to break the loop: Analyzing and mitigating repetitions for neural text generation. In *Advances in Neural Information Processing Systems*.

Weijia Xu, Yuwei Yin, Shuming Ma, Dongdong Zhang, and Haoyang Huang. 2021. Improving multilingual neural machine translation with auxiliary source languages. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3029–3041, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Huiyun Yang, Huadong Chen, Hao Zhou, and Lei Li. 2022. Enhancing cross-lingual transfer by manifold mixup. In *International Conference on Learning Representations*.

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.

Zhilin Yang, Ruslan Salakhutdinov, and William W. Cohen. 2016. Multi-task cross-lingual sequence tagging from scratch. *CoRR*, abs/1603.06270.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Mengjie Zhao, Yi Zhu, Ehsan Shareghi, Ivan Vulić, Roi Reichart, Anna Korhonen, and Hinrich Schütze. 2021. A closer look at few-shot crosslingual transfer: The choice of shots matters. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5751–5767, Online. Association for Computational Linguistics.

A Language Identification Scores

	FR		ES	
	LID_{DE}	LID_{FR}	LID_{DE}	LID_{ES}
DE	67.7	20.5	73.6	18.8
DE+ZH	9.8	88.2	11.2	86.4
DE+EN	15.2	76.2	14.9	78.4

Table 9: Language identification confidence scores on the title generation task fine-tuned on single and multiple source languages.

Task	Template
Seq Tagging (UDPOS)	Input: <extra_id_0>In <extra_id_2>my <extra_id_3>view <extra_id_4>it <extra_id_5>is <extra_id_6>significant Output: <extra_id_0>ADP <extra_id_2>PRON <extra_id_3>NOUN <extra_id_4>PRON <extra_id_5>AUX <extra_id_6>ADJ
Classification (PAWS-X)	Input: The original version was skipped in favor of the mild edition. <extra_id_0>The mild version was skipped in favor of the original version. Output: <extra_id_0>No.
QA (TyDiQA)	Input: What is the surface area of the human cortex? <extra_id_0> Output: <extra_id_0>1.3 square feet
Generation (ByteCup)	Input: story: {News article on Philadelphia Flower Show} title: <extra_id_0> Output: <extra_id_0>philly flower show will treat visitors to sights, sounds and scents of rainforest

Table 10: Templates for casting tasks into a text-to-text format.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Yes, in the section "Limitations"

☒ A2. Did you discuss any potential risks of your work?
We believe that our study is an Engineering study and is not applicable to this question.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Yes. Under the "Abstract" and "Introduction"

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 3.1 Models and Datasets

☑ B1. Did you cite the creators of artifacts you used?
Section 3.1 Models and Datasets

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
The datasets we used are very common open-sourced datasets. This information can be found in the citations we provided.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 3.1

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
The datasets we used are very common open-sourced datasets.

☒ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No. Since the datasets we used are very popular open-sourced datasets, we believe that

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 3.1

C ☑ Did you run computational experiments?

Yes. In Section 6

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
We did report the number of parameters. But we did not discuss the computational budget because the experiment we do are fine-tuning, which should be fairly lightweight.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 3.1

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
We stated that we report the mean result averaged across three runs. But we did not provide error bars.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 3.1

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.