Enhancing Few-shot Cross-lingual Transfer with Target Language Peculiar Examples

Hwichan Kim and **Mamoru Komachi***
Tokyo Metropolitan University
6-6 Asahigaoka, Hino, Tokyo 191-0065, Japan
kim-hwichan@ed.tmu.ac.jp

Abstract

Few-shot cross-lingual transfer, fine-tuning Multilingual Masked Language Model (MMLM) with source language labeled data and a small amount of target language labeled data, provides excellent performance in the target language. However, if no labeled data in the target language are available, they need to be created through human annotations. In this study, we devise a metric to select annotation candidates from an unlabeled data pool that efficiently enhance accuracy for few-shot cross-lingual transfer. It is known that training a model with hard examples is important to improve the model's performance. Therefore, we first identify examples that MMLM cannot solve in a zero-shot cross-lingual transfer setting and demonstrate that it is hard to predict *peculiar* examples in the target language, i.e., the examples distant from the source language examples in cross-lingual semantic space of the MMLM. We then choose high *peculiarity* examples as annotation candidates and perform few-shot cross-lingual transfer. In comprehensive experiments with 20 languages and 6 tasks, we demonstrate that the high *peculiarity* examples improve the target language accuracy compared to other candidate selection methods proposed in previous studies. The code used in our experiments is available at https://github.com/hwichan0720/fewshot_transfer_with_peculiarity.

1 Introduction

Sufficient labeled data is essential to train an accurate model. However, few languages have abundant language resources for both labeled and unlabeled data like English (Joshi et al., 2020). In addition, constructing a large amount of labeled data through human annotators is costly and time-consuming. The use of Multilingual Mask Language Models (MMLMs) is one way to overcome this problem as they show good zero-shot cross-lingual performance in target languages by fine-tuning with only task-specific labeled data of source language, such as English. While this zero-shot cross-lingual transfer ability is promising for the target languages with no or limited task-specific resources, there is a divergence in accuracy between the source and target languages, meaning the zero-shot cross-lingual transfer ability is imperfect.

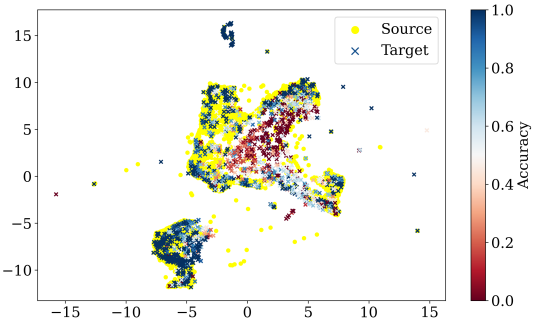


Figure 1: UMAP (McInnes et al., 2018) visualization of hidden states of source (English) and target (Arabic) language examples (NER task). We indicate the source language examples in yellow and target language examples in blue to red based on their accuracy in the zero-shot cross-lingual transfer setting. We use BOS hidden states of the last layer from XLM-R, which is an MMLM.

To analyze the characteristics of hard examples for predictions in zero-shot cross-lingual transfer, we visualize representations of the source and target languages' examples and highlight the target ones with accuracy in the zero-shot setting, as shown in Figure 1. We can observe that the accuracies of target language examples distant from the source language examples are low compared to others. We refer to these as *peculiar* examples of the target language. We should address *peculiar* examples to further enhance performance in the target language.

Few-shot cross-lingual transfer, adapting

*Now at Hitotsubashi University

Findings of the Association for Computational Linguistics: ACL 2023, pages 747–767
July 9-14, 2023

MMLM with a small number (0.1k–1k) of task-specific labeled examples in the target language, is a promising approach to enhance performance for target languages. Specifically, Lauscher et al. (2020) showed that a small number of examples randomly selected from labeled dataset significantly improves the accuracy for the target languages. However, if there are no labeled examples at all, we should create them by selecting annotation candidates from an unlabeled data pool. From this perspective, Kumar et al. (2022) evaluated candidate selection methods proposed in active learning research for selecting the annotation candidates for few-shot cross-lingual transfer. They split the target languages based on their zero-shot cross-lingual transfer performance into "good" and "poor" (or "good", "fair", and "poor") language groups[1] and demonstrated that effective methods vary for each language group. Our preliminary analysis (Figure 1) suggests that adapting MMLM for *peculiar* examples is crucial to improve performance for the target language and recommends selecting *peculiar* examples as annotation candidates.

Therefore, in this study, we first propose a metric to measure *peculiarity* of the target language examples. Note that *peculiarity* is defined without labels of the downstream tasks. Then, we select high *peculiarity* examples as annotation candidates and conduct few-shot cross-lingual transfer using languages with "good" and "poor" zero-shot cross-lingual performance. Our experiments show that the proposed metric *peculiarity* is very simple yet effective in selecting candidates for few-shot cross-lingual transfer. Our contributions in this study are threefold:

1. We propose a simple metric to measure *peculiarity* and show that the prediction accuracy of high *peculiarity* examples is low compared to others in the zero-shot setting.

2. We conduct few-shot cross-lingual transfer using high *peculiarity* examples and demonstrate that these examples can improve accuracy compared to other candidate selection methods regardless of language groups in few (2–3) label classification tasks. In addition, our analysis shows that peculiarity is robust for hyperparameters and brings consistent performance in few-shot cross-lingual transfer.

3. However, high *peculiarity* examples do not work well for the "poor" language group in many-label (5) classification and sequence-tagging tasks. Our analysis indicates that these examples are redundant to fine-tuning MMLM. Therefore, we design a method that combines existing methods to select diverse examples. Our experiments demonstrate that our method enhances accuracy across target languages.

[1]They referred the language groups as "C1" and "C2" (or "C1", "C2", and "C3"), respectively.

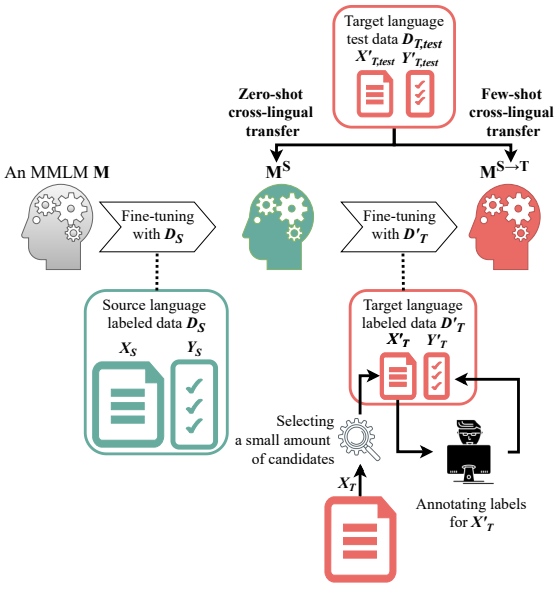


Figure 2: Overview of zero-shot and few-shot cross-lingual transfer.

2 Notation and Task Setting

In this section, we define notations and explain our task setting. We denote the source and target language as S and T, respectively. For the source language S, we assume that labeled data exist for downstream tasks $D_S = (X_S, Y_S)$, where $X_S = \{x_S^1, ..., x_S^i\}$ are monolingual data and $Y_S = \{y_S^1, ..., y_S^i\}$ are corresponding labels. For the target language T, we only have monolingual data $X_T = \{x_T^1, ..., x_T^j\}$. We denote an MMLM as $\mathbf{M}$ and one fine-tuned by D_S as $\mathbf{M^S}$. We use $\mathbf{M^S}$ to the target language inputs in the zero-shot cross-lingual transfer setting.

In this study, we conduct few-shot cross-lingual transfer. In this setting, we select annotation candidates $X'_T \subset X_T$ and limit $|X'_T|$ to n. Then, human

annotators annotate labels Y_T' for X_T'[2]. We additionally fine-tune $\mathbf{M^S}$ using $D_T' = (X_T', Y_T')$ and denote the model as $\mathbf{M^{S \to T}}$. To support understanding of zero-shot and few-shot cross-lingual transfer, we provide an overview of zero-shot and few-shot cross-lingual transfer in Figure 2. The objective of our task is to select the candidates X_T' that lead to better performance of $\mathbf{M^{S \to T}}$ in the target language T.

3 Related Works

Zero-shot cross-lingual transfer. mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) are the representative MMLMs. mBERT is an extension of BERT that is pre-trained on Wikipedia data in various languages. XLM-R is trained on 2.5T data of more than 100 languages extracted from Common Crawl (Wenzek et al., 2020). The most interesting property of these MMLMs is that they show a strong zero-shot cross-lingual transfer performance even though they did not explicitly use bilingual data.

Several studies have analyzed their zero-shot cross-lingual transfer ability and indicated that source-target languages' similarity is important for the transfer. Pires et al. (2019) analyzed the ability of mBERT in NER and POS tagging tasks and demonstrated that more overlap in WALS (Dryer and Haspelmath, 2013) features between the source and target languages, better transfer. Lauscher et al. (2020) quantitatively measured the similarity between the source and target languages using LANG2VEC (Littell et al., 2017) and showed that it has a strong correlation with the zero-shot cross-lingual transfer performance.

Yang et al. (2022) analyzed relations of alignment quality between the source and target languages and the transfer performance. Specifically, they measured the alignment quality between languages in cross-lingual semantic space of an MMLM using CKA (Kornblith et al., 2019) and showed a strong correlation between the CKA scores and the zero-shot cross-lingual transfer performances. They also proposed a method to improve the alignment quality using pseudo-bilingual data and demonstrated that it enhances the zero-shot cross-lingual transfer performance. In addition, there have been several attempts to enhance the zero-shot cross-lingual transfer ability using additional bilingual resources (Lample and Conneau, 2019; Cao et al., 2020; Chi et al., 2021; Dou and Neubig, 2021; Yang et al., 2021).

Few-shot cross-lingual transfer. Few-shot cross-lingual transfer is another approach to improve the performance for target languages. Lauscher et al. (2020) randomly selected annotation candidates X_T' from X_T and conducted few-shot cross-lingual transfer. They demonstrated that these candidates surprisingly boost performance compared to the zero-shot cross-lingual transfer setting (e.g., gains are 27.3 points in POS tagging task).

The most relevant previous work is Kumar et al. (2022) who conducted few-shot cross-lingual transfer and evaluated representative candidate selection methods used in active learning research, such as entropy and BADGE (Ash et al., 2020). BADGE is a method to select diverse and uncertain examples. It first calculates gradient embeddings (GEs) for each example $x_T \in X_T$, which are vectors of hidden states multiplied by $\mathbf{M^S}$'s confidences about each example, and then selects the most typical examples in the GEs space. However, BADGE is expensive for sequence-tagging tasks takes because it has to calculate the GE for each token. Therefore, Kumar et al. (2022) simplified the GEs to loss embeddings (LEs), which are vectors consisting of cross-entropy loss for each token, by taking $\mathbf{M^S}$'s prediction as actual labels. Their experiments showed that the method using LEs obtains consistent gains over other methods for sequence-tagging tasks.

The methods using embeddings, such as GEs and LEs, can select diverse candidates from X_T. However, we consider that these methods also select the candidates well aligned with the source language examples X_S, which are predicted accurately in the zero-shot cross-lingual transfer setting (Figure 1). Therefore, in this study, we propose a metric, *peculiarity*, which isolates examples that cannot be covered by source language data X_S, and select the candidates based on our metric.

4 How to Measure Language Peculiarity

Sorscher et al. (2022) revealed that training with hard examples for a neural model can improve the model performance exponentially beyond power law scaling, both in theory and practice. In terms of few-shot cross-lingual transfer, we need to fine-tune $\mathbf{M^S}$ using target language examples that will not be predicted correctly. Therefore, we study

[2]In our experiments, since we employ previously annotated labels, we do not annotate the candidates ourselves.

how to extract these examples as the annotation candidates from X_T without knowing their labels.

Yang et al. (2022) showed that there is a correlation between cross-lingual transfer performance and language-level alignment quality. Based on the preliminary analysis (Figure 1), we infer that there are correlations not only at the language-level but also at the example-level, which means that the accuracy of the target language examples depends on whether they are aligned with the source language examples. This analysis indicates that we can detect hard examples for $\mathbf{M^S}$ from X_T if we can measure the example-level alignment quality.

Therefore, we devise a simple but effective metric to isolate the examples based on their alignment quality without additional language resources. Intuitively, the examples that have no source language examples in the neighborhood in cross-lingual semantic space are not well aligned[3]. In addition, Figure 1 shows that accuracies of these examples are low compared to the others. We use this aspect instead of the example-level alignment quality as a proxy to detect the hard examples for $\mathbf{M^S}$. We refer to this aspect of whether there are source language examples in the neighborhood as *peculiarity* and define *peculiarity* of $x_T^j \in X_T$ as:

$$\begin{aligned}&\text{Peculiarity}(x_T^j; \mathbf{M^S}) \\ &= \frac{1}{k} \sum_{x_S^i \in k\text{-NN}(x_T^j)} distance(\boldsymbol{x_T^j}, \boldsymbol{x_S^i}) \end{aligned} \quad (1)$$

where $distance$ serves as a metric for measuring the distance between input hidden states, encompassing options such as Cosine and Euclidean distance. $x_S^i \in X_S$ is a source language example of k nearest neighbor based on $distance$ and k is a hyperparameter. $\boldsymbol{x_T^i}$ and $\boldsymbol{x_S^j}$ are hidden states of BOS tokens from the final layer of $\mathbf{M^S}$. *Peculiarity* is higher when there are no source language examples in the neighborhood of the input target language example. In experiments, we first confirm whether the *peculiarity* can isolate the examples with low or high prediction accuracy in the zero-shot setting. Then, we select high *peculiarity* examples as annotation candidates and conduct few-shot cross-lingual transfer.

[3]To strictly measure the example-level alignment quality for each $x_T \in X_T$, we should translate x_T into the source language and calculate similarities between their representations. However, this method is costly because it requires a human translator or a high-quality machine translation system.

	All	Lowest	Highest
XNLI	72.4	73.5	64.7
PAWS-X	81.4	99.6	67.0
MARC-2	89.9	99.4	53.2
MARC-5	54.2	70.6	50.8
NER	76.9	89.3	50.3
POS	74.4	79.0	69.5
Average	75.0	85.1	59.2

Table 1: Accuracy of each task in the zero-shot setting. "Lowest" and "Highest" are the scores of validation data, and "Lowest" and "Highest" are the scores of 10% examples of the lowest and highest peculiarity. The scores are the average for all languages, except for English (source language).

5 Experimental Settings

We explain experimental settings, which are largely similar to those used in previous works such as Lauscher et al. (2020) and Kumar et al. (2022), including the dataset and training hyperparameters.

5.1 Tasks and datasets

We experiment with XNLI (Conneau et al., 2018) for a classification task, and NER, POS-tagging for sequence tagging tasks. We use WikiANN (Rahimi et al., 2019) and Universal Dependency tree banks (UD, Nivre et al. (2016)) for NER and POS-tagging, respectively. In addition, we use PAWS-X (Yang et al., 2019) and Multilingual Amazon Review Classification (MARC, Keung et al. (2020)) as datasets for classification tasks. There are two- and five-label classification settings in MARC, and we refer to them as MARC-2 and MARC-5, respectively.

WikiANN, UD, and XNLI include several languages. We use the same languages used in Kumar et al. (2022)'s experiments. For PAWS-X and MARC, we use all languages. The previous studies (Lauscher et al., 2020; Kumar et al., 2022) indicated that the effectiveness of candidate selection methods in each language depends on the accuracy of downstream tasks in the zero-shot setting. Therefore, we evaluate each method by splitting the target languages into "good' and "poor" groups, where the target languages achieve the median or higher accuracy and lower than the median accuracy, respectively. We detail the languages of each group and the datasets in Appendix B.

		100-shot			500-shot			1,000-shot		
		All	Good	Poor	All	Good	Poor	All	Good	Poor
XNLI	Random	0.97	0.79	1.40	1.16	0.75	1.70	1.24	1.11	1.62
	Entropy	1.02	0.79	1.53	1.17	0.85	1.64	1.41	1.17	1.87
	KM	1.23	0.86	1.83	0.99	0.69	1.48	1.27	1.06	1.60
	GE-KM	0.94	0.79	1.24	0.89	0.51	1.32	1.39	0.94	1.92
	peculiarity	**1.31**	**1.04**	**1.84**	**1.25**	**0.94**	**1.75**	**1.52**	**1.35**	**1.95**
PAWS-X	Random	2.99	2.61	2.97	3.90	2.86	4.32	3.47	2.33	4.09
	Entropy	3.87	3.00	4.22	3.95	3.16	4.14	2.92	2.73	2.91
	KM	3.69	3.01	3.97	3.78	2.88	4.04	3.96	3.11	4.16
	GE-KM	3.83	2.85	4.24	3.66	2.95	3.79	3.67	3.20	3.72
	peculiarity	**4.01**	**3.45**	**4.28**	**4.23**	**3.45**	**4.52**	**4.18**	**3.45**	**4.36**
MARC-2	Random	0.54	0.17	1.10	0.80	0.45	1.32	1.09	0.77	1.57
	Entropy	0.96	0.45	**1.72**	0.68	0.43	1.05	1.15	0.87	1.57
	KM	0.71	0.25	1.35	0.79	0.42	1.35	1.16	0.77	1.75
	GE-KM	0.72	0.30	1.35	0.75	0.50	1.12	1.11	0.78	1.60
	peculiarity	**0.99**	**0.55**	1.10	**1.11**	**0.78**	**1.60**	**1.42**	**1.18**	**1.77**

Table 2: Evaluation on few-label classification tasks. These scores are differences in accuracy between 0-shot and each n-shot model, averaged across languages in each group. They are the average of three models. We indicate the best improvement scores in bold.

5.2 Model and training

We use XLM-R *Base*[4] as the MMLM in all experiments[5]. Following Devlin et al. (2019); Pires et al. (2019), we attach token-level and sentence-level classifiers to the last layer of XLM-R to train sequence-tagging and classification models, respectively.

We use English as the source language and fine-tune XLM-R with English dataset of each task. We limit the sequence length to 128 subword tokens and set the batch size as 32. For the sequence-tagging tasks, we fix the number of training epochs to 20 and the learning rate as $2 \cdot 10^{-5}$. For the classification tasks, we set the training epochs to 3 and learning rate as $3 \cdot 10^{-5}$.

For few-shot cross-lingual transfer, we conduct additional fine-tuning with combined English data and sampled target language examples. We change the training epochs to 1 for the classification tasks and use the same hyperparameters mentioned above.

[4] https://huggingface.co/xlm-roberta-base

[5] In the previous studies (Lauscher et al., 2020; Kumar et al., 2022), mBERT *Base cased* and XLM-R *Base* were used in the experiments, with XLM-R achieving better cross-lingual performances. In addition, they showed these MMLMs performing with the same trend.

5.3 Candidate selection methods

We select n annotation candidates, and set n to 100, 500, and 1,000. We compare our method with the methods (Random, Entorpy, GE-KM, and LE-KM) used in Kumar et al. (2022)[6] and an additional method (KM).

Random. We select the candidates randomly.

Entropy. We select the candidates with the highest entropy. We average the entropy per each token for the sequence-tagging tasks.

KM. We cluster `BOS` hidden states of the last layer using k-means++ (k is the same number of n) and select medoids, most typical candidates, from each cluster following Chang et al. (2021); Hacohen et al. (2022). We use scikit-learn[7] for performing k-means++.

GE-KM or LE-KM. We use the gradient embeddings (GEs) for the classification tasks and the loss embeddings (LEs) for the sequence tagging tasks instead of the hidden states and apply the same

[6] They also employed a method named Data Cross-Entropy (DCE). However, their experiments showed DCE is not effective for few-shot cross-lingual transfer compared to the other methods. Therefore, we exclude DCE in our experiments.

[7] https://scikit-learn.org/stable/

		1,000-shot		
		All	Good	Poor
MARC-5	Random	1.74	1.26	2.45
	Entropy	0.88	1.00	0.70
	KM	1.81	1.35	**2.50**
	LE-KM	1.61	1.16	2.27
	peculiarity	**1.83**	**1.71**	2.00
NER	Random	12.27	4.79	21.16
	Entropy	11.43	4.75	19.36
	KM	**12.39**	4.88	**21.35**
	LE-KM	12.38	4.86	21.30
	peculiarity	11.84	**5.11**	20.22
POS	Random	16.84	6.28	25.72
	Entropy	16.30	5.88	25.16
	KM	16.86	6.23	**25.93**
	LE-KM	16.94	6.24	25.92
	peculiarity	**17.08**	**6.66**	25.87

Table 3: Evaluation on many-label classification and sequence-tagging tasks.

		1,000-shot	
		Good	Poor
XNLI	KM	0.10	0.08
	peculiarity	0.09	0.08
PAWS-X	KM	0.10	0.09
	peculiarity	0.10	0.08
MARC-2	KM	0.14	0.09
	peculiarity	0.16	0.09
MARC-5	KM	0.10	0.10
	peculiarity	0.08	0.75
NER	KM	0.24	0.14
	peculiarity	0.20	0.09
POS	KM	0.22	0.13
	peculiarity	0.19	0.10

Table 4: Token type ratio per each langauge group.

steps with KM[8]. We use Yuan et al. (2020)'s implementation for GE-KM and our re-implementation for LE-KM.

***Peculiarity* (ours).** We select the candidates with the top-n *peculiarity* according to Equation 1. We use Faiss (Johnson et al., 2019) library[9] to search the k-NN source language examples. We set 20 and Euclidean distance as k and $distance$, respectively, in all experiments.

6 Experimental Results

In this section, we confirm the effectiveness of *peculiarity*. We first show that *peculiarity* isolates the examples based on their accuracy in the zero-shot cross-lingual transfer setting (Subsection 6.1), and then demonstrate that the examples of high *peculiarity* are useful for few-shot cross-lingual transfer (Subsection 6.2).

6.1 Zero-shot transfer for *peculiar* examples

First, we examine whether *peculiarity* (Equation 1) can isolate the examples predicted correctly and incorrectly in the zero-shot setting. We construct two subsets by extracting 10% examples of bottom and top *peculiarity* from each language's validation data and measure their accuracy. We show the average accuracy of all languages for each subset in Table 1 and more details about each language in Appendix B. This table shows that the examples with high *peculiarity* cannot be predicted accurately in the zero-shot setting. Specifically, the average score is 85.1% in "Lowest", but is 59.2% in "Highest". This result indicates that *peculiarity* can extract the examples of low accuracy in the zero-shot setting without their labels.

6.2 Enhancing few-shot transfer with high *peculiarity* candidates

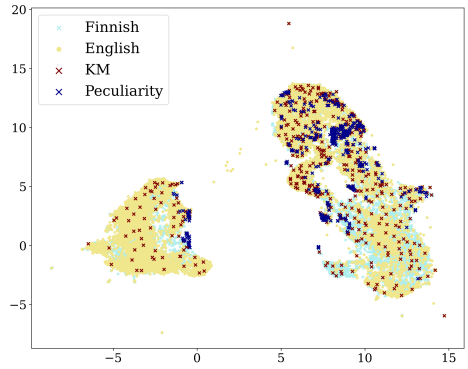
We conduct few-shot cross-lingual transfer using candidates extracted by *peculiarity*. We show the experimental results for few-label (2–3) classification tasks (XNLI, PAWS-X, and MARC-2) in Table 2 and many-label (5) classification and sequence-tagging tasks (MARC-5, NER, and POS) in Table 3[10]. We report delta scores in accuracies between 0-shot and n-shot models following Kumar et al. (2022). These scores are the average across languages in each group.

Table 2 shows that the methods proposed in the previous studies enhance accuracy for each language group but do not outperform random baseline consistently. However, *peculiarity* consistently achieves the highest scores regardless of the language groups. Therefore, we conclude that *peculiarity* is useful for selecting annotation candidates

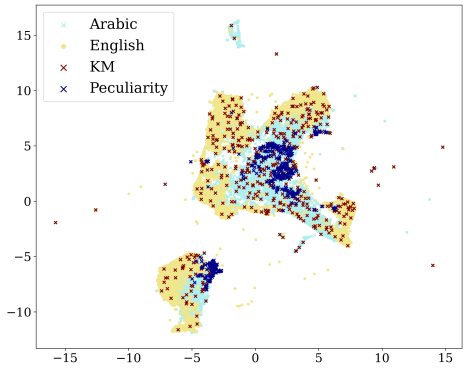
[8] GE-KM and LE-KM correspond to the methods proposed by Ash et al. (2020) and Kumar et al. (2022), respectively.

[9] https://github.com/facebookresearch/faiss

[10] In Appendices B and C, we provide more detailed results including statistical significance tests.



(a) Finnish ("good" group)



(b) Arabic ("poor" group)

Figure 3: Visualization of the source (English) and target (Finish or Arabic) examples. We indicate the source language in yellow and the target language in light blue. We color the candidates extracted by *peculiarity* and KM in blue and red, respectively.

		All	Good	Poor
MARC-5	Best	1.83	**1.71**	2.50
	peculiarity-KM	**1.88**	1.47	**2.62**
NER	Best	12.39	5.11	21.35
	peculiarity-KM	**13.13**	**5.50**	**22.27**
POS	Best	17.08	6.66	25.93
	peculiarity-KM	**17.26**	**6.75**	**26.17**

Table 5: Evaluation of *peculiarity*-KM under 1,000-shot setting. The scores of Best for "good" and "poor" groups are those of *peculiarity* and KM, respectively, and are the highest scores in Table 3.

that could improve the performance for these few-label classification tasks.

On the contrary, for MARC-5, NER, and POS the best methods differ in each language group. In the "good" group, *peculiarity* consistently achieves the highest improvements in the all tasks and n-shots. However, in the "poor" group, *peculiarity* does not work as well as in the "good" and is worse than random in some settings. In addition, the scores show that the methods considering the diversity of candidates (KM, GE-KM, and LE-KM) are the best choice for the "poor" group[11].

7 Analysis

7.1 Redundancy of high *peculiarity* examples

As mentioned above, *peculiarity* outperforms the other methods in the few-label classification tasks (XNLI, PAWS-X, and MARC-2), but the KM-based methods perform better in the "poor" group for sequence-tagging and many-label classification tasks (NER, POS, and MARC-5). The KM-based methods select candidates by considering their diversity, but *peculiarity* does not.

Therefore, we assume that *peculiarity* selects more redundant candidates than KM-based methods. To confirm our assumption, we measure the token type ratio (TTR) of the candidates extracted by KM and *peculiarity*. Table 4 shows the results and that the TTRs of KM and *peculiarity* are almost the same in the few-label classification tasks. By contrast, the TTRs of *peculiarity* are lower than that of KM in the sequence-tagging and many-label classification tasks, which means that *peculiarity* selects redundant candidates in these tasks. Intuitively, training a model with only similar examples harms generalization performance because the model is optimized only for limited data. We consider that the redundancy is one of the causes of *peculiarity* not working well across languages.

Then, we verify why *peculiarity* works well in the "good" group but not the "poor" group. To do this, we chose an NER task and Finish and Arabic from the "good" and "poor" groups, respectively, and visualize the hidden states of the candidates extracted by KM and *peculiarity* ($n = 1,000$). We compress the hidden states to two dimensions using UMAP (McInnes et al., 2018) and show the results in Figure 3. This figure shows that *peculiarity* selects local candidates compared to KM in the both languages. For Finnish, these candidates are enough to cover the Finish examples that are not covered by English (source language). However,

[11]We confirm that GE-KM and LE-KM achieve the highest scores in 500-shot settings. The results for 100- and 500-shot experiments are in Appendix C.

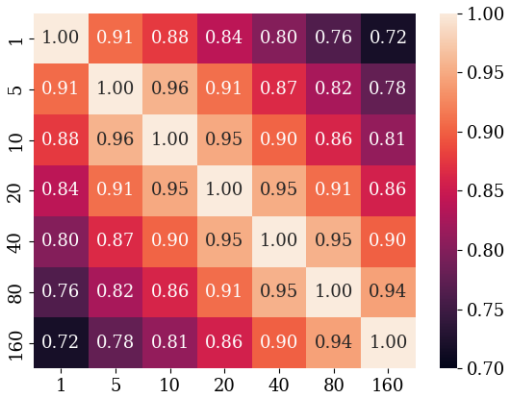


Figure 4: Overlap rates of candidates extracted by *peculiarity* measured at each k.

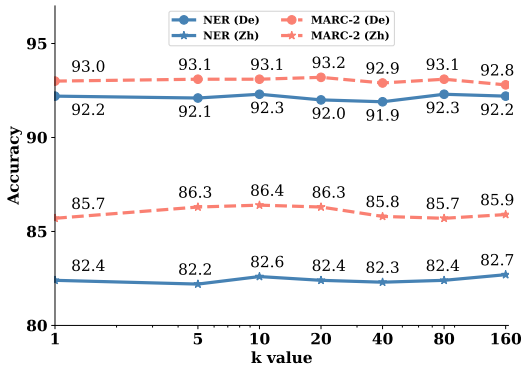


Figure 5: Accuracy of the models fine-tuned by the examples selected by *peculiarity* measured by each k.

for Arabic, the candidates selected by *peculiarity* cannot complement all of the Arabic examples because almost all of the examples are not covered by English examples. Several studies (Yuan et al., 2020; Hacohen et al., 2022) have suggested that training a model with representative and diverse examples, in other words, examples that approximate the entire data pool, is beneficial to enhance the model's performance when the training examples are limited. In the case of few-shot cross-lingual transfer, we should fine-tune $\mathbf{M}^{\mathbf{S}}$ using source and target language examples that approximate an unlabeled data pool in the target language X_T. Figure 3 indicates that *peculiarity* selects ideal candidates in Finish, but not in Arabic. Therefore, it is recommended to select examples by considering diversity for the "poor" group, such as Arabic.

Motivated by the previous analyses, we design a new candidate selection method that considers *peculiarity* and diversity of candidates. Specifically, first, we extract 50% of candidates $X_T^{\prime\prime} \subset X_T$ using *peculiarity*. Then, we select candidates $X_T^{\prime} \subset X_T^{\prime\prime}$ by adapting KM to $X_T^{\prime\prime}$. We denote this method as *peculiarity*-KM and provide evaluation results on NER, POS, and MARC-5 tasks in Table 5. *Peculiarity*-KM achieves the highest scores in the both "good" and "poor" groups for the NER and POS tasks, which means that we can mitigate the weakness of *peculiarity* by considering the diversity. However, the scores drop 0.24 points compared to *peculiarity* in the "good" group of MARC-5. Therefore, there are still challenges to address in the candidate selection method that can efficiently enhance few-shot cross-lingual performance across languages and tasks.

7.2 Robustness of *peculiarity*

Peculiarity (Equation 1) has two hyperparameters k and $distance$. Finally, we analyze the robustness of the hyperparameters of *peculiarity*.

We measure *peculiarity* using various k (1, 5, 10, 20, 40, 80, and 160) and extract 1,000 candidates based on each *peculiarity*. Figure 4 shows overlap rates between each subset of extracted candidates in MARC-2 and NER tasks averaged across all languages. We indicate the overlap rates for the other tasks in Appendix C. The overlap rates gradually decrease as the k value changes. However, even the lowest overlap rates (between 1 and 160) are 0.72 and 0.74 in MARC-2 and NER, respectively. Therefore, this figure reveals that *peculiarity* selects almost same examples regardless of k. We also conduct few-shot cross-lingual transfer with these candidates. In this experiment, we chose German (De) and Chinese (Zh) from the "good" and "poor" groups and use MARC-2 and NER tasks. Figure 5 shows the evaluation results. The figure demonstrates that the accuracy are almost consistent between each k value. This result reveals that *peculiarity* can lead to robust performance regardless of k.

In previous experiments, we adopted Euclidean distance as $distance$; however, Cosine distance is another option. We extract 1,000 candidates based on Euclidean and Cosine distances and measure the overlap rates between each subset of extracted candidates. The overlap rates averaged across all languages are 0.93, 0.99, 0.98, 0.97, 0.96, and 0.93 in XNLI, PAWS-X, MARC-2, MARC-5, NER, and POS tasks, respectively. This indicates that *peculiarity* extract consistent candidates regardless of distance metric.

8 Conclusion

In this study, we proposed a simple metric called *peculiarity*, which measures whether source language examples exist in the neighborhood of target language examples. We showed that high *peculiarity* examples are not predicted correctly in the zero-shot setting and demonstrated that these examples can enhance accuracy for few-shot cross-lingual transfer regardless of language groups.

In addition, we showed that *peculiarity*-KM, the candidate selection method that considers *peculiarity* and diversity, further boosts few-shot cross-lingual transfer performance compared to *peculiarity* or KM alone. However, there are some configurations in which *peculiarity*-KM does not work well, such as the "good" group in MARC-5. Therefore, we would like to attempt to analyze the causes of this result and develop a new candidate selection method that enhances the few-shot cross-lingual performance across languages and tasks.

Limitations

Although we demonstrated that the proposed metric *peculiarity* is useful for selecting candidates for few-shot cross-lingual transfer, our current work has the following limitations.

Lack of evaluations to argue the usefulness of *peculiarity*. We demonstrated that *peculiarity* selects candidates to efficiently enhance few-shot cross-lingual performance in several tasks and languages. In addition, *peculiarity* is robust for hyperparameter k. However, further verification is required to evaluate the usefulness of *peculiarity*.

In this study, we only used XLM-R as the MMLM in the experiments, because previous works (Lauscher et al., 2020; Kumar et al., 2022) have demonstrated that mBERT and XLM-R show the same trend and XLM-R achieves better zero-shot and few-shot cross-lingual performance. However, it is not obvious that *peculiarity* will work well in mBERT. In addition, recently, Lin et al. (2022) proposed XGLM, a pre-trained multilingual causal language model, that demonstrates strong multilingual capabilities. We would like to experiment using these pre-trained multilingual models to show the usefulness of *peculiarity* regardless of models.

We fine-tuned the MMLM using a standard training objective, predicting true labels or tags for inputs. On the contrary, Zhao and Schütze (2021) revealed that fine-tuning in a prompting format encourages better zero-shot and few-shot cross-lingual transfer than the standard fine-tuning. It is worthwhile to examine few-shot cross-lingual transfer performance when fine-tuning the MMLM with high *peculiarity* examples in a prompting format because it may be possible to achieve higher accuracy in the target languages with a smaller amount of examples.

We experimented using English as the source language. However, if possible, it is better to use a language that is linguistically close to the target language as the source language (Pires et al., 2019; Lauscher et al., 2020; Chai et al., 2022). In our experiments, we did not show that *peculiarity* works well regardless of source languages. Therefore, verifying this aspect is also a remaining challenge.

Definition of annotation cost. In this study, we defined annotation cost in terms of the number of candidates following previous studies (Pires et al., 2019; Lauscher et al., 2020; Chai et al., 2022). However, a small number of candidates does not necessarily mean less work for annotators. If a candidate (sentence) length is long or hard, it is considered to take longer to understand. On the other hand, if the candidate length is short or easy, annotation time per candidate will be shorter, and the annotators can annotate more candidates in the same time. Therefore, we should evaluate candidate selection methods based on total time required for annotation.

In addition, aligning the cross-lingual representations between source and target languages using bilingual data is one approach to enhance accuracy for the target languages (Lample and Conneau, 2019; Cao et al., 2020; Chi et al., 2021; Dou and Neubig, 2021; Yang et al., 2021). To align the representations, we should create bilingual data through a human or automatic translator. Verification whether labeling or translating is less labor intensive and further boosting performance is one of the future goals.

Developing a better *peculiarity*-based candidate selection method. In this study, we used the `BOS` hidden states to measure *peculiarity*; in other words, it measures example-level *peculiarity*. In classification tasks, using example-level *peculiarity* to select candidates is intuitive because we predict labels based on the `BOS` hidden states. On the other hand, in the sequence-tagging tasks, we

predict token tags based on hidden states of each token. In addition, we consider that it is necessary to fine-tune $\mathbf{M}^{\mathbf{S}}$ with *peculiar* tokens, tokens that are not covered by the source language, to ensure that the model predicts tags of these tokens correctly. Therefore, we will attempt to select candidates that contain *peculiar* tokens by using token-level *peculiarity* and conduct few-shot cross-lingual transfer in the sequence-tagging tasks.

We observed that *peculiarity* selects more redundant candidates compared to the KM-based methods and argued that this aspect is the reason that *peculiarity* does not work in the "poor" group. We consider the possibility of other reasons for this behavior. Several studies (Swayamdipta et al., 2020; Sorscher et al., 2022; Hacohen et al., 2022) have suggested that if only a small amount of examples can be used for training, it is important to use not only hard (atypical) examples but also some easy (typical) examples for training in order to improve model performance. In terms of few-shot cross-lingual transfer using *peculiarity*, we should fine-tune $\mathbf{M}^{\mathbf{S}}$ with the both highest and lowest *peculiarity* examples. In addition, using typical examples selected by KM instead of the lowest *peculiarity* examples is one of the approaches. For future work, we would like to verify the effectiveness of these methods for few-shot cross-lingual transfer.

Ethics Statement

Impact of our work. Thanks to the efforts of various researchers, pre-trained models have been proposed that can solve NLP tasks with high accuracy. However, labeled data is essential for fine-tuning these models, and few languages have abundant language resources like English. In addition, construction of labeled data is not easy. Therefore, attempts to train high-quality models with little effort, as in our study, are very important for low-resource languages. Although our method and study have the limitations mentioned above, our experiments provided useful insights into selecting annotation candidates for few-shot cross-lingual transfer. In addition, we will publish the code used in our experiments, which will facilitate the reproduction of our experiments and contribute to further research.

Potential risks for bias. In recent years, bias in data has become an issue. Training a model on such data can lead to unwarranted predictions or generate negative sentences for a particular person or group. When the training data is small and contains biases, such problems may be more pronounced because the model is optimized only for the provided data. In this study, we did not take into account this issue, and our proposed method is not designed to select bias-less candidates. Therefore, when using the proposed method, sufficient attention should be paid to the problem of bias.

Acknowledgments

This work was supported by TMU research fund for young scientists and JST, the establishment of university fellowships towards the creation of science technology innovation, Grant Number JPMJFS2139.

References

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2020. Deep batch active learning by diverse, uncertain gradient lower bounds. In *International Conference on Learning Representations*.

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

Yuan Chai, Yaobo Liang, and Nan Duan. 2022. Cross-lingual ability of multilingual masked language models: A study of language structure. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4702–4712, Dublin, Ireland. Association for Computational Linguistics.

Ernie Chang, Xiaoyu Shen, Hui-Syuan Yeh, and Vera Demberg. 2021. On training instance selection for few-shot neural text 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 2: Short Papers)*, pages 8–13, Online. Association for Computational Linguistics.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised

cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, 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.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online. Association for Computational Linguistics.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Guy Hacohen, Avihu Dekel, and Daphna Weinshall. 2022. Active learning on a budget: Opposite strategies suit high and low budgets. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 8175–8195. PMLR.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. The multilingual Amazon reviews corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4563–4568, Online. Association for Computational Linguistics.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, pages 3519–3529.

Shanu Kumar, Sandipan Dandapat, and Monojit Choudhury. 2022. "Diversity and uncertainty in moderation" are the key to data selection for multilingual few-shot transfer. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1042–1055, Seattle, United States. Association for Computational Linguistics.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*.

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.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona T. Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online and Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. 2018. UMAP: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

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.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. Association for Computational Linguistics.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari S. Morcos. 2022. Beyond neural scaling laws: beating power law scaling via data pruning. In *Advances in Neural Information Processing Systems*.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

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.

Ziqing Yang, Wentao Ma, Yiming Cui, Jiani Ye, Wanxiang Che, and Shijin Wang. 2021. Bilingual alignment pre-training for zero-shot cross-lingual transfer. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 100–105, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Michelle Yuan, Hsuan-Tien Lin, and Jordan Boyd-Graber. 2020. Cold-start active learning through self-supervised language modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7935–7948, Online. Association for Computational Linguistics.

Mengjie Zhao and Hinrich Schütze. 2021. Discrete and soft prompting for multilingual models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8547–8555, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

	WikiANN			UD		
	train	dev	test	train	dev	test
En	23,234	11,624	11,594	359,550	56,194	76,817
Ar	20,912	10,460	10,438	553,404	73,707	96,888
Bg	23,422	11,628	11,670	143,528	18,615	18,095
De	24,834	12,398	12,430	3,459,957	379,232	437,084
El	23,658	11,827	11,837	48,402	12,086	12,106
Es	21,153	10,548	10,565	844,454	91,608	91,776
Eu	12,126	12,144	11,966	69,285	22,793	23,095
Fi	22,968	11,421	11,472	309,406	36,006	77,710
He	22,931	11,468	11,414	141,995	11,716	12,625
Hi	5,279	1,067	1,060	592,684	74,192	107,210
It	22,732	11,266	11,361	–	–	–
Ja	53,242	26,551	26,988	221,320	15,982	74,428
Ko	22,444	11,185	11,241	146,605	15,297	41,244
Ru	21,904	10,934	10,942	1,293,061	181,397	203,482
Sv	25,156	12,523	12,643	149,979	34,271	70,270
Th	103,972	50,880	52,213	–	–	–
Tr	21,926	11,095	11,046	143,180	22,997	55,59
Ur	20,148	1,001	1,003	231,161	31,000	31,667
Vi	20,704	10,321	10,371	–	–	–
Zh	40,354	20,470	19,983	31,667	17,780	54,261

Table 6: Example numbers for each datasets.

A Experimental Settings

We experiment with WikiANN[12], UD[13], XNLI[14], PAWS-X[15], and MARC (MARC-2 and MARC-5)[16] datasets. Example numbers of train, dev, and test data are the same across all languages in the XNLI (392,702, 2,418, and 5,010), PAWS-X (49,401, 2,000, and 2,000), MARC-2 (160,000, 4,000, and 4,000), and MARC-5 (200,000, 5,000, and 5,000) datasets. We show example numbers for the other datasets in Table 6.

B Zero-shot Cross-lingual Transfer

We show the accuracy of the test and validation datasets for each target language in zero-shot cross-lingual transfer in Tables 7, 8, 9, 10, and 11. Low and High are subsets of extracted 10% examples of the lowest and highest *peculiarity* from the validation dataset. These scores are the average accuracy of three models. The accuracies of High are obviously lower than others in all languages and tasks. Therefore, these tables indicate that *peculiarity* can isolate the languages depending on their zero-shot cross-lingual performance. We also indicate the median (Mdn.) and macro-average (Avg.) scores (excluding the English score). We underline the scores higher than the median score in the validation dataset, and we use those languages as "good" group and the others as "poor" group.

C Few-shot Cross-lingual Transfer

We show the accuracy of the test datasets for each target language in few-shot cross-lingual transfer in Tables 12, 13, 14, 15, and 16. We perform pairwise t-Test to measure statistical significance. The † (or ‡) indicates the statistical significance ($p < 0.1$)[17] between *peculiarity* (or *peculiarity*-KM) and the underlined method that achieves the highest average score except for the *peculiarity*-based methods. In addition, the ⋆ indicates the statistical significance between *peculiarity* and *peculiarity*-KM.

In PAWS-X, MARC-2, and MARC-5 (Tables 15 and 16) tasks, the *peculiarity*-based methods consistently achieves the highest accuracy. In NER, POS, and XNLI tasks (Tables 12, 13, and 14), the *peculiarity*-based methods bring the best performances in almost all languages. We also show differences in accuracy between 0-shot and each n-shot model averaged across all languages in Table 17.

Finally, we show overlap rates of examples extracted by *peculiarity* measured at each k and observe that they are higher than 0.70. Therefore, *peculiarity* is robust for hyperparameter k.

[12] https://huggingface.co/datasets/wikiann

[13] https://huggingface.co/datasets/universal_dependencies

[14] https://github.com/facebookresearch/XNLI

[15] https://huggingface.co/datasets/paws-x

[16] https://huggingface.co/datasets/amazon_reviews_multi

[17] We set p value following Kumar et al. (2022).

	En	Ar	Bg	De	El	Es	Eu	Fi	Fr	He	Hi	It	Ja	Ko	Ru	Sv	Th	Tr	Ur	Vi	Zh	Mdn.	Avg.
Test	91.7	67.5	89.2	88.5	87.5	83.4	79.8	88.8	84.3	75.8	75.6	88.2	68.9	74.6	80.8	89.2	21.2	83.9	60.1	83.5	69.0	82.1	76.9
Valid	92.1	67.3	89.5	88.7	87.2	83.5	79.8	88.5	85.3	75.4	75.1	88.5	67.8	75.1	81.3	88.8	21.5	83.7	61.8	83.0	68.8	82.1	76.9
Low	–	93.7	96.7	97.8	95.6	98.1	96.8	97.9	98.0	90.1	96.0	98.6	76.9	85.7	93.7	98.1	25.6	99.1	95.6	98.1	74.4	96.3	89.3
High	–	37.0	57.5	62.1	62.8	49.5	65.3	64.1	56.7	35.9	53.8	55.4	36.6	43.9	53.0	66.9	25.3	54.1	51.3	42.2	32.8	53.4	50.3

Table 7: Zero-shot cross-lingual transfer performance on NER.

	En	Ar	Bg	De	El	Es	Eu	Fi	He	Hi	Ja	Ko	Ru	Sv	Tr	Ur	Zh	Mdn.	Avg.
Test	96.0	67.9	91.3	89.0	86.7	89.7	72.9	88.3	56.9	78.1	43.8	61.2	91.6	93.9	73.8	65.8	57.5	75.9	75.2
Valid	96.5	56.2	91.1	87.7	86.7	89.5	72.8	86.4	58.2	79.0	45.8	59.0	91.5	94.7	73.0	64.7	57.2	76.0	74.4
Low	–	65.7	91.7	93.0	87.5	92.0	76.9	89.7	69.8	81.7	53.1	64.9	93.3	95.6	78.4	70.2	61.6	80.0	79.0
High	–	50.5	88.8	82.2	82.9	79.3	68.1	73.8	47.9	79.3	46.2	54.0	83.2	92.0	62.8	67.4	56.0	70.9	69.5

Table 8: Zero-shot cross-lingual transfer performance on POS tagging.

	En	Ar	Bg	De	El	Es	Fr	Hi	Ru	Sw	Th	Tr	Ur	Vi	Zh	Mdn.	Avg.
Test	83.0	69.5	76.6	74.7	74.4	77.5	77.0	68.3	75.1	62.6	70.4	71.3	64.1	73.6	75.5	74.4	72.4
Valid	82.4	69.2	74.4	74.5	73.5	76.8	75.5	68.0	74.2	62.5	69.7	71.0	63.5	73.1	76.2	73.5	71.8
Low	–	73.9	76.0	78.7	74.8	84.7	86.4	74.4	77.3	64.2	65.2	72.5	45.7	69.8	86.4	74.8	73.5
High	–	62.5	74.4	74.4	72.9	77.4	74.4	50.8	76.5	46.1	59.6	53.9	37.9	66.0	74.4	72.9	64.7

Table 9: Zero-shot cross-lingual transfer performance on XNLI.

	En	De	Es	Fr	Ja	Ko	Zh	Mdn.	Avg.
Test	92.9	85.6	86.2	86.2	74.6	71.5	77.7	81.6	80.3
Valid	92.6	83.8	86.5	87.6	75.1	76.5	79.2	81.5	81.4
Low	–	99.4	99.8	99.4	99.9	99.0	99.9	99.6	99.6
High	–	69.8	71.7	68.5	65.8	65.4	60.9	67.1	67.0

Table 10: Zero-shot cross-lingual transfer performance on PAWS-X.

		En	De	Es	Fr	Ja	Zh	Mdn.	Avg.
MARC-2	Test	94.0	92.9	92.0	91.8	89.1	85.1	91.8	90.2
	Valid	93.5	91.7	91.5	91.5	89.5	85.1	91.5	89.9
	Low	–	99.9	100.0	100.0	99.5	97.7	99.9	99.4
	High	–	52.6	55.0	52.9	53.5	52.4	52.9	53.2
MARC-5	Test	60.2	59.8	55.0	54.9	52.8	49.6	54.9	54.4
	Valid	59.5	59.0	54.4	54.2	53.8	48.6	54.2	54.2
	Low	–	66.8	74.9	69.4	64.2	77.7	69.4	70.6
	High	–	54.6	57.2	55.9	43.4	43.3	54.6	50.8

Table 11: Zero-shot cross-lingual transfer performance on MARC.

		Good										Poor									
		Bg	De	El	Es	Fi	Fr	It	Sv	Tr	Vi	Ar	Eu	He	Hi	Ja	Ko	Ru	Th	Ur	Zh
100-shot	Random	90.5	90.0	**90.0**	89.3	90.3	89.2	90.2	91.8	90.7	86.9†	**84.6**†‡	92.0†	82.6	84.5	**81.3**†‡	**82.5**†‡	87.3	73.5	87.0	83.8†
	Entropy	90.2	89.9	89.8	89.7	90.7	88.6	90.3	90.1	89.8	86.7	83.8	87.5	82.1	84.3	78.1	80.7	87.2	63.5	87.3	78.1
	KM	90.7	90.3	89.6	89.9†	91.1	88.5	90.4	91.9	90.9	86.9†	83.6	91.8	**83.1**†‡	**86.8**†‡	80.3	81.2	87.3	73.6	88.9†	83.6
	LE-KM	89.2	89.5	89.6	89.0	90.6	87.8	89.9	92.2	**91.3**†‡	85.2	83.2	91.6	81.9	83.2	80.8	81.0	86.8	74.4	88.8	81.6
	peculiarity	**91.5**†	90.8†	89.8	89.5	**91.6**†	**89.5**†*	**90.7**†	92.1	90.3	86.5	83.3	87.9	82.3	84.6	80.1	80.2	88.0†	74.7	87.3	80.2
	peculiarity-KM	91.2‡	**91.3**‡*	89.9	**92.3**‡*	91.5	89.0	90.5	**93.3**‡*	89.9	**87.1**‡*	83.7*	**92.8**‡*	82.3	84.5	80.8*	81.1*	**88.5**‡*	**75.3**‡*	**89.7**‡*	**84.0**‡*
500-shot	Random	92.1	91.5	90.5	91.2†	92.4	89.8	91.1	94.2	92.3	87.6	86.9	92.8	85.9	88.8	83.3	86.2	89.0	80.1	90.1	85.4
	Entropy	92.0	91.3	90.6	88.1	92.5	89.9	91.6	92.6	91.8	88.2	85.7	88.1	84.4	85.7	80.3	85.0	88.1	77.3	90.6	70.7
	KM	91.3	91.5	**91.8**	89.7	91.8	90.0	91.0	94.1	92.1	88.5	**87.6**†	89.5	86.1†	**88.9**†‡	82.3	85.3	89.5	83.1†	91.0	86.4†
	LE-KM	90.6	91.5	90.4	90.4	92.2	89.3	91.7	94.3	92.4	88.1	86.8	93.4†	86.1†	85.5	83.8†	**86.5**†‡	87.0	81.6	**91.3**†	85.3
	peculiarity	**92.4**†*	91.6	91.6	90.3	**92.6**	**90.3**†*	**92.4**†	94.3	92.6	88.9†	87.3	93.0	85.6	87.3	82.3	85.1	**89.6**	80.9	89.6	81.6
	peculiarity-KM	92.2	**91.9**‡	91.7	**94.2**‡*	92.4	89.9	92.2‡	**95.1**‡*	**93.3**‡*	**90.0**‡	87.5	**95.0**‡*	**86.8**‡*	86.5	**84.0***	85.0	89.4	**83.9**‡*	91.2*	**87.2**‡*
1,000-shot	Random	92.6	91.9	91.6	91.5†	92.6	89.6	91.7	95.0	93.3	89.0	87.5	94.2	87.3	87.6	85.0†	86.5	90.0	84.0	91.7	87.2
	Entropy	91.8	92.3	92.0	90.1	93.2	90.0	91.6	95.2	92.6	89.4	85.7	93.2	86.7	88.8	82.0	86.4	89.6	79.9	91.2	81.0
	KM	92.8	91.9	90.9	91.0	92.4	89.9	92.3	95.3	**93.9**	89.8	**89.0**†‡	94.5†	87.1	87.2	84.8	87.3†	90.4	83.3	92.2†	86.6
	LE-KM	**93.4**†‡	91.8	91.9	91.1	93.0	90.0	91.7	95.1	93.4	88.5	88.4	94.1	87.8†	86.7	84.5	86.4	89.6	**85.4**†‡	90.6	88.2†
	peculiarity	92.8*	92.1	**92.4**†*	90.3	**93.3**	**90.6**†*	**92.9**†*	95.4	93.6	89.8	86.5	90.3	86.4	89.2†	83.8	85.4	90.3	83.3	90.8	82.4
	peculiarity-KM	92.2	**92.4***	92.0	**94.5**‡*	93.2	90.1	92.0	**96.1**‡*	93.7	**91.1**‡*	88.4*	**94.6***	**87.9***	**89.8**‡*	**85.1***	**88.5**‡*	**91.1**‡*	84.4*	**92.8**‡*	**88.9**‡*

Table 12: Few-shot cross-lingual transfer on NER.

		Good								Poor							
		Bg	De	El	Es	Fi	Hi	Ru	Sv	Ar	Eu	He	Ja	Ko	Tr	Ur	Zh
100-shot	Random	94.7	90.8	94.2	93.6	89.0	87.6	92.8	94.5	**87.9**	**87.4**†‡	94.4	88.3	**75.3**†‡	82.0	88.9	86.8
	Entropy	92.9	89.6	92.9	92.3	88.0	87.4	92.4	94.7	87.1	85.2	93.6	84.8	68.5	81.3	88.6	87.0
	KM	94.5	91.5	93.7	93.5	89.9	88.4	94.2	95.0	87.2	85.5	94.1	86.5	72.6	81.7	88.9	87.4
	LE-KM	95.2	91.8	93.7	93.8	90.0	88.6	94.0	84.3	87.5	85.6	**95.0**‡	88.3	74.3	**82.7**†‡	89.1	87.1
	peculiarity	**96.0**†*	91.7	94.3	**94.2**†	**90.5**†	88.7	94.2	**95.1**	87.7	86.7	94.9*	**88.4**	74.6	81.8	**90.9**†*	87.5
	peculiarity-KM	95.0	**92.2**‡*	**94.8**‡*	94.0	90.2	**89.1**‡*	**94.6**‡*	95.0	87.7	86.5	94.5	88.3	74.8	82.3*	90.3‡	**88.6**‡*
500-shot	Random	99.1	94.0	95.9	94.9	90.1	89.6	94.8	95.8	87.3	90.0	**96.4**	90.5‡	77.0	85.4	91.2	90.3
	Entropy	99.1	91.4	95.4	94.2	89.6	89.3	94.8	96.0	88.1	89.1	95.9	89.9	76.1	85.1	90.8	90.6
	KM	99.1	94.4	96.1	95.3	90.4	89.8	95.6	96.1	87.7	91.0	96.1	89.8	78.0	83.9	91.4	90.7
	LE-KM	97.2	94.7	95.9	95.0	90.7‡	**89.9**†	95.1	96.0	87.9	**91.8**†‡	**96.4**	90.1	**79.3**†‡	85.8	**92.2**	91.1†
	peculiarity	**99.2**	94.5	**96.2**	**95.9**†*	**90.8***	89.5	95.3	**96.4**†	**88.4**	89.9	**96.4**	**90.6***	76.4	85.5	**92.2**	90.4
	peculiarity-KM	**99.2**	**94.8**	**96.2**	95.6	90.2	**89.9***	**95.9**‡*	96.2	88.2	91.2*	**96.4**	90.0	78.2*	**86.3**‡*	92.0	**91.5**‡*
1,000-shot	Random	99.1	94.6	96.5	95.4	90.4	90.4	95.5	96.7	88.2	92.3	96.6	90.9	80.3	86.4	92.8	91.7†
	Entropy	**99.2**	92.2	96.4	94.5	90.1	90.9	95.8	96.6	87.9	92.6	96.4	90.3	78.0	86.4	91.7	90.7
	KM	**99.2**	94.3	96.2	95.3	90.8	90.7	95.7	96.3	88.4	92.4	96.6	90.6	80.5†	87.6	92.6	90.4
	LE-KM	98.3	94.4	96.5	95.4	91.0	90.8	95.5	96.6	**88.5**	92.6	96.7	90.6	80.4	87.6	92.5	91.5
	peculiarity	**99.2**	94.9	**96.7**	96.0†	**91.5**†	91.0	95.8	**97.1**†*	**88.5**	92.7	96.7	**91.5**†	79.8	87.7	92.8	90.4
	peculiarity-KM	**99.2**	**95.2**‡*	96.5	**96.2**‡	**91.5**‡	**91.9**‡*	**96.3**‡*	96.8	88.4	**93.0**‡	**96.9**	91.3‡	**80.8***	**88.1**‡*	**93.2**‡*	**92.2**‡*

Table 13: Few-shot cross-lingual transfer performance on POS tagging.

		Good							Poor						
		Bg	De	El	Es	Fr	Ru	Zh	Ar	Hi	Sw	Th	Tr	Ur	Vi
100-shot	Random	78.1	**76.1**	75.6	78.4	78.1	**76.3**†	73.7	72.2	69.2	61.4	71.9	72.4	66.7	75.0
	Entropy	77.9	75.9	76.0	78.7	77.9	76.0	73.9	72.5	69.1	63.3	71.9	72.1	66.5	74.7
	KM	**78.6**	76.0	75.9	78.8	77.9	76.1	73.4	72.4†	69.0	63.6	71.7	**72.7**†	**67.9**†	75.0
	GE-KM	77.9	76.0	75.9	78.8	77.9	75.6	**74.1**	71.5	69.4	60.5	72.1	72.3	66.5	74.9
	peculiarity	**78.6**	**76.1**	**76.2**	**78.9**	**78.3**†	75.7	**74.1**	**72.6***	**69.9**†	**64.0**†	**72.5**†	72.4	66.9	**75.1**
500-shot	Random	77.9	76.1	75.4	78.6	78.1	**76.0**	73.2	71.6	69.2	64.1	72.5	72.0	66.0	75.2
	Entropy	78.0	76.6	**75.9**	78.5	78.3	75.4	**74.0**	71.7	69.1	63.3	**72.6**	72.3	66.4	75.1
	KM	77.9	**77.2**†	75.1	78.6	78.0	75.3	73.4	71.7	69.6	63.4	71.8	71.8	66.0	75.0
	GE-KM	77.8	76.1	75.4	78.5	77.9	75.4	73.3	70.6	69.5	62.1	72.2	**72.4**	66.2	74.9
	peculiarity	**78.2**	76.9	75.6	**78.7**	**78.4**	75.7	73.9	**71.9**	**69.9**†	**64.6**†	72.5	72.3	**66.7**†	**75.4**
1,000-shot	Random	78.3	76.5	75.9	78.8	78.2	75.5	74.7	72.5	69.9	65.0	72.8	72.2	62.5	75.5
	Entropy	78.0	77.2	75.6	79.1	78.4	76.1	74.6	72.2	69.5	62.6	**73.2**	72.7	64.2	75.3
	KM	78.2	77.2	**76.0**	79.1	78.0	75.6	74.5	71.5	69.1	64.5	73.0	72.1	64.8	74.9
	GE-KM	78.2	76.3	75.5	79.0	78.0	75.5	74.2	71.6	69.6	63.8	72.2	**73.0**†	66.6	75.2
	peculiarity	**78.5**	**77.3**	75.9	**79.4**†	**78.8**†	**76.2**	**74.9**	**72.9**†	**70.2**†	**65.2**	**73.2**	72.7	**67.2**†	**75.6**

Table 14: Few-shot cross-lingual transfer performance on XNLI.

		Good			Poor		
		De	Es	Fr	Ja	Ko	Zh
100-shot	Random	87.4	89.3	89.1	**77.3**	76.5	78.8
	Entropy	87.9	89.3	89.8	77.0	77.2	82.2
	KM	88.1	89.0	89.9	**77.3**	76.8	81.6
	GE-KM	87.7	89.0	89.8	**77.3**	77.3	81.8
	peculiarity	**88.8**†	**89.4**	**90.3**†	**77.3**	**77.4**	**82.7**†
500-shot	Random	87.6	89.4	89.5	77.4	78.1	80.8
	Entropy	87.7	89.8	89.9	77.0	78.2	81.2
	KM	87.4	89.5	89.6	77.0	78.0	80.9
	GE-KM	87.5	89.6	89.7	75.5	77.7	**81.9**
	peculiarity	**88.4**†	**90.0**†	**90.1**	**77.8**†	**78.9**†	**81.9**
1,000-shot	Random	87.5	88.4	89.1	77.0	77.6	81.5
	Entropy	87.9	89.0	89.2	75.2	75.7	81.5
	KM	87.6	89.9	89.8	76.6	77.7	**81.7**
	GE-KM	88.0	89.9	89.7	75.7	**78.0**	81.1
	peculiarity	**88.7**†	**90.0**	**90.3**†	**77.4**†	77.9	**81.7**

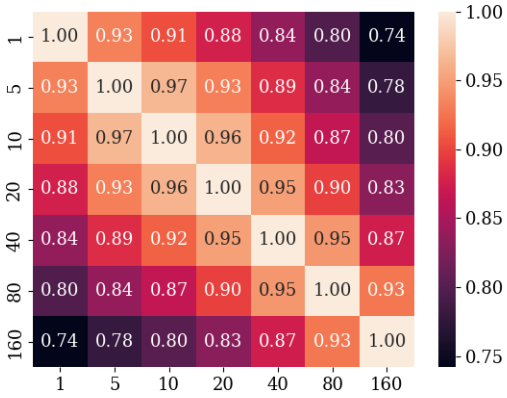
Table 15: Few-shot cross-lingual transfer performance on PAWS-X.

			Good			Poor	
			De	Es	Fr	Ja	Zh
MARC-2	100-shot	Random	93.0	92.0	92.3	90.2	86.1
		Entropy	93.1	**92.6**	92.3	**91.0**	86.0
		KM	93.2	92.5	91.9	90.6	86.4
		GE-KM	93.2	92.2	92.3	90.4	86.4
		peculiarity	**93.2**	**92.6**	**92.5**	**91.0**	**86.6**
	500-shot	Random	92.9	92.4	92.4	90.5	86.2
		Entropy	93.1	92.5	92.6	91.0	86.2
		KM	93.1	92.6	92.3	90.6	86.3
		GE-KM	93.2	92.4	92.8	89.9	**86.4**
		peculiarity	**93.4**†	**93.0**†	**92.9**	**91.3**†	**86.4**
	1,000-shot	Random	93.7	92.6	92.8	90.7	86.6
		Entropy	93.5	92.0	92.6	90.6	86.7
		KM	93.9	92.9	92.4	90.8	86.8
		GE-KM	93.5	93.0	92.7	90.8	86.1
		peculiarity	**94.1**†	**93.2**	**93.4**†	**91.2**†	**86.9**
MARC-5	100-shot	Random	60.3	55.3	55.2	53.8	50.5
		Entropy	60.2	55.6	54.9	52.7	49.0
		KM	60.0	55.7	55.4	54.0	50.1
		GE-KM	60.2	55.7	55.1	53.7	50.6†
		peculiarity	**60.6**†*	**56.2**†	**55.7**†	**54.3**†	49.9
		peculiarity-KM	60.1	56.1‡	55.5	54.2	**50.7***
	500-shot	Random	60.6	55.8	55.3	55.0†	51.4
		Entropy	60.5	55.7	55.7	52.4	50.0
		KM	60.8‡	56.3	55.9	54.4	51.3
		GE-KM	60.8‡	56.3	55.5	54.5	**51.7**†
		peculiarity	**61.1**†*	**56.5**	55.9	54.3	51.0
		peculiarity-KM	60.1	56.4	**56.0**	**56.6**‡*	**51.7***
	1,000-shot	Random	60.9	56.4	55.9	55.4	**51.9**†
		Entropy	60.8	56.0	55.8	53.1	50.7
		KM	61.2‡	56.5	56.0	55.6†	51.8
		GE-KM	61.0	56.2	56.0	55.1	**51.9**†
		peculiarity	**61.6**†*	**56.9**†*	56.5†	55.1	51.4
		peculiarity-KM	60.0	56.2	**56.6**‡	**55.8***	**51.9***

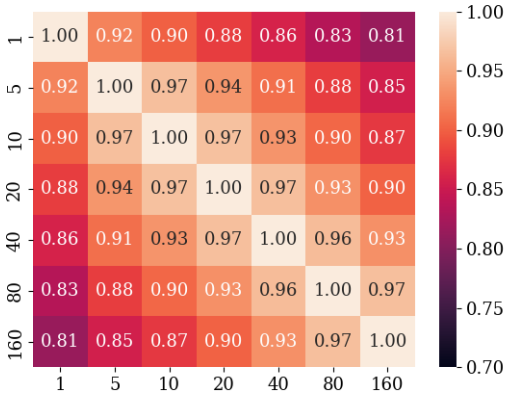
Table 16: Few-shot cross-lingual transfer performance on MARC.

		100-shot			500-shot			1,000-shot		
		All	Good	Poor	All	Good	Poor	All	Good	Poor
NER	Random	9.32	2.93	16.81	11.42	4.19	20.07	12.27	4.79	21.16
	Entropy	7.91	2.67	14.33	9.63	3.83	16.86	11.43	4.75	19.36
	KM	9.42	3.06	16.91	11.57	4.11	20.46	12.39	4.88	21.35
	LE-KM	8.82	2.53	16.17	11.19	3.99	19.76	12.38	4.86	21.30
	peculiarity	8.87	3.23	15.83	10.91	4.56	18.81	11.84	5.11	20.22
	peculiarity-KM	**9.88**	**3.76**	**17.00**	**12.19**	**5.09**	**20.54**	**13.13**	**5.50**	**22.27**
POS	Random	13.73	3.57	22.30	15.96	5.76	24.53	16.84	6.28	25.72
	Entropy	12.36	2.70	20.62	15.44	5.15	24.11	16.30	5.88	25.16
	KM	13.46	4.05	21.60	16.02	6.03	24.60	16.86	6.23	25.93
	LE-KM	13.89	4.15	22.25	16.28	5.71	**25.23**	16.94	6.24	25.92
	peculiarity	14.20	4.36	22.55	16.13	**6.51**	24.59	17.08	6.66	25.87
	peculiarity-KM	**14.26**	**4.47**	**22.68**	**16.35**	6.01	25.02	**17.26**	**6.75**	**26.17**
XNLI	Random	0.97	0.79	1.40	1.16	0.75	1.70	1.24	1.11	1.62
	Entropy	1.02	0.79	1.53	1.17	0.85	1.64	1.41	1.17	1.87
	KM	1.23	0.86	1.83	0.99	0.69	1.48	1.27	1.06	1.60
	GE-KM	0.94	0.79	1.24	0.89	0.51	1.32	1.39	0.94	1.92
	peculiarity	**1.31**	**1.04**	**1.84**	**1.25**	**0.94**	**1.75**	**1.52**	**1.35**	**1.95**
PAWS-X	Random	2.99	2.61	2.97	3.90	2.86	4.32	3.47	2.33	4.09
	Entropy	3.87	3.00	4.22	3.95	3.16	4.14	2.92	2.73	2.91
	KM	3.69	3.01	3.97	3.78	2.88	4.04	3.96	3.11	4.16
	GE-KM	3.83	2.85	4.24	3.66	2.95	3.79	3.67	3.20	3.72
	peculiarity	**4.01**	**3.45**	**4.28**	**4.23**	**3.45**	**4.52**	**4.18**	**3.45**	**4.36**
MARC-2	Random	0.54	0.17	1.10	0.80	0.45	1.32	1.09	0.77	1.57
	Entropy	0.96	0.45	**1.72**	0.68	0.43	1.05	1.15	0.87	1.57
	KM	0.71	0.25	1.35	0.79	0.42	1.35	1.16	0.77	1.75
	GE-KM	0.72	0.30	1.35	0.75	0.50	1.12	1.11	0.78	1.60
	peculiarity	**0.99**	**0.55**	1.10	**1.11**	**0.78**	**1.60**	**1.42**	**1.18**	**1.77**
MARC-5	Random	0.59	0.36	0.92	1.20	0.65	2.02	1.74	1.26	2.45
	Entropy	0.07	0.35	-0.35	0.45	0.76	-0.02	0.88	1.00	0.70
	KM	0.63	0.48	0.95	0.63	1.10	1.65	1.81	1.35	**2.50**
	GE-KM	0.66	0.46	0.85	**1.36**	0.96	**1.95**	1.61	1.16	2.27
	peculiarity	0.75	**0.91**	0.50	1.21	1.23	1.17	1.83	**1.71**	2.00
	peculiarity-KM	**0.92**	0.71	**1.22**	1.33	1.10	1.67	**1.88**	1.47	**2.62**

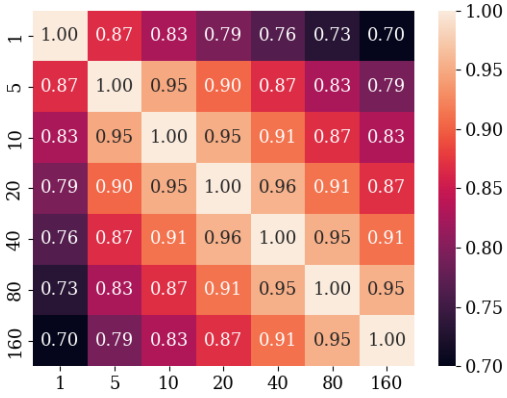
Table 17: Differences in accuracy between 0-shot and each n-shot model.



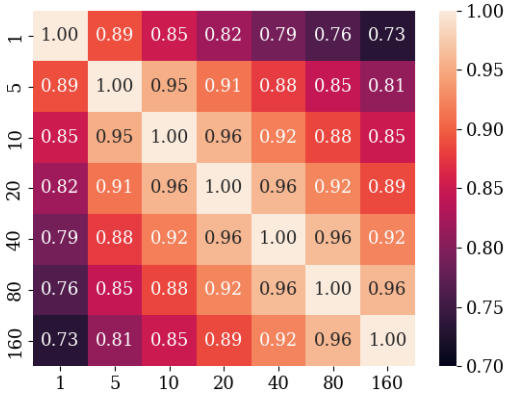
(a) NER



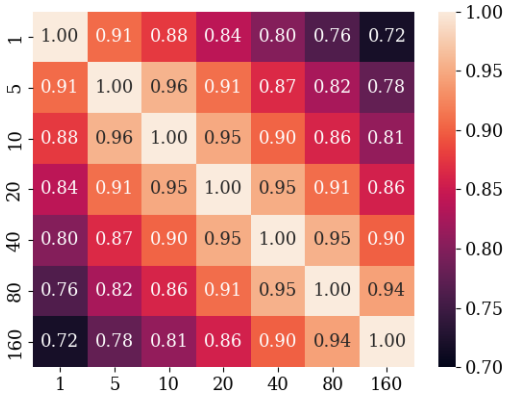
(b) POS



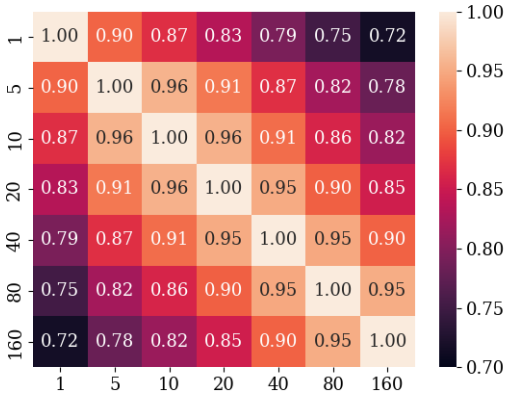
(c) XNLI



(d) PAWS-X



(e) MARC-2



(f) MARC-5

Figure 6: Overlap rates of the examples extracted by *peculiarity* measured at each k. They are averaged across all languages.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section named Limitations

☑ A2. Did you discuss any potential risks of your work?
Section named Ethics Statement

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Sections 1 and 8

☑ A4. Have you used AI writing assistants when working on this paper?
We use DeepL and Grammarly in all sections.

B ☑ Did you use or create scientific artifacts?

Sections 5, 6, and 7

☑ B1. Did you cite the creators of artifacts you used?
Section 5

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☐ 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)?
Not applicable. Left blank.

☐ 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?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Appendices A and B

☑ 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.
Appendix A

C ☑ Did you run computational experiments?

Sections 6 and 7

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 5

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 5

☑ 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?
Section 6 and 7

☑ 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 5

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.