“Diversity and Uncertainty in Moderation” are the Key to Data Selection for Multilingual Few-shot Transfer

Shanu Kumar[1] **Sandipan Dandapat**[1] **Monojit Choudhury**[2]
[1] Microsoft R&D, Hyderabad, India
[2] Microsoft Research, India
{shankum,sadandap,monojitc}@microsoft.com

Abstract

Few-shot transfer often shows substantial gain over zero-shot transfer (Lauscher et al., 2020), which is a practically useful trade-off between fully supervised and unsupervised learning approaches for multilingual pretained model-based systems. This paper explores various strategies for selecting data for annotation that can result in a better few-shot transfer. The proposed approaches rely on multiple measures such as data entropy using n-gram language model, predictive entropy, and gradient embedding. We propose a loss embedding method for sequence labeling tasks, which induces diversity and uncertainty sampling similar to gradient embedding. The proposed data selection strategies are evaluated and compared for POS tagging, NER, and NLI tasks for up to 20 languages. Our experiments show that the gradient and loss embedding-based strategies consistently outperform random data selection baselines, with gains varying with the initial performance of the zero-shot transfer. Furthermore, the proposed method shows similar trends in improvement even when the model is fine-tuned using a lower proportion of the original task-specific labeled training data for zero-shot transfer.

1 Introduction

Language resource distribution, for both labeled and unlabeled data, across the world’s languages is extremely skewed, with more than 95% of the languages having hardly any task-specific labeled data (Joshi et al., 2020). Therefore, *cross-lingual zero-shot transfer* using pretrained deep multilingual language models has received significant attention from the NLP community. During cross-lingual zero-shot transfer, first a multilingual model (Devlin et al., 2019; Conneau and Lample, 2019; Conneau et al., 2020; Liu et al., 2020; Xue et al., 2020; Ouyang et al., 2020) is created using only unlabelled data from a large number of languages (typically in the range of 100) with some self-supervised learning objectives. These pretrained models are then fine-tuned with task-specific labeled data from one or more languages (we refer to these as the *pivot* languages) and tested on all the other languages (here referred to as the *target* languages) for which no annotated data was used during fine-tuning.

Many recent work (Pires et al., 2019; Karthikeyan et al., 2019; Wu and Dredze, 2019; Artetxe et al., 2020; Lauscher et al., 2020) have studied the efficacy of zero-shot cross-lingual transfer across languages and factors influencing it. Other work have shown that a few-shot transfer, where very little labeled data in the *target* language is also used during fine-tuning, can result in substantial gains over the zero-shot transfer. For instance, Lauscher et al. (2020) show that zero-shot transfer does not hold much promise for transfer across typologically different languages or when there is not enough unlabeled data in the target language during model pretraining. In such cases, the gap in the cross-lingual transfer can be effectively reduced by fine-tuning it on a little annotated data in the target language. However, very few languages have readily available annotated resources for different NLP tasks, and collecting annotated data for a large set of target languages can be expensive and time-consuming (Dandapat et al., 2009; Sabou et al., 2012; Fort, 2016). Therefore, it is essential to carefully select and annotate target language data for a few-shot transfer, reducing the transfer gap effectively.

Training data selection has been investigated for several NLP tasks, especially for domain adaptation (Blitzer et al., 2007; Søgaard, 2011; Liu et al., 2019). The majority of these approaches use different techniques to rank the entire data and use top n data points to train the system (Moore and Lewis, 2010). In addition, active learning (Fu et al., 2013; Settles and Craven, 2008) has been widely used to

Findings of the Association for Computational Linguistics: NAACL 2022, pages 1042 - 1055

improve annotation efficiently by using model predictions to select informative data. Active learning is generally used in an iterative setting, in which a model is learned at each iteration, and samples are selected for labeling to improve performance. However, in this paper, we are trying to select a few samples. Hence we are limiting the training to one iteration. In the past, Chaudhary et al. (2019) have used active learning to annotate only uncertain entity spans for Dutch and Hindi languages. However, to the best of our knowledge, none of these approaches have been studied for a large set of languages in a cross-lingual few-shot transfer setting.

The central goal of this work is to propose specific strategies for data selection (and subsequent annotation) for few-shot learning so that the performance in a target language is maximized, given a data budget. The main contributions of this work are: [1] We propose different data selection strategies based on the notions of cross-entropy, predictive entropy, gradient embedding and loss embedding, and perform various reliability analyses of these strategies. [2] We conduct experiments on a set of 20 typologically diverse languages including some syntactically divergent from the pivot language – English & Chinese. [3] We propose a loss embedding-based method for sequence labeling tasks which incorporates both diversity and uncertainty sampling. [4] Through experiments on three NLP tasks, we show that embedding-based strategies perform consistently better than random data selection baselines, with gains varying with the initial performance of the zero-shot transfer. We also observe several language and data-size dependent trends in the performance across different data selection strategies. [5] Finally, we provide a concrete set of recommendations for data selection based on features such as zero-shot performance and the amount of unlabeled data available for a target-language.

The rest of the paper is organized as follows. The next section introduces the novel data sampling strategies. Section 3, 4, and 5 present the experimental setup, results and related research in the area, respectively. Concluding remarks are made in Section 6.

2 Data Sampling Strategies

Assuming we have a pre-trained multilingual language model and enough labelled data in a particular language such as English (EN) for fine-tuning on a task. We can measure the zero-shot performance on a set of target languages. We observe that zero-shot performances are not uniform and often vary with the tYpological similarity between the target and pivot language as stated by (Pires et al., 2019; Lauscher et al., 2020). Nevertheless, all the target languages show a drop in zero-shot performance compared to the performance achieved in the pivot language. Hence, there is a cross-lingual transfer gap for all the target languages. This gap can be attributed to the inherent linguistic property of the target languages however, Lauscher et al. (2020) have shown that the cross-lingual transfer gap can be reduced by fine-tuning on a little annotated data in a target language.

Consequently, few-shot performance can reduce the transfer gap for all the languages. Given a fixed budget, let say k examples, we want to maximize the few-shot performance in a target language by carefully choosing the effective k examples. To this end, we are proposing several data selection strategies in this section. We compare them with random sampling where k target language examples are randomly selected from task-specific fine-tuning data collection. Note that the sampling strategies are oblivious to the actual labels of the data points, as annotation would follow the data selection step in practice.

2.1 Data Cross-Entropy (DCE)

Cross-entropy (Moore and Lewis, 2010; Axelrod et al., 2011; Dara et al., 2014) has been widely used for domain adaptation by selecting in-domain data from a large non-domain-specific (contains both in- and out-domain data) corpus. In our scenario, like Dara et al. (2014), the target language labeled data acts as the large non-domain-specific corpus and using cross-entropy, novel and diverse data is selected from it. Assuming that there is little overlap between the tokens of the pivot and target language during the zero-shot cross-lingual transfer, we presume that *no* in-domain labeled data for a particular target language is available initially. We further assume that we have access to a non-domain-specific collection of data points, the entire target language unlabeled corpus (may or may not be distinct from the pretraining corpus).

First, two N-gram language models $\mathbf{M_I}$ and $\mathbf{M_O}$ are trained on the sentences selected L_I (initially an *empty set*) and sentences left in the target language corpus L_O (staring with the entire corpus), respectively. We use SRILM[1] (Stolcke, 2002) to build the N-gram (for N=3) language models. We do not want to select sentences which are similar to already picked L_I; hence we measure data cross entropy (DCE) and select sentences from L_O that have high entropy with respect to L_I and low entropy with respect to L_O.

$$H_I(x) = H(\mathbf{M_I}(x)) \tag{1}$$

$$H_O(x) = H(\mathbf{M_O}(x)) \tag{2}$$

$$\text{DCE}(x) = \frac{H_O(x)}{\sum_{s \in L_O} H_O(s)} - \frac{H_I(x)}{\sum_{s \in L_O} H_I(s)} \tag{3}$$

where $H(x)$ is the measure entropy of a sentence x using a N-gram language model. The size of L_O and L_I will vary across the iterations, therefore we appropriately normalize the entropy H_I and H_O for calculating cross-entropy.

Algorithm 1 Sentence Selection using DCE

Input: Target Language Corpus D^t, g, k

$L_I \leftarrow \{\}$, $L_O \leftarrow D^t$
while size$(L_I) < k$ AND $L_O \neq \phi$ **do**
  $\mathbf{M_I} \leftarrow$ TrainLM(L_I)
  $\mathbf{M_O} \leftarrow$ TrainLM(L_O)
  for each $s \in L_O$ **do**
    $H_I(s) \leftarrow H(\mathbf{M_I}(s))$
    $H_O(s) \leftarrow H(\mathbf{M_O}(s))$
  end for
  for each $s \in L_O$ **do**
    Calculate DCE(s)
  end for
  $L_g \leftarrow$ Select top g sentences ranked by DCE$(\cdot)$
  $L_I \leftarrow L_I \cup L_g$
  $L_O \leftarrow L_O - L_g$
end while

Algorithm 1 describes the data selection method using data cross-entropy, where g is the number of data points to be selected in one iteration, and k is the total number of sentences to be selected. The overall time complexity of this method is $\mathcal{O}(nk/g)$, where $n = |D^t|$. For reducing the computation time, we can increase g, which we set to 10 in our experiments.

[1] http://www.speech.sri.com/projects/srilm/

2.2 Predictive Entropy (PE)

We employ predictive entropy to measure the task-specific knowledge of a fine-tuned model. For a sequence labelling task, we define the predictive entropy $E(x_i)$ of a token x_i of a sentence x given a fine-tuned model $\mathbf{M}$ as follows:

$$p(y_i|x_i) = \mathbf{M}(x_i) \tag{4}$$

$$E(x_i) = -\sum_{j=1}^{C} p(y_i = c_j|x_i) * \log p(y_i = c_j|x_i) \tag{5}$$

where $c_1, c_2, \ldots c_C$ are the class labels.

We define the predictive entropy of the sentence using the equation (6):

$$\text{PE}(x) = \frac{1}{N_x} \sum_{i=1}^{N_x} E(x_i) \tag{6}$$

where N_x is the number of tokens in sentence x. For classification tasks, N_x will be 1.

To define the scoring function for data selection using predictive entropy that can generalize to the corpus with different domain-shift, we use the statistics of the predictive entropy from the entire target language corpus. We use μ_{PE} mean and σ_{PE} standard deviation of the predictive entropy of all the sentences in the corpus. Selecting sentences with very low predictive entropy will not help improve the performance as they have less novel information to enhance the knowledge of the model. Furthermore, picking sentences with very high predictive entropy can be harmful to training. It can be high due to either noise or out-of-domain data. As we want to select very few data instances for few-shot learning and improve further upon the zero-shot performance, we consider selection around μ_{PE}, the mean of the predictive entropy. But if the zero-shot performance is excellent, then μ_{PE} will be very low, and selecting data closer to mean may not improve over the zero-shot performance. Therefore, we add σ_{PE}. We formally define the scoring function in equation (7).

$$\text{score}_{PE}(x) = |\text{PE}(x) - (\mu_{PE} + \lambda * \sigma_{PE})| \tag{7}$$

Here, λ controls the distance of the preferred selection zone from μ_{PE}.

2.3 Gradient Embedding (GE)

Most of the data selection strategies use either representative sampling such as DCE or uncertainty

Task	Model	EN	AR △	BG △	DE △	EL △	ES △	EU △	FI △	FR △	HE △	HI △	JA △	KO △	RU △	SV △	SW △	TH △	TR △	UR △	VI △	ZH △
POS	B	96.4	-55.4	-11.4	-11.7	-18.9	-13.27	-37.4	-19.8	-	-46.9	-35.2	-53.0	-49.7	-12.2	-7.3	-	-	-26.3	-43.4	-	-38.5
	X	97.2	-43.6	-9.7	-9.9	-14.6	-13.1	-27.9	-14.7	-	-44.4	-27.6	-74.3	-46.7	-10.2	-6.3	-	-	-20.4	-37.2	-	-63.9
NER	B	84.2	-45.3	-7.4	-6.7	-12.8	-11.9	-24.3	-7.7	-5.7	-28.6	-20.2	-54.9	-25.2	-21.3	-9.9	-	-83.5	-12.4	-47.9	-11.3	-40.7
	X	82.5	-39.6	-5.7	-9.2	-9.8	-8.8	-23.9	–8.7	-5.6	-32.5	-15.3	-60.5	-36.6	-18.7	-13.4	-	-78.1	-8.6	-34.5	-16.4	-53.5
XNLI	B	81.9	-16.7	-13.2	-11.1	-14.8	-7.1	-	-	-7.8	-	-22.1	-	-	-12.9	-	-32.2	-28.8	-20.8	-24.3	-11.7	-13.2
	X	84.1	-12.7	-6.3	-8.3	-8.8	-5.7	-	-	-6.5	-	-15.0	-	-	-9.0	-	-20.4	-12.7	-12.0	-18.6	-9.9	-10.6

Table 1: We report the zero-shot cross-lingual transfer performance drops relative to EN for all the languages on POS, NER, and XNLI tasks with mBERT (**B**) and XLM-R (**X**). The results are medians over three RAND initialization (seeds). The darkness of the cell indicates the drops in zero-shot performance.

sampling such as PE. Recently, Ash et al. (2019) proposed BADGE that combines both diversity and uncertainty sampling. BADGE uses gradient embedding to capture uncertainty from the model, assuming the norm of the gradients will be smaller if the model is highly certain about its predictions and vice versa. As we don't have access to the ground truth labels, the gradient embedding $g_{x_i} \in \mathbb{R}^d$ is computed for a input sentence x_i by taking model's (M) prediction as the true label $\hat{y}_i$.

$$\hat{y}_i = \arg\max \mathbf{M}(x_i) \tag{8}$$

$$g_{x_i} = \frac{\partial}{\partial \theta_{out}} l_{CE}(\mathbf{M}(x_i), \hat{y}_i) \tag{9}$$

where l_{CE} is the cross-entropy loss function, $\theta_{out} \in \mathbb{R}^d$ refers to the parameters of last layer and d is the number of parameters. We have used hidden states of the [CLS] token from last layer classification tasks, hence we have computed the gradients with θ_{out} as the last layer of the pre-trained models.

BADGE selects samples by applying k-MEANS++ (Arthur and Vassilvitskii, 2006) clustering on the gradient embedding. The selection is made on the assumption that examples with gradient embedding of small magnitude will tend to cluster together and not be selected repeatedly. k-MEANS++ tend to select samples that are diverse and highly uncertain. For simplicity, we will call BADGE method as GE.

As we want to select very few data instances for few-shot learning and improve further upon the zero-shot performance, we consider applying GE selection on examples satisfying the following criteria:

$$\text{GE}(\lambda) = \text{GE}(\{x \colon g_x > \mu_g + \lambda * \sigma_g\}) \tag{10}$$

where μ_g and σ_g are the mean and standard deviation of magnitude of the gradient embedding of all the examples in the corpus. λ controls the final value of the selection criteria.

We noticed that in certain cases selecting samples sharing similar context but having different true labels may be more helpful for few-shot learning. To incorporate this, we propose **GE(γ)**, which adds γ similar examples for each k sample selected using the GE method. As gradient embedding loses information about the sentence, we use Multilingual Sentence XLM-R (Reimers and Gurevych, 2020) for calculating similarity based on sentences. We do not apply any constraints to ensure similar examples have different true labels but the gradient embedding can be used for ensuring it.

2.4 Loss Embedding (LE)

Sequence labelling tasks require prediction over all the tokens of a sentence, and therefore we have to calculate the gradient embedding for each token classification. Considering the maximum number of allowed tokens in a sentence to be m, the resulting gradient embedding g_{x_i} will of dimension $d \times m$. Due to its high dimensionality, applying k-MEANS++ will be expensive. We solve this dimensionality issue by proposing the Loss Embedding method, which has a dimension of m, considering lm is usually less than d.

Instead of calculating gradient, we consider only using classification loss at each token. For a sentence x_i, we compute loss embedding $l_{x_i} \in \mathbb{R}^m$ by computing cross-entropy loss for each token by taking the model's prediction as actual labels. As the norm of loss embedding will be smaller if the model is highly certain about its predictions and vice-versa, it satisfies the primary assumption of BADGE method. Another property preferable for sequential tasks is that the sentences with similar syntax will have a similar structure in the loss embedding as it depends upon the position of tokens in a sentence. Therefore applying k-MEANS++ clustering on the loss embedding will induce both diversity and uncertainty sampling.

Similar to GE, we also experiment with selection of examples satisfying the following criteria:

$$\text{LE}(\lambda) = \text{LE}(\{x\colon l_x > \mu_l + \lambda * \sigma_l\}) \qquad (11)$$

where μ_l and σ_l are the mean and standard deviation of magnitude of the loss embedding of all the examples in the corpus.

3 Experimental Setup

We conduct various experiments to evaluate effectiveness of our proposed data sampling techniques in a few-shot transfer setting with up to 20 languages from various language families on two different sequential tasks and one classification task.

3.1 Datasets

We evaluate our methods on three benchmarks datasets on POS-tagging, NER, and NLI. The complete statistics of training and test data available in each language is provided in Appendix A.

Named Entity Recognition (NER). We perform NER experiments using NER Wikiann dataset (Rahimi et al., 2019) on 20 languages. We also remove duplicates data points from the training corpus as these will hinder data selection.

Part-of-speech Tagging (POS). We perform POS experiments using Universal Dependency treebanks (Nivre et al., 2016) on the same set of languages of NER except French (FR), Thai (TH), and, Vietnamese (VI) due to unavailability of substantial amount of training data after removing duplicates.

Cross-lingual Natural Language Inference (XNLI). The XNLI dataset (Conneau et al., 2018) consists of translated train, dev and test sets in 14 languages of English hypothesis-premise pairs.

3.2 Training Details

We conduct all our experiment using the 12 layer multilingual mBERT *Base cased* (Devlin et al., 2019) and XLM-R (Conneau et al., 2020). We use the standard fine-tuning technique as described in (Devlin et al., 2019; Pires et al., 2019) for all the experiments. We limit the sentence length to 128 subword tokens and set the batch size as 32. Following (Lauscher et al., 2020), we fix the number of training epochs to 20 and the learning rate as 2.10^{-5} for NER and POS. For XNLI, we set the training epochs to 3 for zero-shot and 1 for few-shot training, and learning rate as 3.10^{-5}. We report F_1-score for NER and POS, and accuracy for XNLI. All the reported results are medians over three random initializations (seeds).

3.3 Zero-Shot Transfer

Throughout our experiments, we assume EN as the pivot language. We report the zero-shot cross-lingual transfer results in Table 1. We observe similar trends in zero-shot performance as reported in (Lauscher et al., 2020), where there are significant drops in performance for TH, JA, AR, ZH, UR, KO, VI. In TH, we observe the highest transfer gap with nearly 0 F_1-score, which indicates no cross-lingual transfer has happened.

3.4 Few-Shot Transfer

We add k additional examples from a target language and report the improvement of few-shot performance over the zero-shot performance reported in Section 3.3, where k examples are chosen according to the proposed strategies in Section 3, namely random sampling (RAND), DCE, PE, GE, and LE. We use similar training and evaluation setups for the few-shot transfer experiments as we used in the zero-shot setting and repeat the experiments with three random seeds. We consider three RAND baselines and report the average for all the data selection experiments.

4 Results

We calculated the difference between the F_1-scores of few-shot and zero-shot setups, *deltas*($\triangle$), for each language separately, but we observed different sampling strategies to work better depending upon the cross-lingual transfer gap. Therefore, we present the experimental results after categorizing languages by the transfer gap as indicated by the zero-shot performance, shown in Table 1. We categorize the languages in three groups: C_1, C_2 and C_3, and are coloured as light grey, dark grey and very dark grey respectively in Table 1. For NER task, groups are defined as $C_1 \in \{$BG, DE, EL, ES, EU, FI, FR, HI, RU, SV, TR, VI$\}$, $C_2 \in \{$AR, HE, JA, KO, UR, ZH$\}$, and $C_3 \in \{$TH$\}$. For POS, groups are defined as $C_1 \in \{$BG, DE, ES, EL, FI, RU, SV, TR$\}$, and $C_2 \in \{$AR, EU, HE, HI, JA, KO, JA, UR, ZH$\}$. For XNLI, the groups are different for XLM-R and mBERT, hence we have mentioned them in the Appendix.

We report the *deltas*, for NER and POS tasks in Table 2 and 3, respectively. The reported deltas are averaged across all the target languages for each language group. All the reported values are positive, which means in all cases, performance for the few-shot is higher than that for the zero-shot. The proposed methods require two parameters λ

		$k=10$			$k=50$			$k=100$			$k=500$			$k=1000$		
	Method	$\triangle^{C_1}$	$\triangle^{C_2}$	$\triangle^{C_3}$	$\triangle^{C_1}$	$\triangle^{C_2}$	$\triangle^{C_3}$	$\triangle^{C_1}$	$\triangle^{C_2}$	$\triangle^{C_3}$	$\triangle^{C_1}$	$\triangle^{C_2}$	$\triangle^{C_3}$	$\triangle^{C_1}$	$\triangle^{C_2}$	$\triangle^{C_3}$
mBERT	RAND	2.9	9.9	0.5	6.4	15.8	1.3	7.7	17.4	1.3	12.1	26.9	18.6	14.0	30.4	**31.2**
	DCE	1.8	8.4	4.0	5.1	12.8	2.8	5.2	11.8	3.3	9.9	23.0	18.8	12.3	28.5	29.2
	PE ($\lambda=1$)	3.1	10.0	**5.7**	6.8	14.5	**4.7**	7.5	16.1	**3.9**	12.4	24.5	**19.8**	14.2	27.4	27.4
	LE	5.5	**11.3**	2.5	7.4	**18.4**	1.1	**8.9**	**18.9**	0.7	**13.0**	**27.6**	18.9	**14.9**	**30.6**	31.0
	LE ($\lambda=0$)	**5.6**	11.0	0.7	**8.4**	18.3	0.1	8.7	18.4	-0.0	12.9	26.9	15.1	14.8	30.0	29.8
XLM-R	RAND	1.4	8.0	0.3	6.7	15.3	0.4	7.8	16.8	1.5	12.8	**26.1**	**20.7**	**14.6**	29.4	27.7
	DCE	-3.8	0.5	2.4	3.4	10.0	0.9	4.3	10.5	-0.2	10.5	23.8	19.2	13.2	27.8	26.3
	PE ($\lambda=1$)	**4.4**	**8.6**	3.6	7.0	**16.5**	1.2	7.9	**17.8**	0.5	12.3	**26.1**	19.0	14.2	**29.9**	**28.4**
	LE	3.0	8.1	**5.7**	**7.9**	15.7	3.4	**9.0**	16.7	2.4	**13.0**	**26.1**	18.2	14.5	29.1	23.4
	LE ($\lambda=0$)	2.4	8.2	1.4	7.4	16.0	**5.0**	8.5	16.9	**4.0**	**13.0**	**26.1**	16.5	14.5	29.3	23.2

Table 2: Few-shot cross-lingual transfer performance on NER tasks with varying number of target language examples k using EN as the pivot language. We have reported the $\triangle$ delta between few-shot and zero-shot performance averaged across the languages in each category C_1, C_2, and C_3.

		$k=10$		$k=50$		$k=100$	
	Method	$\triangle^{C_1}$	$\triangle^{C_2}$	$\triangle^{C_1}$	$\triangle^{C_2}$	$\triangle^{C_1}$	$\triangle^{C_2}$
mBERT	Rand	4.1	22.6	6.7	27.5	7.3	28.0
	DCE	2.3	18.7	5.2	24.3	6.0	25.9
	PE ($\lambda=1$)	4.4	**23.4**	**7.0**	27.8	7.4	28.1
	LE	3.9	20.1	6.3	26.3	7.1	26.9
	LE ($\lambda=0$)	**4.5**	21.9	6.8	27.3	**7.5**	27.9
	LE ($\lambda=0.5$)	4.1	23.3	6.8	**28.2**	**7.5**	**28.6**
XLM-R	RAND	3.1	24.6	5.2	28.5	5.6	28.8
	DCE	1.8	22.1	3.7	26.0	4.5	27.2
	PE ($\lambda=1$)	3.4	24.7	5.6	29.1	6.1	**29.2**
	LE	2.9	22.1	5.5	27.6	6.2	28.5
	LE ($\lambda=0$)	3.1	24.9	5.6	28.6	6.1	28.8
	LE ($\lambda=0.5$)	**3.5**	**25.2**	**5.8**	**29.2**	**6.4**	**29.2**

Table 3: Few-shot cross-lingual transfer performance on POS tasks with varying number of target language examples k using EN as the pivot language. We have reported the $\triangle$ delta between few-shot and zero-shot performance averaged across the languages in each category C_1 and C_2.

and γ for data selection. We perform experiments for $\lambda \in \{0, 0.5, 1\}$ and $\gamma \in \{1, 2, 3\}$, which are reported in Appendix B. In Table 2 and 3, we are reporting the best setups for PE and LE, where we observe the highest gains for NER and POS tasks, respectively. The methods based on PE and LE consistently outperform the baseline RAND and DCE for all values of k on POS task, and most of the cases in the case of NER. DCE performs worse compared to RAND for all the languages from groups C_1 and C_2. In general, the gains obtained through PE and LE compared to RAND are higher for C_1 than C_2. Similarly, the proposed approaches are more useful compared to RAND for small values k, and the advantage of our sampling strategies diminishes as k approaches to 1000. For TH ($\in C_3$), due to deficient zero-shot transfer performance in NER, the gains are not consistent across models. However, all three approaches outperform RAND for small values of k.

	Method	10	100	500	$1k$	$5k$	$10k$
mBERT	RAND	-0.5	-0.9	-0.2	0.6	2.4	3.4
	DCE	0.0	-0.1	-0.1	0.5	2.8	3.9
	PE ($\lambda=1$)	-0.4	-1.0	-0.3	0.4	2.6	3.4
	GE	-0.1	-0.9	-0.5	0.2	2.6	3.1
	GE ($\gamma=1$)	**0.1**	**-0.1**	**0.7**	**1.4**	**3.3**	**4.2**
XLM-R	RAND	-0.1	-0.2	0.0	0.2	1.2	1.5
	DCE	**0.2**	**0.7**	0.5	0.7	1.6	2.1
	PE ($\lambda=1$)	-0.4	-0.0	0.0	0.3	1.3	1.5
	GE	-0.3	-0.1	0.2	0.4	1.5	**2.6**
	GE ($\gamma=1$)	0.1	0.5	**0.6**	**1.2**	**1.9**	2.1

Table 4: Averaged few-shot performance on XNLI tasks with varying number of target language examples k using EN as the pivot language.

For XNLI, the averaged deltas across all languages are reported in Table 4. As DCE requires a sentence to train n-gram language model, hence we represent a sentence in XNLI by joining the hypothesis and the premise of an instance with a separator (-). The few-shot improvements are less pronounced than the sequence labeling tasks; noticeable gains start after seeing $k = 500$ target-language examples. As the size of the target-language corpus in XNLI is enormous compared to POS and NER, we also evaluated the methods for $k = 10000$. Surprisingly, GE ($\gamma = 1$) and DCE outperforms RAND. As DCE selects examples in batches of 10, it selects examples having similar contexts similar to GE ($\gamma = 1$), which benefits the few-shot learning. Since GE ($\gamma = 1$) also includes uncertainty sampling, it outperforms DCE for most of the values of k. Due to the large corpus size of XNLI, diversity becomes crucial during sampling.

We observe low few-shot gains for PE as it does not induce diversity. To measure the impact of pivot size, we trained a zero-shot model with 40k EN examples and observe similar trends for both DCE and GE (see Table 13 in Appendix).

Method	XNLI		NER		POS	
	C_1 (24)	C_2 (4)	C_1 (24)	C_2 (12)	C_1 (16)	C_2 (16)
PE	1	1	8	1	1	7
LE	-	-	**15**	**5**	**3**	**10**
GE	**18**	**2**	-	-	-	-

Table 5: Pairwise t-Test is performed using the proposed methods against RAND. We have reported the number of languages in each group having significance level of 0.1 using both XLM-R and mBERT models.

4.1 Effect of λ and γ parameters

In Appendix B, we have provided detailed results by varying λ and γ. For LE, a higher value of λ is required for the POS task due to the higher number of class labels than NER. The number of classes is 18 for POS and 7 for the NER task. Due to the higher number of class labels, the norm of loss embedding distribution has a higher tail. Hence, a higher value of λ is required for POS. We limited the value of λ to 0.5 as beyond that, very few examples were left for selection.

We incorporate γ parameter to include examples similar in context. As sentences with similar context will also have similar class labels in the case of POS and NER tasks, further decreasing the diversity in samples. Hence, we only consider experimenting with γ for XNLI. We observe that $\gamma = 1$ provides the best performance on average, suggesting that having two samples of similar contexts provides better few-shot learning.

4.2 Statistical Significance Test

We perform a pairwise t-Test for measuring the statistical significance of the proposed methods against the RAND baseline. We perform t-test for each language using both mBERT and XLM-R and have reported the number of languages having p-value less than the critical point (which is 0.1 in our case) for each language group. We have considered the following methods in our tests: GE ($\gamma = 1$) for XNLI, LE for NER, LE ($\lambda = 0.5$) for POS and PE ($\lambda = 1$) for all the tasks.

In Table 5, we notice that for XNLI, GE provides significant gains than RAND for 18 out of 24 cases from C_1 group, and 2 out of 4 cases from C_2 group. For NER, the gains are significant for 20 out of 32 cases while using LE, but only 9 cases have significant gains using PE. For POS, we observe LE provide significant gains for cases compared to PE. We can conclude that the embedding-based methods provide better gains than uncertainty-based methods for most languages.

	Method	10	50	100	500	1000
mBERT	RAND	3.9	8.8	10.5	18.6	21.2
	DCE	1.8	5.5	6.5	15.3	19.2
	LE	**5.0**	**11.0**	**12.2**	**19.8**	**22.0**
XLM-R	RAND	7.3	15.3	**17.1**	**26.5**	**29.3**
	DCE	-0.8	8.6	10.5	22.9	26.9
	LE	**8.7**	**15.9**	**17.1**	26.1	29.1

Table 6: $\triangle$ Delta between Few-shot and zero-shot performance on NER tasks using **ZH** as the pivot language, averaged across all languages.

5 Impact of Pivot Language

We conduct few-shot experiments considering ZH as the pivot language to validate the effectiveness of our method across different pivots. The delta between the gains using ZH as the pivot have been reported in Table 6 on the NER task. The delta has been averaged across all the languages. LE provides consistent gains over RAND, and gains saturate beyond 500 examples.

5.1 Embedding Visualization

We visualize the loss embeddings for DE language using t-SNE (Van der Maaten and Hinton, 2008) in Figure 1. Most of the samples using RAND (▼) tend to have lower norm of loss embedding, which may not be ideal for few-shot learning. We notice that examples having lower norm of loss embedding are clustered together and highlighted with ocean colour. Hence, samples selected via LE ($\times$) are more likely to have higher norm or higher uncertainty estimates. It is also evident that the samples from LE (cluster centre) will have higher diversity than RAND for few-shot learning.

5.2 Qualitative Analysis of Samples

We have compared sentences selected using RAND and LE for the NER task in Table 7. Random sampling has no constraints due to which it may select examples having very few entities which might not improve the few-shot performance. Since LE uses loss as the measure of uncertainty, it selects

	Raw Text	Translated Text
RAND	Er beschäftigt sich dort hauptsächlich mit dem Auswärtigen Amt und der SPD. In der weiblichen Hauptrolle ist Elżbieta Czyżewska zu sehen. Aus Asien in den Nordwesten , wo die Erfindung lange verharrte , dann an die Ostküste gelangte , um erst rund drei Jahrtausende später den Westen zu erreichen. Er lebt in Fernwald.	There he mainly deals with the Foreign Office and the SPD. Elżbieta Czyżewska stars in the female lead. From Asia to the northwest, where the invention remained for a long time, then reached the east coast, only to reach the west some three millennia later. He lives in Fernwald.
LE	Auch unter Trainer Joachim Löw behielt er den Posten des Managers. Maria de Lourdes Ruivo da Silva Matos Pintasilgo. Weiterleitung Atlantic Coast Hockey League. Chuck Weyant und Al Herman gingen mit Dunn-Rennwagen an den Start , wobei der 13.	He also retained the post of manager under coach Joachim Löw. Maria de Lourdes Ruivo da Silva Matos Pintasilgo. Forwarding Atlantic Coast Hockey League. Chuck Weyant and Al Herman competed in Dunn racers, with 13th.

Table 7: Samples in DE language from NER task using RAND and LE methods for $k = 10$ using XLM-R. Highlighted tokens are entities. We observe that LE tends to pick examples containing more entities than RAND.

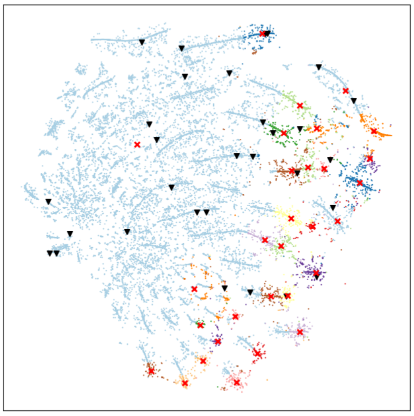

Figure 1: The t-SNE visualization of loss embedding using the complete DE language NER corpus with XLM-R. The clusters of loss embedding are highlighted with ×, while samples from RAND are highlighted with ▼.

Language	RAND	LE
ar	30.15%	**35.21%**
eu	16.53%	**20.17%**
he	34.11%	**34.77%**
hi	28.25%	**31.70%**
ko	31.65%	**34.64%**
ja	86.34%	**86.44%**
ur	**20.29%**	20.26%
zh	73.27	**77.52**

Table 8: Percentage of tokens from sentences sampled using RAND and LE($\lambda = 0.5$) for $k = 10$ from POS task. We calculated the percentage of tokens from class labels that are mispredicted using the zero-shot model for more than 40% on the whole language corpus. We observe that the LE method selects sentences containing more incorrect class labels without access to the ground truth labels. The languages from C_1 group are not considered as the gains are not relatively low.

sentences with a higher number of entities often miss-classified by the zero-shot model. Hence, LE will most probably improve the few-shot performance compared to RAND. Similarly, we observe in Table 8 that for POS, LE selects sentences containing class labels that are often incorrectly tagged by the zero-shot model. In the case of XNLI, we found that the GE-based method selects more competing examples (similar hypotheses for different premises leading to different labels), which effectively can enhance the model capability. Selected examples using different methods for POS and XNLI tasks can be found in Appendix C.

6 Related Work

6.1 Cross-lingual Transfer

In recent years, several pre-trained multilingual language models have been proposed including mBERT (Devlin et al., 2019), XLM (Conneau and Lample, 2019), XLM-R (Conneau et al., 2020), mBART (Liu et al., 2020), mT5 (Xue et al., 2020) and ERNIE-M (Ouyang et al., 2020) for cross-lingual transfer. Pires et al. (2019) show mBERT to have good zero-shot performance on NER and POS tagging tasks and attributed the effectiveness of transfer to the typological similarity between the languages. In contrast, several works (Karthikeyan et al., 2019; Wu and Dredze, 2019) have shown that cross-lingual transfer does not depend on subword vocabulary overlap and joint training across languages. Lauscher et al. (2020) empirically demonstrate that both pre-training corpora sizes and linguistic similarity are strongly correlated with the zero-shot transfer. Target languages with smaller pretraining corpora or higher linguistic dissimilarity with the pivot language have a low zero-shot transfer. Furthermore, they have shown that the gap can be reduced significantly by fine-tuning with a small number of target-language examples. Nooralahzadeh et al. (2020) study the cross-lingual transfer in meta-learning setting and demonstrate improvement in zero-shot and few-shot settings. While (Lauscher et al., 2020; Nooralahzadeh et al., 2020) focus on reducing zero-shot transfer gap using few-shot learning, in this work, we explore the data selection methods to get better cross-lingual transfer than the often used random sampling.

6.2 Training Data Selection

The problem of training data selection has been extensively studied for several NLP tasks, with the most notable ones from area of Machine Translation systems where target-domain data is limited and large non-domain-specific data is available. The task is to pick sentences that are closer to the target domain and also penalize the sentences which are out-of-domain. Moore and Lewis (2010) and Axelrod et al. (2011) address this problem by ranking sentences using the cross-entropy of target-domain-specific and non-domain-specific n-gram language models. Dara et al. (2014) employ an extension of the cross-entropy difference by including a vocabulary saturation filter which removes selection of very similar sentences. Song et al. (2012) have shown the effectiveness of cross-entropy selecting in-domain data for word segmentation and POS tagging tasks. We also use an extension of cross-entropy for selecting training data from the target language corpus for effective few-shot transfer using multilingual transformer models and compare with the proposed methods.

6.3 Active Learning

Active Learning has been widely used to reduce the amount of labeling to learn good models, (Yoo and Kweon, 2019; Fu et al., 2013). Uncertainty sampling methods have been commonly used in AL, where the most uncertain samples are selected for labeling. Various metrics have defined uncertainty using least confidence, sample margin, and predictive entropy. On the other hand, diversity sampling methods (Sener and Savarese, 2018; Gissin and Shalev-Shwartz, 2019) select examples which can act as a surrogate for the entire dataset. Chaudhary et al. (2019) used AL-based approaches to select entity spans for labeling in a cross-lingual transfer learning setting. However, this work was limited to only two languages. Our work focuses on data selection for cross-lingual transfer on a large and diverse set of target languages.

7 Discussion and Conclusion

This work explored various data sampling strategies for few-shot learning for two sequence labeling and a semantic tasks on 20 target languages. Our study shows that the embedding-based strategies, LE and GE, consistently outperform random sampling baseline across languages and sample sizes. Some of the salient observations are as follows. On NER and POS tasks, languages of the group C_2 show significant improvements in few-shot performance, suggesting that the gains from few-shot learning are strongly correlated to the zero-shot transfer gap. LE and GE-based data selection methods show consistent gains over the RAND strategy for each target language group, but these gains saturate as the sample size, k, increases beyond 500. The saturation occurs due to the relatively smaller target-language corpus size (varies between 5k and 20k for NER and POS, respectively) effectively reducing the diversity in the total sample. LE provides better few-shot performance than PE in terms of statistical significance. DCE only performs better than RAND for Thai. As DCE does not use any form of information from the fine-tuned model and if the target-language corpus size is small, it fails to select novel target language examples any better than RAND. However, in TH, for which zero-shot performance is close to zero, DCE selects the highly representative and diverse training examples for small values of k. The trends for XNLI are different from that of the sequence labeling tasks. GE and DCE outperform all other methods, with gains increasing with the value of k, which suggests that the size of the target-language corpus is crucial for data selection. XNLI has about 400k examples in each target-language corpus, much larger than that of NER and POS, signifying the importance of diversity sampling.

Based on our observations, we recommend the LE-based sampling strategy for data selection for cross-lingual few-shot transfer for sequence labeling tasks and GE-based sampling for classification tasks. While the optimal parameter setting for the LE sampling algorithm varies across tasks, we recommend the vanilla LE method without any parameter for most of the tasks. For tasks having higher number of class labels, we recommend using LE variant with λ such as 0 or 0.5.

In future the work can be extended to other high-level tasks such as cross-lingual QA and Machine translation. We would also like to extend this work in a reinforcement learning (Liu et al., 2019) or meta-learning (Tseng et al., 2020) framework, where the parameters can be automatically learnt for various tasks and settings.

References

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

David Arthur and Sergei Vassilvitskii. 2006. k-means++: The advantages of careful seeding. Technical report, Stanford.

Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2019. Deep batch active learning by diverse, uncertain gradient lower bounds. In *International Conference on Learning Representations*.

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 355–362, Edinburgh, Scotland, UK. Association for Computational Linguistics.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 440–447.

Aditi Chaudhary, Jiateng Xie, Zaid Sheikh, Graham Neubig, and Jaime G Carbonell. 2019. A little annotation does a lot of good: A study in bootstrapping low-resource named entity recognizers. 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 5164–5174.

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

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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.

Sandipan Dandapat, Priyanka Biswas, Monojit Choudhury, and Kalika Bali. 2009. Complex linguistic annotation – no easy way out! a case from Bangla and Hindi POS labeling tasks. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 10–18, Suntec, Singapore. Association for Computational Linguistics.

Aswarth Abhilash Dara, Josef van Genabith, Qun Liu, John Judge, and Antonio Toral. 2014. Active learning for post-editing based incrementally retrained mt. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 185–189.

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.

Karën Fort. 2016. Collaborative annotation for reliable natural language processing: Technical and sociological aspects.

Yifan Fu, Xingquan Zhu, and Bin Li. 2013. A survey on instance selection for active learning. *Knowledge and information systems*, 35(2):249–283.

Daniel Gissin and Shai Shalev-Shwartz. 2019. Discriminative active learning. *arXiv preprint arXiv:1907.06347*.

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.

K Karthikeyan, Zihan Wang, Stephen Mayhew, and Dan Roth. 2019. Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*.

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.

Miaofeng Liu, Yan Song, Hongbin Zou, and Tong Zhang. 2019. Reinforced training data selection for domain adaptation. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1957–1968.

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.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224, Uppsala, Sweden. Association for Computational Linguistics.

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).

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.

Xuan Ouyang, Shuohuan Wang, Chao Pang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Erniem: Enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora. *arXiv preprint arXiv:2012.15674*.

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.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525.

Marta Sabou, Kalina Bontcheva, and Arno Scharl. 2012. Crowdsourcing research opportunities: Lessons from natural language processing. In *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies*, i-KNOW '12, New York, NY, USA. Association for Computing Machinery.

Ozan Sener and Silvio Savarese. 2018. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*.

Burr Settles and Mark Craven. 2008. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 1070–1079.

Anders Søgaard. 2011. Data point selection for cross-language adaptation of dependency parsers. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 682–686.

Yan Song, Prescott Klassen, Fei Xia, and Chunyu Kit. 2012. Entropy-based training data selection for domain adaptation. In *Proceedings of COLING 2012: Posters*, pages 1191–1200, Mumbai, India. The COLING 2012 Organizing Committee.

Andreas Stolcke. 2002. Srilm - an extensible language modeling toolkit. In *INTERSPEECH*. ISCA.

Hung-Yu Tseng, Hsin-Ying Lee, Jia-Bin Huang, and Ming-Hsuan Yang. 2020. Cross-domain few-shot classification via learned feature-wise transformation. *arXiv preprint arXiv:2001.08735*.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

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.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Donggeun Yoo and In So Kweon. 2019. Learning loss for active learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 93–102.

A Data Statistics

We report the number of sentences in both training and test data in Table 9 and 10. POS task lower number of training data relative to NER task for most of the languages. XNLI task has enormous amount of training data compared to POS and NER.

Language	POS			NER		
	Train	Test	Language Group	Train	Test	Language Group
EN	11732	15039	-	19632	10000	-
BG	8736	1116	C_1	16235	10000	C_1
DE	149249	56354	C_1	18515	10000	C_1
ES	14092	1278	C_1	17817	10000	C_1
FI	14979	8233	C_1	18933	10000	C_1
FR	-	-	-	18109	10000	C_1
SV	3167	3000	C_1	14495	10000	C_1
TR	7745	9619	C_1	18433	10000	C_1
EL	1637	456	C_1	15908	10000	C_1
EU	5383	1799	C_2	8089	10000	C_1
HI	13291	2000	C_2	3948	1000	C_1
KO	22947	5563	C_2	18796	10000	C_2
RU	3841	3601	C_1	18795	10000	C_1
VI	-	-	-	16066	10000	C_1
AR	5956	2040	C_2	17703	10000	C_2
HE	5174	491	C_2	18329	10000	C_2
JA	7025	2172	C_2	19012	10000	C_2
UR	3892	535	C_2	13043	1000	C_2
ZH	3995	2451	C_2	18310	10000	C_2
TH	-	-	-	17683	10000	C_3

Table 9: We report the statistics of training and test data available in each language for our experiments.

Language	Train	Test	XLMR Group	mBERT Group
AR	392403	5010	C_1	C_1
BG	392335	5010	C_1	C_1
DE	392440	5010	C_1	C_1
EL	392331	5010	C_1	C_1
EN	392568	5010	C_1	C_1
ES	392405	5010	C_1	C_1
FR	392405	5010	C_1	C_1
HI	392356	5010	C_1	C_2
RU	392318	5010	C_1	C_1
SW	391819	5010	C_1	C_2
TH	392480	5010	C_1	C_2
TR	392177	5010	C_1	C_1
UR	388826	5010	C_1	C_2
VI	392416	5010	C_1	C_1
ZH	392251	5010	C_1	C_1

Table 10: We report the statistics of training and test data available in each language for XNLI.

B Study of λ and γ parameters on few-shot transfer

We conduct experiments using the following set of values for $\lambda \in \{0, 0.5, 1\}$ for NER and POS tasks. We have reported the results in Table 14 and 15. We find the parameter $\lambda = 1$ to be providing highest performance on average for PE, while $\lambda = 0.5$ show better performance for C_2 language group when $k = 10$. For LE methods, we observe $\lambda = 0.5$ provides the highest gains in POS tasks. However for NER task, LE method any λ parameter provides best gains on average. The gains start diminishing with higher λ values in general, but for C_2 language group, $\lambda = 0.5$ provides best gains for smaller values of k.

In Table 13, we observe that increasing the value γ beyond 1 hurts the performance for mBERT. $\gamma = 3$ provides higher gains in few cases for XLM-R. But overall, we consider $\gamma = 1$ to provide consistent gains across models.

C Qualitative Analysis of Samples

We have compared sentences selected using RAND and LE for POS task in Table 11. The comparison of examples from XNLI task selected using RAND and GE is shown in Table 12.

	Translated Text
RAND	Study: Domestic violence in the United States affects 25% of women and 7.5 of men The Lebanese authorities, led by the Command of the Emergency Force, contacted the Command and asked them to move to prevent and prevent the violations of Israel, which is carrying out works inside the Lebanese territories, including especially unloading sand or paving slopes, according to what was announced by an official source.
LE	It was also agreed to participate in the Maqama cultural festivals held on both sides by the two sides and to exchange and exchange musical, folk bands, artists, painters, playwrights, and others. Arafat is also scheduled to visit Kuala Lumpur, Jakarta, and Tokyo, according to Palestinian sources.

Table 11: Sample sentences in AR from POS task using RAND and LE ($\lambda = 0.5$) methods for $k = 10$ using XLM-R. The tokens are highlighted having ground truth class labels that are mispredicted using the zero-shot model. In case for AR, we noticed following class labeled are wrongly predicted: Other, Interjection, Particle, Adjective, Determiner, Pronoun and Adverb. LE ($\gamma = 0.5$) select sentences containing these class labels more frequently than RAND.

	Translated Premise	Translated Hypothesis
RAND	The 2000 census is the most important and provides valuable information.	Census Monitoring Board to oversee the 2000 decennial census.
	The guarantee is a really poor tree guarantee.	Yes I know the guy at wolfe told us they cut the tree warranty like six months or less.
	This drama isn't as interesting as nice celebrity meat, according to all the major gossip sites.	Who cares what David Hare did to Arthur Schnitzler's la ronde when there's celebrity meat to be suffered?
	I am a californian.	Oh yes yes yes i'm from i'm from up around i'm a new yorker myself.
	They would try to kill him like a pack of savages.	They would come down on him like a pack of blood thirsty wolves.
	Egyptians travel to surrounding countries to visit casinos.	Egyptian nationals are not allowed to gamble so casinos are only open to foreign guests over the age of 21 (you will be asked for ID).
GE	People from France are usually very boring.	For a people so proud of their identity, the French are a rich mix.
	The French are proud, dynamic patriots.	For a people so proud of their identity, the French are a rich mix.
	They like to sit in fancy places.	'Cause they're fancy places and stuff.
	That's because they're fancy seats and all.	'Cause they're fancy places and stuff.
	If they apologize for forgetting their name, the awkwawrdness will all go away.	If nothing is done to alleviate the situation , you can say bluntly , I 'm so sorry I 've forgotten her name.
	If you can't do better, just apologize for forgetting her name.	If nothing is done to alleviate the situation, you can say, bluntly, I'm so sorry I can't remember her name.

Table 12: Sample sentences in DE from XNLI task using RAND and GE ($\gamma = 1$) methods for $k = 10$ using XLM-R. We observe that GE ($\gamma = 1$) select two examples having similar context but different labels.

		$S = 400k$						$S = 40k$					
	Method	10	100	500	$1k$	$5k$	$10k$	10	100	500	$1k$	$5k$	$10k$
mBERT	RAND	-0.5	-0.9	-0.2	0.6	2.4	3.4	-1.2	-0.7	0.0	0.9	2.8	3.8
	DCE	0.0	-0.1	-0.1	0.5	2.8	3.9	0.2	-0.3	0.2	0.7	3.6	4.3
	PE ($\lambda = 1$)	-0.4	-1.0	-0.3	0.4	2.6	3.4	-	-	-	-	-	-
	GE	-0.1	-0.9	-0.5	0.2	2.6	3.1	-	-	-	-	-	-
	GE ($\gamma = 1$)	**0.1**	**-0.1**	**0.7**	**1.4**	**3.3**	**4.2**	**0.3**	**0.7**	**0.6**	**1.9**	**3.9**	**4.6**
	GE ($\gamma = 2$)	-0.3	-0.7	-0.1	1.3	3.1	4.1	-	-	-	-	-	-
	GE ($\gamma = 3$)	-0.6	-0.2	0.2	1.3	**3.3**	4.1	-	-	-	-	-	-
mBERT	RAND	-0.1	-0.2	0.0	0.2	1.2	1.5	-0.7	-0.6	-0.6	0.1	1.1	1.5
	DCE	0.2	**0.7**	0.5	0.7	1.6	2.1	-0.4	-0.9	-0.2	0.1	1.5	1.9
	PE ($\lambda = 1$) -0.4	-0.0	0.0	0.3	1.3	1.5	-	-	-	-	-	-	
	GE	-0.3	-0.1	0.2	0.4	1.5	**2.6**	-	-	-	-	-	-
	GE ($\gamma = 1$)	0.1	0.5	0.6	**1.2**	1.9	2.1	**-0.1**	**0.5**	**-0.1**	**0.6**	**1.6**	**2.0**
	GE ($\gamma = 2$)	0.4	0.3	0.7	1.3	2.1	2.3	-	-	-	-	-	-
	GE ($\gamma = 3$)	**0.7**	0.3	**1.1**	1.0	**2.1**	2.2	-	-	-	-	-	-

Table 13: Few-shot performance on XNLI tasks with varying number of target language examples k using EN as the pivot language. We have reported the $\triangle$ delta between few-shot and zero-shot performance averaged across all languages. S denotes the size of the pivot-language corpus.

		$k = 10$			$k = 50$			$k = 100$			$k = 500$			$k = 1000$		
	Method	$\triangle^{C_1}$	$\triangle^{C_2}$	$\triangle^{C_3}$	$\triangle^{C_1}$	$\triangle^{C_2}$	$\triangle^{C_3}$	$\triangle^{C_1}$	$\triangle^{C_2}$	$\triangle^{C_3}$	$\triangle^{C_1}$	$\triangle^{C_2}$	$\triangle^{C_3}$	$\triangle^{C_1}$	$\triangle^{C_2}$	$\triangle^{C_3}$
mBERT	RAND	2.9	9.9	0.5	6.4	15.8	1.3	7.7	17.4	1.3	12.1	26.9	18.6	14.0	30.4	31.2
	DCE	1.8	8.4	4.0	5.1	12.8	2.8	5.2	11.8	3.3	9.9	23.0	18.8	12.3	28.5	29.2
	PE ($\lambda = 0$)	3.9	12.0	1.6	6.3	16.7	0.4	7.6	17.8	0.3	12.2	27.1	18.3	13.7	30.0	**32.5**
	PE ($\lambda = 0.5$)	4.4	13.2	0.2	7.6	18.1	1.0	8.2	18.8	0.0	12.5	**27.8**	18.5	14.2	30.3	32.1
	PE ($\lambda = 1$)	3.1	10.0	**5.7**	6.8	14.5	**4.7**	7.5	16.1	**3.9**	12.4	24.5	**19.8**	14.2	27.4	27.4
	LE	5.5	11.3	2.5	7.4	**18.4**	1.1	**8.9**	18.9	0.7	**13.0**	27.6	18.9	14.9	**30.6**	31.0
	LE ($\lambda = 0$)	**5.6**	11.0	0.7	**8.4**	18.3	0.1	8.7	18.4	-0.0	12.9	26.9	15.1	14.8	30.0	29.8
	LE ($\lambda = 0.5$)	4.4	**13.9**	0.4	7.6	17.3	-0.1	8.6	**19.1**	-0.4	12.7	27.5	12.9	**15.0**	30.0	27.9
XLM-R	RAND	1.4	8.0	0.3	6.7	15.3	0.4	7.8	16.8	1.5	12.8	**26.1**	**20.7**	**14.6**	29.4	27.7
	DCE	-3.8	0.5	2.4	3.4	10.0	0.9	4.3	10.5	-0.2	10.5	23.8	19.2	13.2	27.8	26.3
	PE ($\lambda = 0$)	1.1	8.5	1.5	6.2	14.7	**7.3**	7.8	15.6	6.4	**13.0**	25.3	22.1	14.7	29.0	28.1
	PE ($\lambda = 0.5$)	2.5	10.9	-1.2	7.6	15.4	2.3	8.6	17.0	-1.0	**13.0**	25.4	20.1	14.9	28.7	**28.8**
	PE ($\lambda = 1$)	**4.4**	8.6	3.6	7.0	16.5	1.2	7.9	**17.8**	0.5	12.3	**26.1**	19.0	14.2	**29.9**	28.4
	LE	3.0	8.1	**5.7**	**7.9**	15.7	3.4	**9.0**	16.7	2.4	**13.0**	**26.1**	18.2	14.5	29.1	23.4
	LE ($\lambda = 0$)	2.4	8.2	1.4	7.4	16.0	5.0	8.5	16.9	**4.0**	**13.0**	**26.1**	16.5	14.5	29.3	23.2
	LE ($\lambda = 0.5$)	2.4	**11.0**	2.8	7.6	**16.6**	5.2	8.8	16.0	2.5	12.6	25.9	16.3	14.4	28.6	22.8

Table 14: Few-shot cross-lingual transfer performance on NER tasks with varying number of target language examples k using EN as the pivot language. We have reported the $\triangle$ delta between few-shot and zero-shot performance averaged across the languages in each category C_1 and C_2, and C_3.

		$k = 10$		$k = 50$		$k = 100$		$k = 500$		$k = 1000$	
	Method	$\triangle^{C_1}$	$\triangle^{C_2}$	$\triangle^{C_1}$	$\triangle^{C_2}$	$\triangle^{C_1}$	$\triangle^{C_2}$	$\triangle^{C_1}$	$\triangle^{C_2}$	$\triangle^{C_1}$	$\triangle^{C_2}$
mBERT	Rand	4.1	22.6	6.7	27.5	7.3	28.0	11.9	46.5	12.5	48.4
	DCE	2.3	18.7	5.2	24.3	6.0	25.9	11.8	46.1	12.4	48.3
	PE ($\lambda = 0$)	4.9	22.5	6.9	27.7	7.4	28.1	10.2	38.8	10.9	40.5
	PE ($\lambda = 0.5$)	**5.2**	22.5	7.0	28.0	**7.5**	28.3	10.4	38.9	11.2	40.5
	PE ($\lambda = 1$)	4.4	**23.4**	**7.0**	27.8	7.4	28.1	10.5	38.8	11.2	40.4
	LE	3.9	20.1	6.3	26.3	7.1	26.9	**12.1**	46.9	**12.8**	48.7
	LE ($\lambda = 0$)	4.5	21.9	6.8	27.3	**7.5**	27.9	**12.1**	**47.0**	12.0	48.7
	LE ($\lambda = 0.5$)	4.1	23.3	6.8	**28.2**	**7.5**	**28.6**	11.3	**47.0**	11.5	**48.8**
XLM-R	RAND	3.1	24.6	5.2	28.5	5.6	28.8	9.5	42.3	9.8	44.6
	DCE	1.8	22.1	3.7	26.0	4.5	27.2	9.4	42.0	9.8	44.5
	PE ($\lambda = 0$)	4.0	24.7	5.6	28.6	5.9	29.0	8.5	39.4	8.9	41.3
	PE ($\lambda = 0.5$)	**4.1**	25.1	**5.9**	28.8	6.2	**29.2**	8.8	39.4	9.1	41.2
	PE ($\lambda = 1$)	3.4	24.7	5.6	29.1	6.1	**29.2**	8.8	39.2	9.1	41.1
	LE	2.9	22.1	5.5	27.6	6.2	28.5	**9.6**	42.6	**9.9**	**44.9**
	LE ($\lambda = 0$)	3.1	24.9	5.6	28.6	6.1	28.8	9.5	42.7	9.3	**44.9**
	LE ($\lambda = 0.5$)	3.5	**25.2**	**5.8**	**29.2**	**6.4**	**29.2**	9.0	**42.8**	8.9	**44.9**

Table 15: Few-shot cross-lingual transfer performance on POS tasks with varying number of target language examples k using EN as the pivot language. We have reported the $\triangle$ delta between few-shot and zero-shot performance averaged across the languages in each category C_1 and C_2.