Bridging Language Gaps: Enhancing Few-Shot Language Adaptation

Anonymous ACL submission

Abstract

The disparity in language resources presents a significant challenge in multilingual NLP. High-resource languages have the advantage of extensive data and advanced models, whereas low-resource languages often lack the data necessary for effective training. Our research aims to bridge this gap by utilizing the task-specific knowledge from multilingual models trained on high-resource languages and efficiently transferring it to lower-resource languages. Our focus is on task-specific fine-tuning, the stage where acquiring labeled data is costly and where methods that optimize data use are most valuable. The primary advantage of our approach is its data efficiency, enabling rapid adaptation to new languages in a task-specific few-shot setting and reducing the need for large labeled datasets. Our Contrastive Language Alignment with Prompting (CoLAP) method integrates contrastive learning with cross-lingual representations, facilitating knowledge transfer from high-resource to lower-resource languages. This effectively narrows the cross-lingual performance gap, contributing to the development of more efficient multilingual NLP techniques.[1]

1 Introduction

The adaptation of pretrained language models (PLMs) to specific downstream tasks is typically resource-intensive, requiring extensive labeled data and computational resources. This becomes particularly challenging in multilingual contexts, especially for languages not represented in the model's pretraining data. These low-resource languages are significantly underrepresented, primarily due to the scarcity of large-scale corpora necessary for self-supervised pretraining and the lack of labeled data required for task-specific fine-tuning (Lauscher et al., 2020; Wu and Dredze, 2020; Yang et al., 2022). Multilingual pretrained language models (mPLMs), such as mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), and mT5 (Xue et al., 2021), have been developed to address these challenges. These models are trained across a wide range of languages using self-supervised data within a unified embedding space, facilitating cross-lingual transfer (XLT) of information. They have shown remarkable results in multilingual tasks. Despite their broad versatility, mPLMs often exhibit representation degradation, particularly in low-resource languages (Yang et al., 2022). This degradation stems from the skewed distribution of self-supervised training data, which disproportionately favors high-resource languages. As a result, the representational quality for low-resource languages, and particularly for those not included in the pretraining data, is substantially diminished, leading to a decline in downstream task performance (Winata et al., 2022).

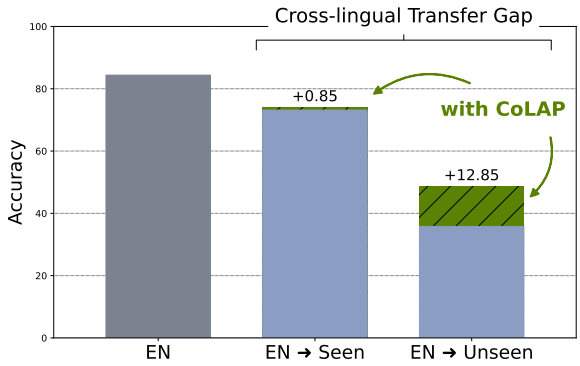

Figure 1: Performance gap for cross-lingual transfer in XLM-R from English to languages seen and unseen during pretraining based on XNLI and AmNLI. Performance gains achieved by few-shot adaptation using our CoLAP method are highlighted in green.

To enhance model representations and downstream task performance, models are commonly fine-tuned using task-specific data. However, acquiring such labeled data at scale is a significant

[1]The code repository is available at https://anonymous.4open.science/r/CoLAP-B2BE.

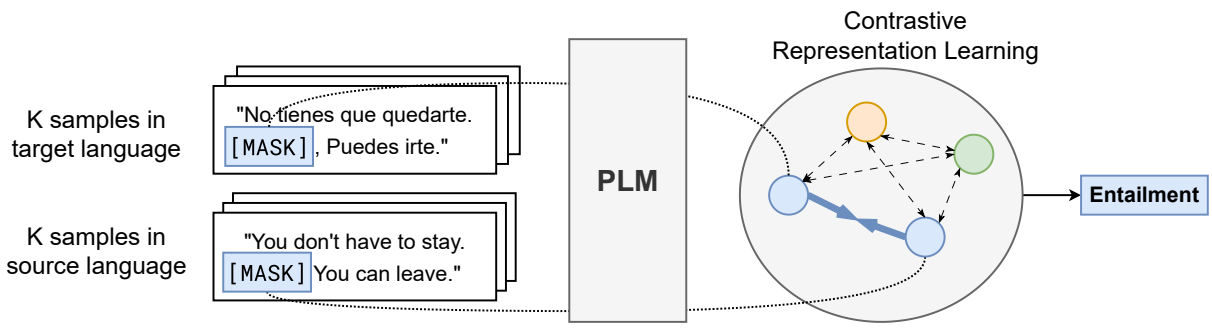

Figure 2: Illustration of our contrastive language alignment with prompting (CoLAP) approach for few-shot cross-lingual transfer applied to the natural language inference task.

challenge, especially for languages already disadvantaged by insufficient pretraining data (Winata et al., 2022). While mPLMs have advanced performance for these languages compared to monolingual models primarily trained on English data, XLT remains challenging as there are substantial performance disparities between different languages. This disparity becomes more evident when fine-tuning mPLMs on task data using high-resource languages and subsequently evaluating the fine-tuned model on other languages (Guo et al., 2023). We display the cross-lingual transfer gap visually for two of the datasets investigated in this study in Figure 1. Few-shot learning techniques address this gap, enabling models to learn effectively from limited data. Recent works in this domain encompass a range of zero-shot cross-lingual transfer (ZS-XLT) and few-shot cross-lingual transfer (FS-XLT) methods, including full fine-tuning (Lauscher et al., 2020), in-context learning (Winata et al., 2022), prompting (Schick and Schütze, 2021), representation mixup (Yang et al., 2022; Xu et al., 2023), and contrastive learning (Chi et al., 2021). Among these, prompting techniques have shown efficacy in low-resource settings. Prompts express tasks as language modeling problems without introducing new model parameters, often articulated as natural language sentences (Qi et al., 2022; Nie et al., 2023). Despite these advances, full fine-tuning remains standard practice in the field (Zhao and Schütze, 2021; Zhou et al., 2023).

In our study, we propose contrastive language alignment with prompting (CoLAP), addressing the cross-lingual transfer gap by aligning representations of underrepresented languages with those from high-resource languages, especially English. Instead of aligning representations during language modeling, we focus on the application of pretrained models on specific downstream tasks. We hypothesize that the transfer of discriminative task-specific information from English to lower-resource languages can be achieved efficiently without relying on abundant self-supervised or task-specific labeled data. This builds upon prior research by Chi et al. (2021) and Yang et al. (2022), which demonstrated the effectiveness of such alignment in general domain representations and ZS-XLT. We posit that representations for downstream classification tasks are less complex compared to representations used for language modeling and can therefore be transferred more data-efficiently between languages. In our contrastive learning approach, we align vector representations derived from prompting mPLMs, aiming to increase their similarity with representations from high-resource languages like English. This effectively leverages the discriminatory power of high-resource language representations to enhance performance in lower-resource languages. Our experiments focus on challenging multilingual tasks including natural language inference and relation extraction, across 27 languages. We evaluate our models in two distinct FS-XLT settings: (i) fine-tuning in a high-resource source language followed by few-shot adaptation to the target language, and (ii) few-shot training directly in the target language without prior task fine-tuning. Our methodology aims to narrow the cross-lingual performance gap, thereby extending the benefits of NLP models to a broader range of languages. We summarize our key contributions as follows:

1. We introduce a novel few-shot cross-lingual transfer approach, leveraging contrastive learning to efficiently transfer knowledge from high- to lower-resource languages.
2. Our method is applicable for both English fine-tuned models and those without prior task-specific training.

③ CoLAP enhances cross-lingual transfer performance for low-resource languages, including those not represented in mPLM pretraining.

④ We propose a few-shot exemplar selection approach based on representation similarity improving data efficiency.

2 Related Work

Cross-lingual transfer involves transferring knowledge from a source language to one or more target languages (Ruder et al., 2019). It is grounded in pretraining language models on multilingual data, embedding languages within a shared universal space to facilitate cross-lingual information transfer. This approach is particularly effective for languages featured in the pretraining dataset (Ruder et al., 2019; Ansell et al., 2021). mPLMs have extended this capability, showing promise in ZS-XLT even for languages not directly seen during pretraining (Devlin et al., 2019; Conneau et al., 2020). The transfer's efficacy is influenced by the alignment quality of representations between source and target languages (Wu and Dredze, 2020; Cao et al., 2020), which is especially effective for topologically similar languages with large-scale pretraining corpora (Lauscher et al., 2020). Recent advancements in enhancing XLT performance include projecting target language representations into the English space (Yang et al., 2022; Xu et al., 2023), and employing data augmentation (Zheng et al., 2021). Notably, Chi et al. (2021) improve cross-lingual representation alignment in their InfoXLM mPLM, employing contrastive learning and large-scale text corpora during pretraining.

Few-shot Learning and FS-XLT involve training models with a minimal number of labeled examples, challenging models to generalize from this limited data (Nie et al., 2023). Techniques explored for FS-XLT include fine-tuning (Lauscher et al., 2020), in-context learning (Winata et al., 2022), and prompting (Schick and Schütze, 2021). Prompting has shown notable performance, mitigating hyperparameter sensitivity and performance variance issues prevalent in fine-tuning and in-context learning techniques (Zhao et al., 2021; Schmidt et al., 2022; Winata et al., 2022). Enhancements to prompting techniques have been made through the use of multilingual templates (Qi et al., 2022) and label words (Huang et al., 2022; Zhou et al., 2023), further improving their applicability in multilingual settings. Schmidt et al. (2023) note the difficulty in comparing studies due to varying few-shot settings. They suggest that averaging model checkpoints during the few-shot fine-tuning phase provides a robust baseline for FS-XLT.

3 Methodology

Our methodology is predicated on the hypothesis that the cross-lingual performance gap in downstream tasks for low-resource languages is largely attributable to their underdeveloped representation spaces within mPLMs. Building on previous studies, we propose that aligning the representation spaces of low-resource languages with the more robust English representation space can significantly enhance task performance (Yang et al., 2022; Xu et al., 2023). This approach aims to transfer the discriminatory information embedded within the English representation space to these lower-resource languages. Prior research, including the works of Chi et al. (2021) and Yang et al. (2022), supports the effectiveness of this alignment during language modeling. However, these methods often rely on large-scale parallel corpora for extensive pretraining, making them infeasible for low-resource scenarios. To address this, we introduce CoLAP, a method specifically designed for few-shot cross-lingual transfer in low-resource contexts. CoLAP avoids introducing additional model parameters through prompting, and does not require large-scale labeled or self-supervised corpora. Our approach includes a task-agnostic representation contrastive learning objective, $XRCL$, which aligns multilingual vector representations of translated inputs. Unlike InfoXLM's contrastive learning objective, our method does not use a memory queue for negative pairs or mixup sampling, enhancing resource efficiency (Chi et al., 2021). In addition, we introduce a classification task-specific contrastive learning objective, $XCCL$, designed to transfer discriminative, class-specific features from high-resource to lower-resource languages, alleviating the dependency on parallel training datasets. Importantly, the additional computational complexity introduced by our contrastive learning objectives is dependent only on the number of samples per batch. Both CoLAP strategies are model-agnostic, making them applicable to any mPLM.

3.1 Prompt-based Training

In prompt-based training approaches, tasks are reformulated as language modeling problems, where

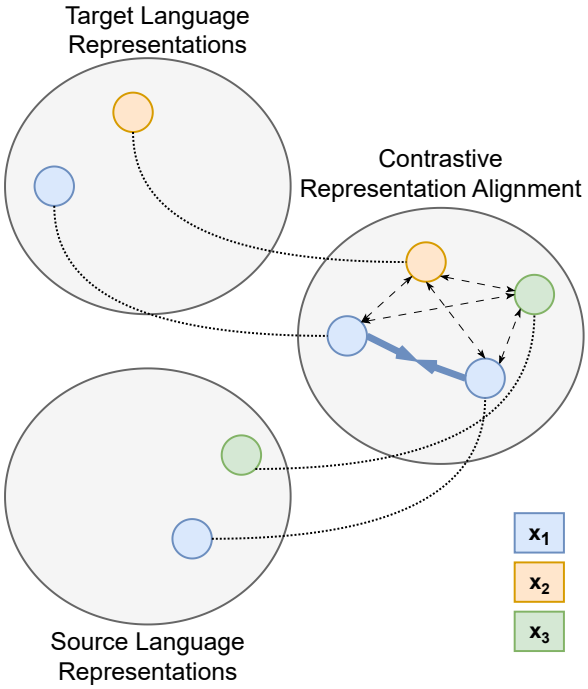

Figure 3: Illustration of cross-lingual contrastive representation alignment using the $XRCL$ objective.

the model predicts tokens that serve as reference labels (Schick and Schütze, 2021). For this approach, a template $\mathcal{T}$ is employed, which transforms inputs x into prompted formats, denoted as $x_{\text{prompt}} = \mathcal{T}(x)$. For language models trained on masked language modeling each prompt contains at least one `[MASK]` token. For models utilizing causal masking, the sentence representation can be derived from the hidden state of the last token. The language model predicts the token based on its context within x_{prompt}. To facilitate this prediction, labels are mapped to tokens in the model's vocabulary, represented as $\mathcal{R} \mapsto \mathcal{V}_{\mathcal{M}}$. Given an input x, the probability of predicting label y is denoted as:

$$p(y \mid x) = p(\texttt{[MASK]} = w_v \mid x_{\text{prompt}}) = \frac{\exp\left(w_v \cdot h_{\texttt{[MASK]}}\right)}{\sum_{v' \in \mathcal{V}} \exp\left(w_{v'} \cdot h_{\texttt{[MASK]}}\right)},$$

where $h_{\texttt{[MASK]}}$ is the hidden vector of the `[MASK]` token, and w_v denotes the pre-softmax vector corresponding to $v \in \mathcal{V}$ (Schick and Schütze, 2021). We employ English label words and language agnostic prompt templates, included in Appendix A.

3.2 Contrastive Learning Objectives

The **cross-lingual representation contrastive loss** term, abbreviated as $\mathcal{L}_{XRCL}$, aims to align the latent representations of instances in the target language with their counterparts in the source language. To achieve this, we utilize a model $\mathcal{M}$ to generate vector representations $r_i = \mathcal{M}(x_i)$ for each input sequence x_i. These sequences, along with their corresponding class labels y_i, are sampled from the training dataset. We obtain `[MASK]` representations by prompting XLM-R[2]. We treat the target language training dataset (D_T) and the source language training dataset (D_S) as parallel corpora. This implies that every instance present in D_T has a corresponding instance in D_S.

For a given instance representation in the target language $r_{i,T}$, its direct translation in the source language $r_{i,S}$, forms a positive pair. Consequently, all other instances within the target language dataset are treated as negative pairs. This is formulated as follows:

$$r_i^+ = \{r_j \mid j = i, j \in D_S\}$$
$$r_i^- = \{r_j \mid j \neq i, j \in D_S\}$$

The $XRCL$ objective maximizes the similarity between these cross-lingual instance representations (positive pairs) while minimizing their similarity to other instances (negative pairs) (van den Oord et al., 2019). We compute the cross-lingual representation contrastive loss, $\mathcal{L}_{XRCL}$, for each instance representation in the target language $r_{i,T}$ as follows:

$$\mathcal{L}_{XRCL} = \sum_{i=1}^{N} -log \frac{exp\left(\phi(r_{i,T}, r_i^+)/\tau\right)}{exp\left(\phi(r_{i,T}, r_i^-)/\tau\right)}, \quad (1)$$

where N is the number of instances in the target language's training set D_T. The parameter τ is a temperature scaling factor, and $\phi(r_i, r_i^-)$ calculates the cosine similarity between representation r_i and each representation in r_i^- using the formula $\sum_{j=1} r_i \cdot r_j / |r_i||r_j|$.

We also introduce the **cross-lingual class contrastive loss**, denoted as $\mathcal{L}_{XCCL}$. This loss is a variant of the previously discussed $\mathcal{L}_{XRCL}$, designed to align latent representations of instances within the same classes during fine-tuning for downstream tasks. Specifically, for each representation $r_{i,T}$ from the target language's training dataset, we construct positive pairs with instances of the same class included in the training dataset of the source language. Conversely, we consider instances from different classes in the source language negative

[2]This can straightforwardly be transferred to other sentence representations including the beginning-of-sequence token `[s]` or end-of-sequence token `[/s]` in causal masked models.

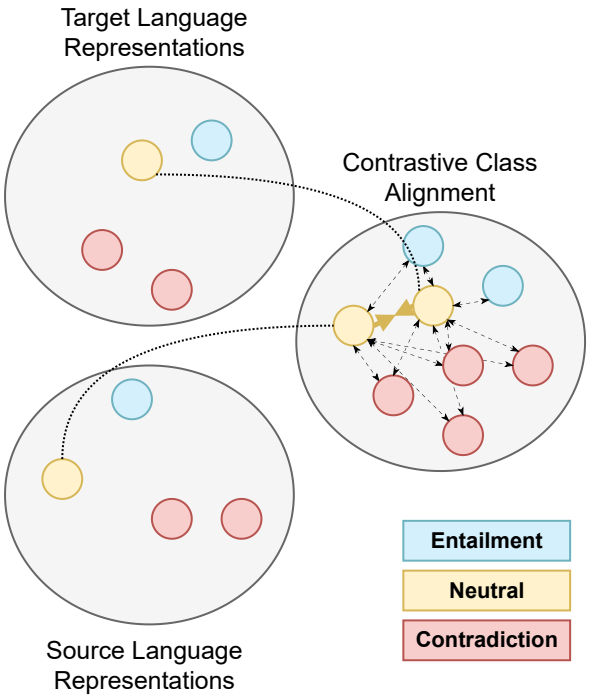

Figure 4: Illustration of cross-lingual class contrastive representation alignment using the $\mathcal{L}_{XCCL}$ objective.

pairs. The formulation is as follows:

$$r_i^+ = \{r_j \,|\, y_j = y_i, j \in D_S\}$$
$$r_i^- = \{r_j \,|\, y_i \neq y_j, j \in D_S\}$$

This approach aims to maximize the similarity between cross-lingual instance representations within the same class while minimizing similarity with different class representations. The key distinction of $\mathcal{L}_{XCCL}$ from $\mathcal{L}_{XRCL}$ lies in its assumption that shared class-specific information exists and can effectively be leveraged across languages. It posits that embeddings from instances within the same class should exhibit higher similarity compared to those from instances of different classes. The computation of the class-contrastive loss is analogous of Equation 1, as shown below:

$$\mathcal{L}_{XCCL} = \sum_{i=1}^{N} -log \frac{exp\left(\phi(r_{i,T}, r_i^+)/\tau\right)}{exp\left(\phi(r_{i,T}, r_i^-)/\tau\right)} \quad (2)$$

4 Experiments

4.1 Models and Training

We investigate two distinct FS-XLT settings. In the sequential setting abbreviated as `Source`→`Target`, models are initially fine-tuned on labeled data in a high-resource source language, typically English. During this phase, we employ the cross-entropy loss $\mathcal{L}_{CE} = -\log(z_y)$, where z_y represents the post-softmax probability of the vocabulary token corresponding to label y. Subsequently, the model is adapted to the target language under FS-XLT. While this two-step approach requires full fine-tuning on source language data, the resulting model can be used for various target languages with FS-XLT (Schmidt et al., 2022). In contrast, the `Target` setting directly adapts the model to the target language without any prior training on the task-specific data in the source language. During few-shot adaptation in both `Source`→`Target` and `Target` settings, we integrate our contrastive loss terms. For the variant incorporating the cross-lingual representation contrastive objective $XRCL$, the total loss is calculated as $\mathcal{L} = \mathcal{L}_{CE} + \mathcal{L}_{XRCL}$. Similarly, for the class contrastive objective variant $XCCL$, the total loss is $\mathcal{L} = \mathcal{L}_{CE} + \mathcal{L}_{XCCL}$.

Unlike some related studies, we do not use a dedicated validation set during few-shot fine-tuning. This approach aligns with the few-shot premise, avoiding the need for additional labeled data that a validation set would require (Schmidt et al., 2023). Consequently, all models are trained for a fixed number of epochs on the few-shot data. We argue that this reflects real-world conditions more accurately, as performance is not skewed by the size or quality of a validation dataset. For our experiments, we use the XLMR-Base model (Conneau et al., 2020), comprising 270 million parameters. Unless specified otherwise, we adhere to training parameters that include a batch size of 64, a learning rate of 2e-5, and the AdamW optimizer (Loshchilov and Hutter, 2017). In the `Source`→`Target` setting, models are trained for 5 epochs on task data in the source language. During few-shot adaptation in both `Source`→`Target` and `Target` settings, models are trained for 10 epochs using data from the source and target language. We include relevant benchmarks proposed in related studies. To ensure fair model comparisons, we integrate relevant benchmarks in our XLT settings, with all models being reimplemented for consistency.

In the regular fine-tuning (**FT**) approach the beginning-of-sequence token `[s]` is used as the sentence representation (Lauscher et al., 2020; Schmidt et al., 2022). The model is fine-tuned using the $\mathcal{L}_{CE}$ objective.

In the prompt-learning from cross-lingual templates (**PCT**) framework Qi et al. (2022) augment model inputs with multilingual prompt templates, optimizing the model using both the $\mathcal{L}_{CE}$ objective

for non-augmented inputs and the $\mathcal{L}_{CE'}$ objective for augmented inputs. Additionally, they introduce a consistency loss term based on the Kullback-Leibler divergence between predicted token probabilities. In contrast to CoLAP, PCT is applied during both supervised fine-tuning and few-shot adaptation phases.

Checkpoint averaging (**CA**), proposed by Schmidt et al. (2023), represents a robust baseline for FS-XLT. Their implementation is based on regular fine-tuning and averages model weights during few-shot adaptation.

4.2 Episode Sampling

In the few-shot task adaptation, episodes are randomly sampled from the training dataset of the target language. Each episode consists of K input sequences along with corresponding labels. The value of K is varied among the set $\{5, 10, 50, 100, 250\}$. Episodes comprise K uniformly sampled instances from the available class labels N_Y. In scenarios where $K \mod N_Y \neq 0$ the remaining instances are randomly selected.

For the computation of our cross-lingual contrastive loss terms, an episode is constructed to include K instances from the training dataset in the target language D_T and an additional K instances from the training dataset in the source language D_S. It is important to note that we assume D_T and D_S to be parallel corpora, meaning that all instances present in the training dataset of the target language D_T are also available in the training dataset of the source language D_S.

4.3 Datasets and Languages

The **XNLI** dataset contains 15 languages, each containing 7,500 examples. In natural language inference, the goal is to determine if a given hypothesis entails, contradicts, or is neutral relative to a premise. Initially annotated in English, the dataset was expanded to other languages using machine translation (Conneau et al., 2018). We assess model performance using the accuracy metric.

AmericasNLI extends the XNLI dataset by adding 10 indigenous languages from the Americas, categorized as very low-resource languages due to their minimal presence in large-scale text corpora. The dataset is human-translated based on the Spanish XNLI. It includes 750 development and 750 test instances per language, which we utilize as our training and testing sets, respectively (Ebrahimi et al., 2022). Performance evaluation on AmericasNLI is based on accuracy[3].

MultiTACRED is a multilingual relation extraction dataset covering 12 languages. The primary task in sentence-level relation extraction involves categorizing a given set of entities into one of 41 distinct relation types. This dataset has been created through machine translation from English annotations (Hennig et al., 2023). We utilize instances from the development set to construct few-shot episodes and assess model performance based on accuracy.

5 Results and Discussion

The results, summarized in Table 1 are obtained by averaging the accuracy metrics across 5 random seeds for each language in the datasets, with each experiment repeated three times to yield 15 evaluation outcomes per target language.

The results in Table 1 validate our initial hypothesis that discriminative task-specific information can be transferred from English to lower-resource languages using only few labeled examples. CoLAP consistently improves downstream performance for natural language inference and relation extraction tasks across all evaluated settings. This includes target languages included during pretraining of XLM-R (XNLI and MultiTACRED) and unseen languages, such as the ones included in AmNLI. It demonstrates superior performance in low-resource settings, such as the `Target` setting when labeled task data in English is not available. Moreover, CoLAP outperforms other approaches across all datasets and settings when source and target data is limited to $K = 5$ exemplars. Even in the most resource-abundant scenario of $K = 250$ exemplars, CoLAP either matches or exceeds performance of its benchmarks. Furthermore, the results of CoLAP employing the $XCCL$ objective prove that class-specific features can effectively be transferred from English to lower-resource languages. In comparison to the $XRCL$ variant, $XCCL$ only sees a 1.31% reduction in average few-shot performance ($K > 0$) across all datasets and settings. While CoLAP is the most data-efficient approach, all evaluated few-shot models provide substantial performance improvements over zero-shot settings with $K = 50$ exemplars. While regular fine-tuning demonstrates strong performance on NLI tasks,

[3]Otomí "oto" has less than 250 samples and therefore is not included in the $K = 250$ setting.

Model	K	XNLI		AmNLI		MultiTACRED	
		Target	Source→Target	Target	Source→Target	Target	Source→Target
FT	0	33.34 ± 0.01	**73.21** ± 0.01	33.37 ± 0.01	36.91 ± 0.01	2.40 ± 0.15	24.57 ± 0.01
CA		33.34 ± 0.01	**73.21** ± 0.01	33.37 ± 0.01	36.91 ± 0.01	2.40 ± 0.01	24.57 ± 0.01
PCT		30.76 ± 0.01	72.97 ± 0.01	33.17 ± 0.01	**37.98** ± 0.01	12.92 ± 0.01	45.87 ± 0.01
CoLAP w/ XRCL		**33.45** ± 0.01	73.15 ± 0.01	**34.63** ± 0.01	35.79 ± 0.01	**48.00** ± 0.01	**51.58** ± 0.01
CoLAP w/ XCCL		**33.45** ± 0.01	73.15 ± 0.01	**34.63** ± 0.01	35.79 ± 0.01	**48.00** ± 0.01	**51.58** ± 0.01
FT	5	33.43 ± 0.15	71.74 ± 1.79	33.37 ± 0.36	39.57 ± 1.36	2.91 ± 0.24	42.65 ± 0.95
CA		33.38 ± 0.09	73.21 ± 1.25	33.25 ± 0.17	37.94 ± 0.93	2.71 ± 0.17	38.94 ± 1.53
PCT		33.55 ± 0.71	72.52 ± 1.28	33.41 ± 0.23	39.11 ± 1.68	36.10 ± 16.56	69.01 ± 1.24
CoLAP w/ XRCL		**33.84** ± 0.68	**73.56** ± 1.22	**34.60** ± 0.55	**40.01** ± 1.43	**43.86** ± 8.93	**69.26** ± 2.62
CoLAP w/ XCCL		33.82 ± 0.65	73.04 ± 1.20	34.43 ± 0.63	39.71 ± 1.01	43.10 ± 8.97	69.18 ± 2.63
FT	10	33.18 ± 0.19	72.47 ± 1.39	33.60 ± 0.35	**42.17** ± 1.43	2.74 ± 0.24	43.91 ± 1.14
CA		33.24 ± 0.22	73.22 ± 0.95	33.49 ± 0.24	40.28 ± 0.48	2.53 ± 0.13	40.82 ± 0.84
PCT		32.99 ± 0.95	72.92 ± 0.65	33.29 ± 0.58	41.26 ± 0.86	38.52 ± 15.77	70.19 ± 3.24
CoLAP w/ XRCL		**34.34** ± 0.47	**73.69** ± 0.88	34.98 ± 0.34	40.69 ± 1.00	**50.48** ± 9.53	70.16 ± 3.80
CoLAP w/ XCCL		34.31 ± 0.53	73.15 ± 0.95	**35.08** ± 0.29	41.09 ± 1.21	50.01 ± 9.63	**70.32** ± 3.51
FT	50	33.49 ± 0.22	73.54 ± 0.78	33.95 ± 0.27	**44.76** ± 0.54	2.79 ± 0.23	46.88 ± 1.52
CA		33.46 ± 0.26	73.55 ± 0.83	33.64 ± 0.12	42.31 ± 0.55	2.67 ± 0.17	46.22 ± 0.54
PCT		34.10 ± 0.49	73.72 ± 0.47	34.07 ± 0.74	43.38 ± 0.63	56.91 ± 2.01	**75.36** ± 1.34
CoLAP w/ XRCL		**34.48** ± 0.29	**73.88** ± 0.67	**34.75** ± 0.29	41.65 ± 0.51	**59.08** ± 2.56	74.02 ± 2.15
CoLAP w/ XCCL		34.47 ± 0.24	73.81 ± 0.54	34.71 ± 0.30	41.98 ± 0.40	58.94 ± 2.50	73.98 ± 2.13
FT	100	33.61 ± 0.24	73.87 ± 0.58	33.95 ± 0.38	**45.94** ± 0.43	3.06 ± 0.36	48.87 ± 1.74
CA		33.57 ± 0.17	73.95 ± 0.49	33.88 ± 0.37	45.39 ± 0.37	2.88 ± 0.32	48.41 ± 1.58
PCT		**35.37** ± 0.54	73.80 ± 0.49	35.55 ± 0.53	45.23 ± 0.37	59.92 ± 3.19	**78.26** ± 8.74
CoLAP w/ XRCL		35.09 ± 0.20	**74.09** ± 0.31	35.63 ± 0.57	44.79 ± 0.27	**61.62** ± 3.94	77.29 ± 3.46
CoLAP w/ XCCL		35.03 ± 0.15	74.03 ± 0.44	**35.90** ± 0.40	45.18 ± 0.24	61.50 ± 3.74	77.06 ± 3.58
FT	250	34.22 ± 0.38	73.35 ± 0.68	34.87 ± 0.47	48.53 ± 1.18	5.52 ± 1.79	55.12 ± 5.08
CA		33.93 ± 0.28	73.63 ± 0.59	34.44 ± 0.19	48.20 ± 1.12	4.39 ± 1.61	53.69 ± 4.31
PCT		38.22 ± 0.71	73.03 ± 0.59	38.15 ± 0.81	47.73 ± 1.08	68.34 ± 1.02	**82.19** ± 4.52
CoLAP w/ XRCL		**38.25** ± 0.76	**74.00** ± 0.29	**39.28** ± 1.54	**48.64** ± 1.24	69.86 ± 1.26	82.16 ± 7.74
CoLAP w/ XCCL		38.08 ± 0.73	73.60 ± 0.64	39.19 ± 1.79	48.44 ± 0.74	**69.95** ± 1.45	81.99 ± 7.49

Table 1: Average accuracy across languages, excluding English, in XNLI, AmNLI, and MultiTACRED datasets. Models are evaluated in two settings: Target, adapting models to the target language without prior fine-tuning, and Source→Target, indicating few-shot adaptation to the target language following fine-tuning on English task data.

the MultiTACRED dataset, with its large number of relation types, proves to be more suitable to prompting-based approaches like CoLAP or PCT.

5.1 Ablation Studies

In Table 3, we investigate the integration of various strategies from related research with our CoLAP method. We assess these combined model variants in the Source→Target setting, examining both small ($K = 5$) and large ($K = 250$) few-shot episodes. Additionally, we consider a variant of CoLAP that utilizes InfoXLM (Chi et al., 2021), which already includes contrastive representation learning in its pretraining phase. The analysis shows that the default implementation of CoLAP performs well for the NLI tasks, while integrating CoLAP with PCT exceeds the best performing models on MultiTACRED. Table 4 reveals that the CE_{Source} loss term for prompts in the source language, also included in PCT, is a significant factor improving performance on MultiTACRED. CoLAP can effectively be applied to InfoXLM, providing the best performance on the AmNLI and MultiTACRED datasets for small few-shot episodes with $K = 5$. Further, we evaluate the impact of CoLAP's individual components on overall model performance in Table 4 included in the Appendix. This reveals that models integrating both the $XRCL$ and $XCCL$ objectives match or improve results relative to those utilizing a single contrastive loss term.

5.2 Layer Selection for Contrastive Learning

To determine the most effective layer for contrastive representation alignment, we conduct experiments with the CoLAP XRCL model in the Source→Target configuration, selecting $K = 50$ exemplars. The findings, illustrated in Figure 5, reveal that applying contrastive representation learning at the 10th layer of XLM-R yields the best performance for NLI tasks. This observation suggests that contrastive learning objectives benefit from the general-domain information captured in the middle layers of the model. In contrast, initial layers process low-level syntactic features, whereas the final layers are more focused on extracting discriminative features relevant for the classification task at hand. This finding diverges from that of Chi et al. (2021), who identified the 8th layer as

PLM	Model	K	XNLI			AmNLI			MultiTACRED		
			Random	High	Low	Random	High	Low	Random	High	Low
XLM-R	CoLAP w/ XRCL	5	73.56	**74.41**	74.22	**40.01**	38.86	38.81	69.26	**72.22**	69.62
	CoLAP w/ XCCL		73.04	74.38	74.12	39.71	39.32	39.33	69.18	71.72	69.36
InfoXLM	CoLAP w/ XRCL	5	72.88	**75.48**	74.91	**41.76**	41.06	41.17	70.94	74.91	72.81
	CoLAP w/ XCCL		72.60	75.32	74.86	41.59	40.61	41.16	71.03	**75.05**	72.54
XLM-R	CoLAP w/ XRCL	10	73.69	73.87	74.08	40.69	39.52	40.01	70.16	74.73	70.87
	CoLAP w/ XCCL		73.15	**74.10**	73.85	**41.09**	39.40	40.33	70.32	**74.74**	69.82
InfoXLM	CoLAP w/ XRCL	10	73.61	**75.18**	**75.18**	43.02	43.25	42.12	72.65	76.03	73.67
	CoLAP w/ XCCL		73.47	75.17	74.95	43.33	**43.60**	42.22	72.71	**76.30**	73.66

Table 2: Evaluation of exemplar selection using representation similarity scores, comparing "High" and "Low" similarity exemplars against random selection, with accuracy as the evaluation metric.

PLM	Model	K	XNLI	AmNLI	MultiTACRED
XLM-R	CoLAP w/ XRCL	5	73.56	**40.01**	69.26
	w/ FT		72.60	39.78	42.72
	w/ CA		**73.75**	39.03	64.79
	w/ PCT		72.95	38.16	**69.87**
InfoXLM	CoLAP w/ XRCL	5	**72.88**	**41.76**	70.94
	CoLAP w/ XCCL		72.60	41.59	**71.03**
XLM-R	CoLAP w/ XRCL	250	**74.00**	**48.64**	82.16
	w/ FT		73.78	48.28	55.23
	w/ CA		73.89	48.11	81.62
	w/ PCT		73.45	47.68	**84.97**
InfoXLM	CoLAP w/ XRCL	250	74.13	47.29	**83.38**
	CoLAP w/ XCCL		**74.23**	**48.50**	83.00

Table 3: Model variants with (w/) or without (w/o) indicated architectural changes. Accuracy (%) is evaluated in the `Source`→`Target` setting.

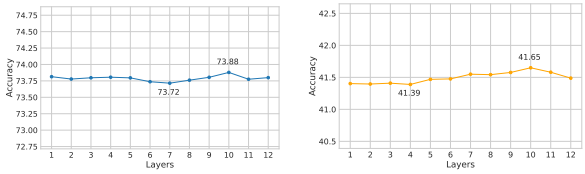

Figure 5: Comparative performance of utilizing different layer representations for contrastive learning in the CoLAP XRCL model with K=50 exemplars. Results are shown for XNLI (left) and AmNLI (right) datasets.

optimal for the contrastive learning objectives used during the pretraining of InfoXLM.

5.3 Few-Shot Exemplar Selection

Building upon the performance of our class contrastive learning objective, we investigate the potential of utilizing class representation similarity for selecting few-shot exemplars. We hypothesize that transferring information from English to other languages via contrastive learning is more sample efficient when selecting exemplars with high intra-class similarity and low inter-class similarity. Employing the XLM-R model, fine-tuned on English task data through prompt-based training, we extracted `[MASK]` representations from English training instances. We then created class prototypes by averaging these embeddings for each class. For each instance x_i, we compute the cosine similarity to its own class prototype P^+ and dissimilarity to other class prototypes P^-, using the formula $s_i = \phi(x_i, P^+) + N_Y - \phi(x_i, P^-)$, where N_Y is the total number of classes. The computational complexity of this procedure is primarily determined by the pairwise similarity calculations, which depend on the number of instances and classes. We selected exemplars based on the highest or lowest similarity scores, focusing on sets of $K = 5$ and $K = 10$. Few-shot episodes consist of these selected exemplars in the source language and their corresponding translations in the target language. Table 2 shows that selecting exemplars based on representation similarity enhances data efficiency for languages in XNLI and MultiTACRED, improving sample efficiency by at least 50%. Interestingly, even exemplars with low similarity scores outperform randomly selected few-shot episodes. Notably, with just $K = 5$ exemplars selected based on our similarity criteria, CoLAP exceeds the performance of randomly selected sets of $K = 250$ exemplars for XNLI. However, for languages not included in the pretraining data, selecting exemplars based on similarity in English yields similar or worse results compared to random selection.

6 Conclusion

Our study showcases the effectiveness of contrastive learning in enhancing few-shot adaptation for mPLMs. We find that aligning cross-lingual representations of translated instances substantially improves few-shot cross-lingual transfer performance. In contrast to existing contrastive learning approaches, our CoLAP approach does not require extensive pretraining on parallel multilingual corpora, yet it still manages to enhance performance, even with the smallest set of few-shot exemplars. Additionally, we demonstrate that selecting few-shot exemplars based on representation similarity enhances data efficiency for languages included in the mPLM pretraining copora.

7 Limitations

In CoLAP, we introduce an effective strategy to narrow the cross-lingual transfer performance gap in few-shot cross-lingual transfer (FS-XLT), tackling the challenge of imbalanced language representation in the pretraining data of current multilingual pretrained language models (mPLMs). Although our approach enhances data efficiency in improving downstream tasks, it does not directly resolve the underlying disparities that contribute to the cross-lingual transfer gap in mPLMs.

A key consideration in CoLAP's implementation is the need for translations of the few-shot examples, predicated on the assumption that acquiring (machine) translations, especially in high-resource languages like English, is cost-effective and straightforward. Nonetheless, this requirement introduces an additional step in the process. While the class contrastive learning objective ($XCCL$) in CoLAP diminishes the dependency on parallel corpora, its counterpart, the cross-lingual representation contrastive learning objective ($XRCL$), necessitates translated instance pairs. Furthermore, the advantage of employing CoLAP with the $XCCL$ objective for not necessitating parallel corpora during few-shot training, is confined to classification tasks. This specificity limits the broader applicability of CoLAP's class contrastive objective across different types of downstream tasks.

References

Alan Ansell, Edoardo Maria Ponti, Jonas Pfeiffer, Sebastian Ruder, Goran Glavaš, Ivan Vulić, and Anna Korhonen. 2021. MAD-G: Multilingual adapter generation for efficient cross-lingual transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4762–4781, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

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.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir Meza Ruiz, Gustavo Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando Coto-Solano, Thang Vu, and Katharina Kann. 2022. AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299, Dublin, Ireland. Association for Computational Linguistics.

Yiduo Guo, Yaobo Liang, Dongyan Zhao, Bing Liu, and Nan Duan. 2023. Analyzing and reducing the performance gap in cross-lingual transfer with fine-tuning slow and fast. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4002–4017, Toronto, Canada. Association for Computational Linguistics.

Leonhard Hennig, Philippe Thomas, and Sebastian Möller. 2023. MultiTACRED: A multilingual version of the TAC relation extraction dataset. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3785–3801, Toronto, Canada. Association for Computational Linguistics.

Lianzhe Huang, Shuming Ma, Dongdong Zhang, Furu Wei, and Houfeng Wang. 2022. Zero-shot cross-lingual transfer of prompt-based tuning with a unified multilingual prompt. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language*

Processing, pages 11488–11497, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *The Fifth International Conference on Learning Representations, ICLR 2017*.

Ercong Nie, Sheng Liang, Helmut Schmid, and Hinrich Schütze. 2023. Cross-lingual retrieval augmented prompt for low-resource languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8320–8340, Toronto, Canada. Association for Computational Linguistics.

Kunxun Qi, Hai Wan, Jianfeng Du, and Haolan Chen. 2022. Enhancing cross-lingual natural language inference by prompt-learning from cross-lingual templates. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1910–1923, Dublin, Ireland. Association for Computational Linguistics.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *J. Artif. Int. Res.*, 65(1):569–630.

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

Fabian David Schmidt, Ivan Vulić, and Goran Glavaš. 2022. Don't stop fine-tuning: On training regimes for few-shot cross-lingual transfer with multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10725–10742, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Fabian David Schmidt, Ivan Vulić, and Goran Glavaš. 2023. Free lunch: Robust cross-lingual transfer via model checkpoint averaging. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5712–5730, Toronto, Canada. Association for Computational Linguistics.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation learning with contrastive predictive coding.

Genta Winata, Shijie Wu, Mayank Kulkarni, Thamar Solorio, and Daniel Preotiuc-Pietro. 2022. Cross-lingual few-shot learning on unseen languages. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 777–791, Online only. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Shaoyang Xu, Junzhuo Li, and Deyi Xiong. 2023. Language representation projection: Can we transfer factual knowledge across languages in multilingual language models? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3692–3702, Singapore. Association for Computational Linguistics.

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

Huiyun Yang, Huadong Chen, Hao Zhou, and Lei Li. 2022. Enhancing cross-lingual transfer by manifold mixup. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*.

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.

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

Bo Zheng, Li Dong, Shaohan Huang, Wenhui Wang, Zewen Chi, Saksham Singhal, Wanxiang Che, Ting Liu, Xia Song, and Furu Wei. 2021. Consistency regularization for cross-lingual fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3403–3417, Online. Association for Computational Linguistics.

Meng Zhou, Xin Li, Yue Jiang, and Lidong Bing. 2023. Enhancing cross-lingual prompting with dual prompt augmentation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11008–11020, Toronto, Canada. Association for Computational Linguistics.

A Training Details

Training is conducted on a single NVIDIA A6000 GPU. The process of fine-tuning XLM-R with prompt-based techniques on the English XNLI dataset takes approximately 10 hours. For few-shot adaptation employing CoLAP across 14 languages from the XNLI dataset (English excluded) in the $K = 250$ setting, the duration is about 30 minutes.

We employ language agnostic prompting templates for CoLAP. For NLI tasks, utilize the template "`[Premise]` `[MASK]`, `[Hypothesis]`", where `[Premise]` and `[Hypothesis]` are placeholders replaced with the corresponding instance x inputs. The label words "Yes", "Maybe", and "No" are used, in line with Schick and Schütze (2021). Additionally, we utilize the multilingual prompt templates introduced by Qi et al. (2022), following their PCT approach "`[Premise]` Question:`[Hypothesis]`? Answer:`[MASK]`".

For relation extraction tasks, we adopt the prompt template "`[Sentence]` `[E1]` `[MASK]` `[E2]`", where `[E1]` and `[E2]` are replaced with the respective entities mentioned in the context of the sentence. We adapt the multilingual PCT template to the relation extraction task "`[Sentence]` Relation:`[E1]`, `[E2]`? Answer:`[MASK]`" and machine translate the English template using Google translate (Qi et al., 2022). For MultiTACRED, we augment the input sentences for benchmarked models by including entity marker tokens `[E1]`, `[/E1]`, `[E2]`, and `[/E2]` indicating the position of the entities in the context. We reduce the number relation types from 41 to 31 by merging overlapping class labels for all models evaluated on MultiTACRED. The corresponding label mapping and template files are available at https://anonymous.4open.science/r/CoLAP-B2BE.

To accurately reflect the real-world few-shot performance, we consistently train all models for 10 epochs on the few-shot episodes. We observe that CoLAP models trained on episodes with K greater than 100 instances gain performance when trained up to 50 epochs.

Model	K	XNLI	AmNLI	MultiTACRED
CoLAP w/ XRCL	5	**73.56**	40.01	69.26
w/o CE		73.46	36.24	65.89
w/o XRCL		72.89	39.87	68.90
w/ CE_{Source}		73.34	39.81	**69.85**
w/ XCCL		73.53	**40.14**	69.28
CoLAP w/ XCCL	5	73.52	40.11	69.18
w/o CE		73.48	37.16	62.10
w/o XCCL		72.89	39.87	68.90
w/ CE_{Source}		73.24	39.83	**69.86**
w/ XRCL		**73.53**	**40.14**	69.28
CoLAP w/ XRCL	250	**74.00**	48.49	82.16
w/o CE		73.12	41.74	66.32
w/o XRCL		73.72	47.98	81.95
w/ CE_{Source}		73.61	**48.57**	**83.18**
w/ XCCL		**74.00**	48.26	82.02
CoLAP w/ XCCL	250	73.84	48.22	81.99
w/o CE		72.77	45.58	70.46
w/o XCCL		73.72	47.98	81.95
w/ CE_{Source}		73.61	**48.64**	**82.97**
w/ XRCL		**74.00**	48.26	82.02

Table 4: Model variants with (w/) or without (w/o) indicated architectural changes. Accuracy (%) is evaluated in the `Source`→`Target` setting.

B Ablation Studies

To validate the individual contribution of CoLAP's components, we conduct ablation studies, comparing model performance with indicated architectural modifications in the `Source`→`Target` setting for small $K = 5$ and large few-shot episodes $K = 250$.

The findings, as detailed in Table 4, reveal that incorporating both the $XRCL$ and $XCCL$ objectives within CoLAP, denoted as w/ XCCL and w/ XRCL, achieve on par or superior performance compared to variants employing a singular contrastive loss term. The CE loss term, proves essential for effective few-shot adaptation. Models excluding the CE loss term (w/o CE) substantially reduce few-shot performance. Interestingly, the addition of a CE_{Source} loss term, aimed at retaining the prediction accuracy in the source language as suggested by Schmidt et al. (2022) and Qi et al. (2022), does not yield performance improvements for NLI tasks. However, it improves model performance on MultiTACRED.

C Detailed Results

The aggregated results presented in Table 1 are broken down by individual languages in Tables 5 and 6 for detailed analysis. Given that English serves as the source language in our study, its accuracy scores are excluded from the average performance calculations.

Model	K	en	ar		bg		de		el		es		fr		hi		ru		sw		th		tr		ur		vi		zh		Avg.	
			T	S→T	T	S→T	T	S→T	T	S→T	T	S→T	T	S→T	T	S→T	T	S→T	T	S→T	T	S→T	T	S→T	T	S→T	T	S→T	T	S→T	T	S→T
FT	0	84.07	33.32	71.38	33.45	77.25	33.37	76.05	33.22	75.42	33.36	78.66	33.34	77.23	33.42	69.22	33.35	75.69	33.16	65.1	33.25	72.33	33.40	72.62	33.37	65.66	33.37	74.75	33.36	73.59	33.34	73.21
CA		84.07	33.32	71.38	33.45	77.25	33.37	76.05	33.22	75.42	33.36	78.66	33.34	77.23	33.42	69.22	33.35	75.69	33.16	65.1	33.25	72.33	33.40	72.62	33.37	65.66	33.37	74.75	33.36	73.59	33.34	73.21
PCT		85.31	27.96	72.96	33.19	73.01	32.59	73.00	28.28	73.00	32.97	73.02	32.20	73.02	30.88	72.93	33.13	73.00	32.71	72.86	27.11	72.96	28.60	72.96	33.33	72.89	28.28	72.98	29.32	72.99	30.75	72.97
CoLAP w/ XRCL		84.37	34.00	72.00	33.80	77.23	32.80	75.77	34.21	75.93	33.20	78.74	33.58	77.23	33.27	69.18	33.95	75.67	30.90	65.17	33.21	71.62	35.62	72.46	33.96	65.41	33.49	73.81	32.14	73.95	33.44	73.16
CoLAP w/ XCCL		84.37	34.00	72.00	33.80	77.23	32.80	75.77	34.21	75.93	33.20	78.74	33.58	77.23	33.27	69.18	33.95	75.67	30.90	65.17	33.21	71.62	35.62	72.46	33.96	65.41	33.49	73.81	32.14	73.95	33.44	73.16
FT	5	84.07	33.49	72.04	33.37	76.53	33.31	74.67	33.44	74.50	33.33	76.81	33.51	76.07	34.25	66.73	33.15	74.64	33.45	61.42	33.23	70.68	33.33	71.15	33.78	63.37	33.18	72.81	33.26	72.90	33.43	71.74
CA		84.07	33.28	72.33	33.35	77.57	33.31	75.66	33.37	75.69	33.08	78.2	33.42	77.16	33.49	69.57	33.22	75.83	34.19	63.85	33.17	71.66	33.33	72.35	33.61	66.81	33.45	74.06	33.02	74.26	33.38	73.21
PCT		85.31	33.80	72.38	33.51	77.03	33.41	74.61	33.66	75.21	33.41	78.36	33.9	77.14	33.27	67.88	33.10	75.76	34.16	59.31	33.07	72.26	33.60	71.38	33.56	65.03	34.24	74.49	33.00	74.46	33.55	72.52
CoLAP w/ XRCL		84.37	33.66	71.71	33.56	78.54	33.77	76.47	33.82	76.44	33.32	79.28	35.18	77.56	33.07	69.83	34.16	75.82	34.11	64.25	33.01	73.19	34.49	72.17	33.66	65.63	34.01	74.84	33.96	74.05	33.84	73.56
CoLAP w/ XCCL		84.37	33.71	71.84	33.50	77.74	33.69	76.17	33.77	76.00	33.34	78.33	34.92	76.87	33.11	68.94	34.10	75.23	34.24	63.99	33.03	72.01	34.45	72.10	33.65	65.85	34.10	73.30	33.83	74.23	33.82	73.04
FT	10	84.07	33.05	72.13	33.70	76.72	33.37	74.42	32.91	74.91	33.01	78.15	33.29	75.82	33.17	68.96	33.02	74.28	33.17	62.53	33.27	70.00	33.12	72.06	33.26	66.5	33.09	74.02	33.10	74.14	33.18	72.47
CA		84.07	33.07	72.27	33.64	77.19	33.21	75.20	33.21	75.45	33.33	78.75	33.32	76.68	33.18	70.04	33.03	75.27	32.86	64.0	33.33	70.97	33.42	72.71	33.36	67.13	33.29	74.75	33.16	74.70	33.24	73.22
PCT		85.31	32.47	72.58	32.98	77.08	32.91	75.78	32.94	75.60	32.74	78.55	32.71	77.09	32.63	68.79	33.07	75.41	33.58	61.57	32.13	71.05	33.54	70.83	33.27	66.57	33.59	74.86	33.31	75.05	32.99	72.91
CoLAP w/ XRCL		84.37	33.96	73.35	34.54	78.41	33.76	75.89	35.31	76.19	33.81	78.97	34.66	77.10	33.3	70.59	34.95	76.27	34.08	64.60	33.79	71.12	35.52	72.38	34.08	66.66	34.78	74.74	34.21	75.34	34.34	73.69
CoLAP w/ XCCL		84.37	34.07	72.82	34.51	77.40	33.44	75.90	35.23	75.79	33.88	78.51	34.59	76.66	33.23	69.80	35.00	74.98	34.13	64.37	33.87	70.64	35.36	71.58	34.02	66.71	34.68	74.32	34.37	74.63	34.31	73.15
FT	50	84.07	33.82	72.11	33.27	77.37	33.71	75.89	33.32	75.37	33.16	78.83	34.06	77.10	33.29	69.28	33.90	75.70	33.1	65.55	33.51	72.31	33.92	72.59	33.66	67.43	33.39	75.31	32.69	74.68	33.49	73.54
CA		84.07	33.82	71.86	33.07	77.36	33.50	75.77	33.33	75.54	33.15	78.89	33.73	76.96	33.32	69.28	34.00	75.49	33.32	65.70	33.27	72.27	33.77	72.92	33.52	67.83	33.45	75.14	33.17	74.65	33.46	73.55
PCT		85.31	33.18	72.62	34.39	78.00	33.76	76.20	34.10	75.73	34.02	78.79	34.73	77.95	34.24	69.68	33.68	76.10	34.82	63.71	33.40	73.11	34.55	72.22	34.30	66.66	34.38	75.56	33.86	75.82	34.10	73.72
CoLAP w/ XRCL		84.37	34.32	72.00	34.57	78.23	35.03	75.91	34.10	76.41	34.60	79.21	35.17	77.86	33.99	70.22	34.49	76.29	34.61	65.92	33.35	72.17	35.58	72.26	34.16	67.28	34.86	74.56	33.86	76.06	34.48	73.88
CoLAP w/ XCCL		84.37	34.37	72.54	34.55	77.88	34.91	76.01	33.98	75.95	34.53	79.31	35.21	77.38	34.10	70.12	34.58	76.09	34.62	65.54	33.39	72.44	35.63	72.77	34.10	67.11	34.88	74.69	33.78	75.46	34.47	73.81
FT	100	84.07	33.46	72.71	33.50	77.29	33.81	76.31	33.88	75.64	33.79	79.02	34.66	77.27	33.57	70.07	33.22	75.58	33.45	66.01	33.32	73.07	32.62	72.99	33.96	67.54	33.36	75.82	33.92	74.85	33.61	73.87
CA		84.07	33.31	72.50	33.35	77.37	33.97	76.40	33.90	75.74	33.72	79.15	34.06	77.32	33.72	69.97	33.25	75.80	33.36	65.98	33.38	73.20	32.69	73.01	33.84	67.88	33.21	76.11	34.17	74.86	33.57	73.95
PCT		85.31	35.56	72.46	36.57	77.60	35.10	75.92	35.84	76.04	35.35	78.66	35.50	78.14	35.44	69.61	34.79	76.13	35.26	64.31	35.78	73.48	34.78	72.16	35.15	66.72	35.27	76.29	34.72	75.71	35.36	73.80
CoLAP w/ XRCL		84.37	34.08	72.52	35.42	78.29	35.52	76.49	35.81	76.78	34.32	79.08	36.40	78.18	34.46	70.26	35.64	76.29	34.82	65.71	32.99	72.81	35.43	72.83	35.04	67.54	36.76	75.34	34.57	75.2	35.09	74.09
CoLAP w/ XCCL		84.37	34.11	72.56	35.35	78.09	35.55	76.81	35.70	76.35	34.37	79.11	36.38	77.73	34.54	70.23	35.47	76.08	34.79	65.76	32.98	72.81	35.37	73.04	34.95	67.19	36.69	75.69	34.21	74.93	35.03	74.03
FT	250	84.07	33.80	71.87	34.61	76.89	35.19	75.69	35.88	75.00	33.78	78.26	35.76	76.96	33.40	69.64	34.51	75.41	33.54	65.28	33.88	72.69	33.79	72.93	33.67	66.70	33.57	75.27	33.76	74.29	34.22	73.35
CA		84.07	33.39	72.16	34.34	76.99	34.56	76.00	35.01	75.19	33.96	78.46	35.11	77.07	33.34	70.08	34.17	75.76	33.40	65.58	33.61	72.94	33.75	73.29	33.59	67.15	33.33	75.56	33.51	74.59	33.93	73.63
PCT		85.31	37.71	72.00	39.29	76.61	37.74	75.16	39.68	74.94	38.36	77.97	38.51	76.83	37.52	69.50	39.21	75.11	37.94	63.87	37.77	73.33	37.50	71.45	37.12	65.35	39.75	75.18	36.98	75.12	38.22	73.03
CoLAP w/ XRCL		84.37	38.71	72.74	38.48	77.71	39.15	76.49	39.68	76.13	39.23	79.03	41.74	78.01	37.56	69.69	38.51	75.82	36.73	65.91	34.53	73.46	38.37	72.95	35.74	66.76	39.82	75.79	37.22	75.47	38.25	74.00
CoLAP w/ XCCL		84.37	38.57	72.16	38.29	77.38	38.78	75.93	39.58	75.50	39.05	78.55	41.52	77.19	37.41	69.57	38.74	75.14	36.44	65.94	34.42	73.02	38.33	72.99	35.72	66.44	39.69	75.37	36.63	75.15	38.08	73.60

Table 5: Average accuracy per language in XNLI.

Model	K	aym		bzd		cni		gn		hch		nah		oto		quy		shp		tar		Avg.	
		T	S→T	T	S→T	T	S→T	T	S→T	T	S→T	T	S→T	T	S→T	T	S→T	T	S→T	T	S→T	T	S→T
FT	0	33.33	36.80	33.33	37.60	33.33	36.27	33.33	38.13	33.33	37.20	33.71	37.94	33.32	35.96	33.33	38.27	33.33	35.73	33.33	35.20	33.37	36.91
CA		33.33	36.80	33.33	37.60	33.33	36.27	33.33	38.13	33.33	37.20	33.71	37.94	33.32	35.96	33.33	38.27	33.33	35.73	33.33	35.20	33.37	36.91
PCT		32.80	38.42	32.93	38.16	33.33	38.02	33.33	37.89	33.33	36.96	33.47	40.81	33.02	38.93	33.07	36.82	33.60	36.29	32.80	37.49	33.17	37.98
CoLAP w/ XRCL		34.27	35.47	33.33	35.47	34.80	34.93	34.80	36.53	34.13	34.80	35.09	37.67	33.69	35.83	35.87	36.53	35.20	35.07	35.07	35.60	34.62	35.79
CoLAP w/ XCCL		34.27	35.47	33.33	35.47	34.80	34.93	34.80	36.53	34.13	34.80	35.09	37.67	33.69	35.83	35.87	36.53	35.20	35.07	35.07	35.60	34.62	35.79
FT	5	32.53	39.17	34.37	40.53	33.40	40.37	33.3	38.93	33.27	38.07	33.03	40.07	33.36	40.17	33.57	40.30	33.27	40.63	33.60	37.47	33.37	39.57
CA		32.93	37.60	33.33	38.77	33.30	39.27	33.33	37.27	33.53	36.57	33.30	39.60	32.89	38.34	33.33	36.77	33.43	38.83	33.07	36.43	33.24	37.95
PCT		33.31	37.64	33.60	40.87	33.73	38.58	33.41	39.64	33.17	38.87	34.34	39.48	33.72	40.02	31.95	37.62	33.09	39.27	33.81	39.14	33.41	39.11
CoLAP w/ XRCL		34.00	38.77	33.63	40.69	35.28	39.28	34.88	41.57	34.61	38.88	34.77	40.11	34.47	40.24	35.17	40.29	35.33	39.68	33.89	40.59	34.60	40.01
CoLAP w/ XCCL		34.08	38.77	33.6	40.51	35.25	39.17	34.53	40.69	34.48	37.92	34.72	41.60	34.44	40.43	35.09	39.33	34.75	38.96	33.33	39.73	34.43	39.71
FT	10	33.23	41.43	34.17	40.30	33.83	42.63	33.33	43.60	33.43	39.77	33.27	43.36	34.59	42.91	32.87	41.63	33.77	43.80	33.47	42.30	33.60	42.17
CA		33.47	39.80	34.40	39.53	33.23	40.77	33.33	40.70	33.33	38.03	33.10	42.85	34.29	41.14	33.60	39.37	32.93	41.00	33.20	39.57	33.49	40.28
PCT		34.00	41.32	33.28	41.05	32.64	40.52	34.21	42.09	33.09	39.24	33.58	42.70	32.43	41.69	32.75	41.16	33.79	41.32	33.09	41.50	33.29	41.26
CoLAP w/ XRCL		34.03	38.93	33.89	39.39	35.89	40.03	36.77	42.21	33.81	39.95	34.74	42.20	34.81	41.26	35.2	40.32	36.03	40.96	34.61	41.65	34.98	40.69
CoLAP w/ XCCL		34.40	39.81	34.05	40.43	35.60	40.53	37.07	41.55	34.13	40.19	34.96	42.14	35.13	41.79	35.04	41.04	35.95	41.15	34.45	42.29	35.08	41.09
FT	50	33.90	45.37	34.07	44.43	34.77	44.93	34.07	47.53	32.60	41.67	33.60	45.46	33.32	43.72	34.67	45.50	34.37	45.57	34.10	43.43	33.95	44.76
CA		33.73	41.63	33.53	42.60	34.90	44.13	33.43	44.37	33.43	39.07	33.37	43.63	33.52	41.71	33.93	40.83	33.33	43.8	33.23	41.30	33.64	42.31
PCT		34.96	43.04	34.69	42.86	34.83	45.39	34.75	45.58	33.31	40.19	33.88	45.67	33.88	43.75	33.15	41.68	33.89	44.08	33.41	41.52	34.08	43.38
CoLAP w/ XRCL		34.19	39.17	33.57	41.81	36.75	41.31	34.64	43.65	33.25	40.69	35.47	42.49	36.02	42.38	34.93	39.57	35.07	42.75	33.63	42.64	34.75	41.65
CoLAP w/ XCCL		34.51	39.92	33.52	42.11	36.96	41.79	34.69	44.53	33.28	41.15	35.31	43.22	35.67	41.55	34.59	40.03	35.23	42.77	33.39	42.75	34.72	41.98
FT	100	34.27	47.37	33.50	47.10	34.03	48.47	35.37	48.57	33.13	40.73	33.57	48.27	35.23	44.39	33.10	43.77	33.13	47.4	34.20	43.30	33.95	45.94
CA		33.73	46.40	33.03	46.03	34.60	45.60	34.17	48.20	33.77	42.83	33.60	45.66	34.49	44.05	33.37	45.67	33.60	46.63	34.40	42.80	33.88	45.39
PCT		35.31	44.99	35.39	47.79	35.95	46.03	36.27	47.20	34.27	41.26	36.07	47.43	35.67	45.62	35.41	43.87	35.17	44.70	36.03	43.44	35.55	45.23
CoLAP w/ XRCL		34.93	44.37	33.71	44.53	36.56	47.55	36.51	48.16	35.44	42.88	35.72	45.07	36.36	43.69	36.08	43.97	35.39	45.49	35.65	42.13	35.63	44.78
CoLAP w/ XCCL		34.96	44.83	34.29	45.68	36.64	47.79	36.91	48.13	35.68	42.93	35.99	46.50	36.71	43.32	36.35	44.29	35.44	45.36	36.03	43.01	35.90	45.18
FT	250	34.90	48.70	34.10	52.50	35.27	50.37	37.13	50.47	34.10	43.47	35.20	50.51	-	-	33.40	48.17	35.70	47.73	34.07	44.87	34.87	48.53
CA		33.87	49.20	33.77	50.90	35.40	50.53	35.87	49.77	34.70	42.87	33.71	50.51	-	-	33.77	47.13	34.77	48.63	34.13	44.27	34.44	48.20
PCT		37.33	47.20	38.85	50.93	38.48	48.26	40.40	50.96	35.07	41.25	38.29	51.47	-	-	39.44	47.46	38.19	46.48	37.33	45.57	38.15	47.73
CoLAP w/ XRCL		39.01	48.69	38.53	51.23	43.20	51.69	40.53	51.83	37.49	43.03	38.11	48.95	-	-	38.91	49.46	39.01	46.73	39.41	46.13	39.36	48.64
CoLAP w/ XCCL		37.81	48.56	38.29	51.76	43.10	50.88	39.71	50.11	37.57	42.72	38.89	50.49	-	-	39.07	48.69	39.01	46.69	39.25	46.08	39.19	48.44

Table 6: Average accuracy per language in AmNLI.

Model	K	en	ar		de		es		fi		fr		hi		hu		ja		pl		ru		tr		zh		Avg.	
			T	S→T	T	S→T	T	S→T	T	S→T	T	S→T	T	S→T	T	S→T	T	S→T	T	S→T	T	S→T	T	S→T	T	S→T	T	S→T
FT	0	85.18	2.32	23.20	2.51	26.32	2.51	30.96	2.60	25.61	2.53	34.29	2.32	22.8	2.40	24.80	2.25	12.61	2.32	23.60	2.33	29.58	2.32	17.20	2.44	23.90	2.40	24.57
CA		85.18	2.32	23.20	2.51	26.32	2.51	30.96	2.60	25.61	2.53	34.29	2.32	22.80	2.40	24.80	2.25	12.61	2.32	23.60	2.33	29.58	2.32	17.20	2.44	23.90	2.40	24.57
PCT		87.38	4.40	40.40	6.07	49.39	7.11	54.39	3.25	50.81	20.82	53.88	2.80	38.00	3.60	46.00	36.49	41.44	4.00	50.00	31.25	52.50	4.00	33.60	31.22	40.00	12.92	45.87
CoLAP w/ XRCL		87.94	59.60	55.60	18.22	48.99	55.65	57.74	47.15	50.41	20.41	60.41	58.80	47.60	38.00	46.40	59.01	49.55	48.80	49.20	60.83	60.83	52.40	40.00	57.07	52.20	48.00	51.58
CoLAP w/ XCCL		87.94	59.60	55.60	18.22	48.99	55.65	57.74	47.15	50.41	20.41	60.41	58.80	47.60	38.00	46.40	59.01	49.55	48.80	49.20	60.83	60.83	52.40	40.00	57.07	52.20	48.00	51.58
FT	5	85.18	3.20	42.88	2.75	44.37	2.85	45.44	2.76	43.50	3.18	45.55	2.56	40.16	2.56	41.52	2.79	35.41	3.20	44.08	3.33	45.42	2.96	39.20	2.73	44.29	2.91	42.65
CA		85.18	2.24	37.76	3.08	41.13	2.76	45.10	2.68	40.41	2.53	45.06	2.32	36.32	2.48	34.88	2.43	31.44	2.88	38.24	3.67	43.08	2.64	32.72	2.83	41.07	2.71	38.93
PCT		87.38	31.80	69.80	31.68	71.56	46.86	73.01	36.59	69.61	32.96	73.27	32.80	67.10	31.60	71.60	42.79	64.64	32.20	66.80	41.67	71.56	33.10	65.50	39.15	63.66	36.10	69.01
CoLAP w/ XRCL		87.94	48.16	71.60	36.60	72.15	49.12	73.05	48.46	70.00	35.43	71.76	41.04	66.24	36.80	70.64	51.17	63.96	42.48	69.36	46.83	70.25	36.56	66.96	53.66	65.17	43.86	69.26
CoLAP w/ XCCL		87.94	47.28	72.08	36.36	72.23	48.03	72.72	47.48	70.41	33.63	71.59	40.32	66.00	35.68	70.40	50.36	64.14	42.16	69.04	46.75	69.42	35.68	66.88	53.46	65.27	43.10	69.18
FT	10	85.18	2.32	44.80	2.59	45.43	3.35	46.03	2.52	45.28	2.94	47.18	3.28	43.04	2.72	42.16	2.52	36.49	2.48	44.8	2.50	45.92	2.56	39.60	3.12	46.15	2.74	43.91
CA		85.18	2.16	39.76	2.67	43.08	2.34	44.69	2.44	41.30	3.27	45.71	2.56	39.68	2.32	36.96	2.61	33.87	2.56	40.80	2.25	44.50	2.32	35.84	2.83	43.61	2.53	40.82
PCT		87.38	38.10	68.20	30.06	73.28	44.04	73.54	36.08	69.82	41.73	73.78	34.60	67.50	31.50	72.40	46.62	67.79	28.50	70.30	44.17	72.29	34.00	67.50	52.80	65.85	38.52	70.19
CoLAP w/ XRCL		87.94	54.40	69.36	43.64	74.90	50.38	74.90	52.68	71.14	47.51	72.33	46.48	66.00	50.48	71.12	50.36	62.61	50.80	71.76	52.50	71.08	50.00	68.96	56.59	67.80	50.48	70.16
CoLAP w/ XCCL		87.94	54.32	69.84	42.51	74.74	50.46	74.90	51.87	71.14	46.53	72.82	46.40	65.84	49.60	71.12	49.55	63.42	50.48	72.32	51.50	71.17	50.08	68.96	56.78	67.61	50.01	70.32
FT	50	85.18	2.48	46.64	3.24	48.91	2.34	48.70	2.68	46.26	2.78	50.45	2.72	44.72	3.20	45.92	2.34	40.72	2.88	48.40	2.92	48.50	2.64	43.04	3.32	50.34	2.79	46.88
CA		85.18	2.40	46.0	2.91	47.13	2.26	50.38	3.09	45.85	2.78	50.2	2.80	44.64	2.96	43.76	2.79	40.18	2.32	46.96	2.75	48.42	2.40	40.96	2.54	50.15	2.67	46.22
PCT		87.38	59.10	74.10	51.01	77.73	59.21	80.23	59.35	76.02	56.12	78.16	51.90	72.90	58.10	75.20	55.63	70.61	58.00	75.80	60.62	77.60	60.50	72.80	53.41	73.17	56.91	75.36
CoLAP w/ XRCL		87.94	63.20	75.12	53.60	76.11	60.50	79.50	57.48	76.18	54.37	78.94	58.72	68.00	59.20	74.56	56.85	66.22	62.80	75.68	62.58	75.58	58.88	70.96	60.78	71.41	59.08	74.02
CoLAP w/ XCCL		87.94	62.88	75.12	52.71	75.71	60.08	79.33	57.48	76.26	54.86	78.86	58.64	68.48	58.96	74.64	56.85	66.22	62.72	75.44	62.67	75.08	59.20	71.28	60.29	71.32	58.94	73.98
FT	100	85.18	3.04	48.88	2.83	49.80	2.76	50.21	2.36	48.94	3.27	50.86	2.88	48.80	3.04	48.40	3.42	42.70	3.12	49.84	3.58	49.67	2.32	46.16	4.10	52.20	3.06	48.87
CA		85.18	2.96	48.72	2.91	50.28	2.93	49.71	2.11	47.64	3.02	50.94	3.04	46.88	3.04	47.68	2.43	42.70	2.72	49.52	3.33	50.00	2.32	45.12	3.71	51.71	2.88	48.41
PCT		87.38	61.90	76.60	57.09	81.68	58.58	83.16	59.96	78.66	60.00	80.31	56.90	76.60	57.50	78.50	62.73	72.30	61.30	78.70	61.25	81.56	59.60	75.90	62.20	75.12	59.92	78.26
CoLAP w/ XRCL		87.94	64.00	77.60	61.30	80.16	62.85	80.08	62.68	79.27	56.41	81.80	55.60	72.40	64.08	77.36	58.20	69.91	67.12	81.68	60.67	78.58	64.24	74.40	62.34	74.24	61.62	77.29
CoLAP w/ XCCL		87.94	63.92	77.76	60.81	79.68	61.84	79.33	62.68	78.94	57.31	81.31	55.04	72.80	63.76	77.28	58.74	70.09	66.72	80.88	60.50	78.42	64.48	74.16	62.15	74.05	61.50	77.06
FT	250	85.18	6.64	57.28	3.81	56.6	3.01	56.23	4.15	54.23	5.96	56.49	6.80	55.04	5.20	56.16	6.94	47.03	5.12	56.96	4.75	55.50	4.96	53.04	8.88	56.88	5.52	55.12
CA		85.18	5.04	55.68	2.67	55.14	2.68	54.90	3.66	53.25	4.49	55.27	5.12	53.6	4.56	54.64	4.41	45.32	5.20	55.12	2.83	54.08	4.40	51.04	7.61	56.20	4.39	53.69
PCT		87.38	68.40	81.60	67.51	84.51	71.76	85.46	67.28	81.71	69.29	83.78	68.00	80.20	65.20	83.00	69.71	78.15	67.40	83.40	70.83	82.40	67.50	81.00	67.20	81.10	68.34	82.19
CoLAP w/ XRCL		87.94	73.04	83.92	71.26	86.56	68.45	82.18	67.56	83.50	65.47	81.71	70.48	81.44	71.20	83.20	65.05	75.32	76.72	87.84	69.33	80.83	72.00	81.52	67.80	77.85	69.86	82.16
CoLAP w/ XCCL		87.94	73.60	84.32	71.66	84.45	68.79	82.18	66.75	84.07	66.12	81.63	70.72	81.28	71.44	83.60	66.94	76.40	75.92	86.96	68.08	79.83	71.44	81.84	67.90	77.37	69.95	81.99

Table 7: Average accuracy per language in MultiTACRED.

Dataset	Label	Exemplar
XNLI	Entailment	Premise: "You look like f–ing hell , Brando , the Star reports he said , and he advised the actor to lose maybe a hundred pounds , pallie ." Hypothesis: "Brando looks like f–ing hell , the Star reported him as saying ."
	Neutral	Premise: "3) Their many enemies want to legislate them out of existence ." Hypothesis: "The Democrats have enemies that want to legislate them out of existence ."
	Neutral	Premise: "Only three now , sir ." Hypothesis: "3 pm now , sir ."
	Contradiction	Premise: "That would certainly help i 'm sure." Hypothesis: "No , that does not help me ."
	Contradiction	Premise: "They ate them , uncle ." Hypothesis: "The people were unharmed ."
AmNLI	Entailment	Premise: "That was my messed up." Hypothesis: "I made a mistake."
	Neutral	Premise: "They asked a few questions and I answered them and they said, Get your baggage and leave there immediately, and come to the address you were supposed to when you arrived in Washington." Hypothesis: "They told me to pick up by white suitcase."
	Neutral	Premise: "That was, that was a pretty scary day." Hypothesis: "It was really scary when the tornado came into town."
	Contradiction	Premise: "Yeah yeah i you know i i wouldn't even mind so much if they had a um corporation that is financed" Hypothesis: "It would make me angry to find out that they had financed the corporation."
	Contradiction	Premise: "Uh-huh all right bye now" Hypothesis: "Let's keep talking."
MultiTACRED	per:origin	"Flower has been exiled from the country since 2003 when he and teammate <E1>Henry Olonga</E1> wore black armbands to protest the ' death of democracy ' in <E2>Zimbabwe</E2>."
	org:city_of_headquarters	"Shares of <E1>Millipore Corp</E1> advanced again Wednesday after the <E2>Billerica maker</E2> of life sciences equipment said it had begun an effort to auction itself off to the highest bidder."
	org:founded	"There are 122 centrally-administered SOEs under the <E1>SASAC</E1> after China National Service Corporation for Chinese Personnel Working Abroad merged with China National Pharmaceutical Group Corporation on <E2>Monday</E2>."
	per:country_of_birth	"Bin Laden 's latest video rant sounds very much like <E1>Adam Gadahn</E1> a <E2>US born</E2> loner and leftist who converted to Islam and joined al Qaeda."
	per:stateorprovinces_of_residence	"For instance , the <E1>Thomas More Law Center</E1> in Ann Arbor , <E2>Mich.</E2> , which is appealing a health care ruling it lost in Detroit , is known for its unsuccessful defense of a Pennsylvania school district that hoped to teach " intelligent design " as an alternative to evolution."

Table 8: Exemplars with highest similarity scores for XNLI, AmNLI, and MultiTACRED.