Multilingual Relation Classification via Efficient and Effective Prompting

Yuxuan Chen **David Harbecke** **Leonhard Hennig**
German Research Center for Artificial Intelligence (DFKI)
Speech and Language Technology Lab
{*yuxuan.chen, david.harbecke, leonhard.hennig*}@*dfki.de*

Abstract

Prompting pre-trained language models has achieved impressive performance on various NLP tasks, especially in low data regimes. Despite the success of prompting in monolingual settings, applying prompt-based methods in multilingual scenarios has been limited to a narrow set of tasks, due to the high cost of handcrafting multilingual prompts. In this paper, we present the first work on prompt-based multilingual relation classification (RC), by introducing an efficient and effective method that constructs prompts from relation triples and involves only minimal translation for the class labels. We evaluate its performance in fully supervised, few-shot and zero-shot scenarios, and analyze its effectiveness across 14 languages, prompt variants, and English-task training in cross-lingual settings. We find that in both fully supervised and few-shot scenarios, our prompt method beats competitive baselines: fine-tuning XLM-R_{EM} and null prompts. It also outperforms the random baseline by a large margin in zero-shot experiments. Our method requires little in-language knowledge and can be used as a strong baseline for similar multilingual classification tasks.

1 Introduction

Relation classification (RC) is a crucial task in information extraction (IE), aiming to identify the relation between entities in a text (Alt et al., 2019). Extending RC to multilingual settings has recently received increased interest (Zou et al., 2018; Kolluru et al., 2022), but the majority of prior work still focuses on English (Baldini Soares et al., 2019; Lyu and Chen, 2021). A main bottleneck for multilingual RC is the lack of supervised resources, comparable in size to large English datasets (Riedel et al., 2010; Zhang et al., 2017). The SMiLER dataset (Seganti et al., 2021) provides a starting point to test fully supervised and more efficient approaches due to different resource availability for different languages.

Previous studies have shown the promising performance of prompting PLMs compared to the data-hungry fine-tuning, especially in low-resource scenarios (Gao et al., 2021; Le Scao and Rush, 2021; Lu et al., 2022). Multilingual pre-trained language models (Conneau et al., 2020; Xue et al., 2021) further enable multiple languages to be represented in a shared semantic space, thus making prompting in multilingual scenarios feasible. However, the study of prompting for multilingual tasks so far remains limited to a small range of tasks such as text classification (Winata et al., 2021) and natural language inference (Lin et al., 2022). To our knowledge, the effectiveness of prompt-based methods for multilingual RC is still unexplored.

To analyse this gap, we pose two research questions for multilingual RC with prompts:
RQ1. What is the most effective way to prompt? We investigate whether prompting should be done in English or the target language and whether to use soft prompt tokens.
RQ2. How well do prompts perform in different data regimes and languages? We investigate the effectiveness of our prompting approach in three scenarios: fully supervised, few-shot and zero-shot. We explore to what extent the results are related to the available language resources.

We present an efficient and effective prompt method for multilingual RC (see Figure 1) that derives prompts from relation triplets (see Section 3.1). The derived prompts include the original sentence and entities and are supposed to be filled with the relation label. We evaluate the prompts with three variants, two of which require no translation, and one of which requires minimal translation, i.e., of the relation labels only. We find that our method outperforms fine-tuning and a strong task-agnostic prompt baseline in fully supervised and few-shot scenarios, especially for relatively low-resource languages. Our method also improves over the random baseline in zero-shot settings, and

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 1059–1075
December 7-11, 2022 ©2022 Association for Computational Linguistics

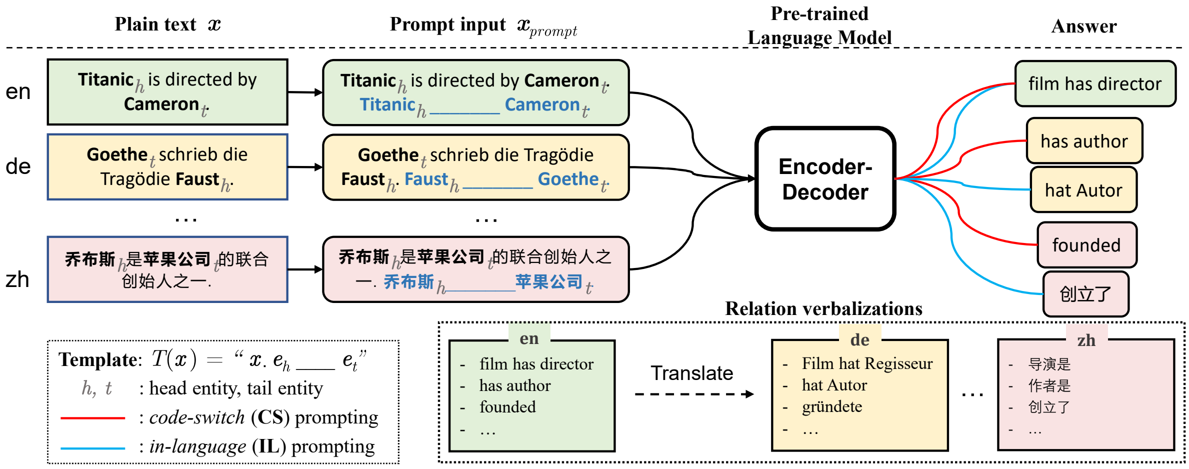
Plain text x
Prompt input x_{prompt}
Pre-trained Language Model
Answer
en
Titanic$_h$ is directed by Cameron$_t$
Titanic$_h$ is directed by Cameron$_t$. Titanic$_h$ ______ Cameron$_t$
de
Goethe$_t$ schrieb die Tragödie Faust$_h$.
Goethe$_t$ schrieb die Tragödie Faust$_h$. Faust$_h$ ______ Goethe$_t$
zh
乔布斯$_h$是苹果公司$_t$的联合创始人之一.
乔布斯$_h$是苹果公司$_t$的联合创始人之一. 乔布斯$_h$ ______ 苹果公司$_t$
Encoder-Decoder
film has director
has author
hat Autor
founded
创立了
Relation verbalizations
en
- film has director
- has author
- founded
- …
Translate
de
- Film hat Regisseur
- hat Autor
- gründete
- …
zh
- 导演是
- 作者是
- 创立了
- …
Template: $T(x)$ = " x. e_h ____ e_t"
h, t : head entity, tail entity
: code-switch (CS) prompting
: in-language (IL) prompting

Figure 1: Overview of our approach. Given a plain text $\boldsymbol{x}$ containing head entity $\boldsymbol{e}_h$ and tail entity $\boldsymbol{e}_t$ from language $\mathcal{L}$, we first apply the template $T(\boldsymbol{x})$ = "$\boldsymbol{x}$. $\boldsymbol{e}_h$ ____ $\boldsymbol{e}_t$" and yield the prompt input with a blank. Then the PLM aims to fill in the relation at the blank. In code-switch prompting, the target sequence is the English relation verbalization. In in-language prompting, the target is the relation name translated into $\mathcal{L}$.

achieves promising cross-lingual performance. The main contributions of this work hence are:

- We propose a simple but efficient prompt method for multilingual RC, which is, to the best of our knowledge, the first work to apply prompt-based methods to multilingual RC (Section 3).
- We evaluate our method on the largest multilingual RC dataset, SMiLER (Seganti et al., 2021), and compare our method with strong baselines in all three scenarios. We also investigate the effects of different prompt variants, including insertion of soft tokens, prompt language, and the word order of prompting (Sections 4 & 5).

2 Preliminaries

We first give a formal definition of the relation classification task, and then introduce fine-tuning and prompting paradigms to perform RC.

2.1 Relation Classification Task Definition

Relation classification is the task of classifying the relationship such as *date_of_birth*, *founded_by* or *parents* between pairs of entities in a given context.

Formally, given a relation set $\mathcal{R}$ and a text $\boldsymbol{x} = [x_1, x_2, \dots, x_n]$ (where $x_1, \cdots, x_n$ are tokens) with two disjoint spans $\boldsymbol{e}_h$ and $\boldsymbol{e}_t$ denoting the head and tail entity, RC aims to predict the relation $r \in \mathcal{R}$ between $\boldsymbol{e}_h$ and $\boldsymbol{e}_t$, or give a *no_relation* prediction if no relation in $\mathcal{R}$ holds. RC is a multilingual task if the token sequences come from different languages.

2.2 Fine-tuning for Relation Classification

In fine-tuning, a task-specific linear classifier is added on top of the PLM. Fine-tuning hence introduces a different scenario from pre-training, since language model (LM) pre-training is usually formalized as a cloze-style task to predict target tokens at `[MASK]` (Devlin et al., 2019; Liu et al., 2019) or a corrupted span (Raffel et al., 2020; Lewis et al., 2020). For the RC task, the classifier aims to predict the target class r at `[CLS]` or at the entity spans denoted by MARKER (Baldini Soares et al., 2019).

2.3 Prompting for Relation Classification

Prompting is proposed to bridge the gap between pre-training and fine-tuning (Liu et al., 2022; Gu et al., 2022). The essence of prompting is, by appending extra text to the original text according to a task-specific template $T(\cdot)$, to reformulate the downstream task to an LM pre-training task such as masked language modeling (MLM), and apply the same training objective during the task-specific training. For the RC task, to identify the relation between "Angela Merkel" and "Joachim Sauer" in the text "Angela Merkel's current husband is quantum chemist Joachim Sauer," an intuitive template for prompting can be "The relation between Angela Merkel and Joachim Sauer is `[MASK]`," and the LM is supposed to assign a higher likelihood to the term *couple* than to e.g. *friends* or *colleagues* at `[MASK]`. This "fill-in the blank" paradigm is well

	Prompt input	Target	Example	
			Input	Target
null prompts	$\boldsymbol{x}$. ____	$\phi^{EN}(r)$	*Goethe* schrieb *Faust*. ____	has author
CS	$\boldsymbol{x}$. $\boldsymbol{e}_h$ ____ $\boldsymbol{e}_t$	$\phi^{EN}(r)$	*Goethe* schrieb *Faust*. *Faust* ____ *Goethe*	has author
SP	$\boldsymbol{x}$. [v$_1$]$\boldsymbol{e}_h$ [v$_2$]____ [v$_3$]$\boldsymbol{e}_t$	$\phi^{EN}(r)$	*Goethe* schrieb *Faust*. [v$_1$]*Faust* [v$_2$]____ [v$_3$]*Goethe*	has author
IL	$\boldsymbol{x}$. $\boldsymbol{e}_h$ ____ $\boldsymbol{e}_t$	$\phi^{\mathcal{L}}(r)$	*Goethe* schrieb *Faust*. *Faust* ____ *Goethe*	hat Autor

Table 1: Overview of the prompts, including null prompts (baseline), and ours with its variants. For each prompt or its variant, we list (1) the prompt input and the target; (2) an example based on the plain text in German "*Goethe* schrieb *Faust*." [v$_\text{i}$]: learnable soft tokens. $\boldsymbol{\phi}^{EN}(r)$: the original (English) relation verbalization. $\phi^{\mathcal{L}}(r)$: the translated relation verbalization into the target language $\mathcal{L}$.

aligned with the pre-training scenario, and enables prompting to better coax the PLMs for pre-trained knowledge (Petroni et al., 2019).

3 Methods

We now present our method, as shown in Figure 1. We introduce its template and verbalizer, and propose several variants of the prompt. Lastly, we explain the training and inference process.

3.1 Template

For prompting (Liu et al., 2022), a prompt often consists of a template $T(\cdot)$ and a verbalizer $\mathcal{V}$. Given a plain text $\boldsymbol{x}$, the template T adds task-related instruction to $\boldsymbol{x}$ to yield the prompt input

$$\boldsymbol{x}_{prompt} = T(\boldsymbol{x}). \qquad (1)$$

Following Chen et al. (2022) and Han et al. (2021), we treat relations as predicates and use the cloze "$\boldsymbol{e}_h$ $\underline{\{relation\}}$ $\boldsymbol{e}_t$" for the LM to fill in. Our template is formulated as

$$T(\boldsymbol{x}) := \text{``}\boldsymbol{x}.\, \boldsymbol{e}_h \;\_\_\_\_\; \boldsymbol{e}_t\text{''}. \qquad (2)$$

In the template $T(\boldsymbol{x})$, $\boldsymbol{x}$ is the original text and the two entities $\boldsymbol{e}_h$ and $\boldsymbol{e}_t$ come from $\boldsymbol{x}$. Therefore, our template does not introduce extra tokens, thus involves no translation at all.

3.2 Verbalizer

After being prompted by $\boldsymbol{x}_{prompt}$, the PLM $\mathcal{M}$ predicts the masked text $\boldsymbol{y}$ at the blank. To complete an NLP classification task, a verbalizer ϕ is required to bridge the set of labels $\mathcal{Y}$ and the set of predicted texts (verbalizations $\mathcal{V}$). For the simplicity of our prompt, we use the one-to-one verbalizer:

$$\phi : \mathcal{Y} \to \mathcal{V}, r \mapsto \phi(r), \qquad (3)$$

where r is a relation, and $\boldsymbol{\phi}(r)$ is the simple verbalization of r. $\phi(\cdot)$ normally only involves splitting r by "-" or "_" and replacing abbreviations such as *org* with *organization*. E.g., the relation *org-has-member* corresponds to the verbalization "organization has member". Then the prediction is formalized as

$$p(r|\boldsymbol{x}) \propto p(\boldsymbol{y} = \phi(r)|\boldsymbol{x}_{prompt}; \theta_{\mathcal{M}}), \qquad (4)$$

where $\theta_{\mathcal{M}}$ denotes the parameters of model $\mathcal{M}$. $p(r|\boldsymbol{x})$ is normalized by the likelihood sum over all relations.

3.3 Variants

To find the optimal way to prompt, we investigate three variants as follows.

Hard prompt vs soft prompt (SP) Hard prompts (a.k.a. discrete prompts) (Liu et al., 2022) are entirely formulated in natural language. Soft prompts (a.k.a. continuous prompts) consist of learnable tokens (Lester et al., 2021) that are not contained in the PLM vocabulary. Following Han et al. (2021), we insert soft tokens before entities and blanks as shown for SP in Table 1.

Code-switch (CS) vs in-language (IL) Relation labels are in English across almost all RC datasets. Given a text from a non-English input $\mathcal{L}$ with a blank, the recovered text is code-mixed after being completed with an English verbalization, corresponding to code-switch prompting. It is probably more reasonable for the PLM to fill in the blank in language $\mathcal{L}$. Inspired by Lin et al. (2022) and Zhao and Schütze (2021), we machine-translate the English verbalizers into the other languages.[1]

[1] See Appendix B for more examples of translated verbalizations. To translate the verbalizer of the SMiLER dataset, we use DeepL by default and Google Translate when the target language is not supported by DeepL (in case of AR, FA, KO and UK).

Task	Dataset	#Class	Verbalizations	# Token in Verb.	
				Mean	Std.
LA	CoLA (Warstadt et al., 2019)	2	`correct, incorrect.` (Gao et al., 2021)	1	0
NER	CoNLL03 (Tjong Kim Sang and De Meulder, 2003)	5	`location, person, not an, ...` (Cui et al., 2021)	1.2	0.4
NLI	MNLI (Williams et al., 2018)	3	`yes, no, maybe.` (Fu et al., 2022)	1	0
NLI	XNLI (Conneau et al., 2018)	3	`yes, no, maybe; Evet, ...` (Zhao and Schütze, 2021)	1	0
PI	PAWS-X (Yang et al., 2019)	2	`yes, no.` (Qi et al., 2022)	1	0
TC	MARC (Keung et al., 2020)	2	`good, {average, bad}.` (Huang et al., 2022)	1	0
RC	TACRED (Zhang et al., 2017)	42	`founded by, city of birth, country of death, ...`	3.23	1.99
	SemEval (Hendrickx et al., 2010)	10	`cause effect, entity origin, product producer, ...`	2.50	0.81
	NYT (Riedel et al., 2010)	24	`ethnicity, major shareholder of, religion, ...`	2.10	1.01
	SciERC (Luan et al., 2018)	6	`conjuction, feature of, part of, used for, ...`	2.17	0.69
	SMiLER (EN) (Seganti et al., 2021)	36	`birth place, starring, won award, ...`	2.58	0.68
	SMiLER (ALL) (Seganti et al., 2021)	36	`hat Genre, chef d'organisation, del país, ...`	3.66	1.44

Table 2: Statistics of the lengths of the verbalizations over several classification tasks. The lengths for non-RC tasks depend on the tokenizers from the respective PLMs in the cited work. The lengths for RC tasks are based on the mT5_{BASE} tokenizer. Mean and std. show that the label space of the RC task is more complex than most few-class classification tasks. The verbalizations of RC datasets are listed in Appendix B. For SemEval, the two possible directions of a relation are combined. For NYT, we use the version from Zeng et al. (2018). For SMiLER, "EN" is the English split; "ALL" contains all data from 14 languages.

Table 1 visualizes both code-switch (CS) and in-language (IL) prompting. For English, CS- and IL-prompting are equivalent, since $\mathcal{L}$ is English itself.

Word order of prompting For the RC task, head-relation-tail triples involve three elements. Therefore, deriving natural language prompts from them requires handling where to put the predicate (relation). In the case of SOV languages, filling in a relation that occurs between $\boldsymbol{e}_h$ and $\boldsymbol{e}_t$ seems less intuitive. Therefore, to investigate if the word order of prompting affects prediction accuracy, we swap the entities and the blank in the SVO-template "$\boldsymbol{x}.\, \boldsymbol{e}_h$ ____ $\boldsymbol{e}_t$" and get "$\boldsymbol{x}.\, \boldsymbol{e}_h\ \boldsymbol{e}_t$ ____" as the SOV-template.

3.4 Training and Inference

The training and inference setups depend on the employed model. Prompting autoencoding language models requires the verbalizations to be of fixed length, since the length of masks, which is identical with verbalization length, is unknown during inference. Encoder-decoders can handle verbalizations of varying length by nature (Han et al., 2022; Du et al., 2022). Han et al. (2021) adjust all the verbalizations in TACRED to a length of 3, to enable prompting with RoBERTa for RC. We argue that for multilingual RC, this fix is largely infeasible, because: (1) in case of in-language prompting on SMiLER, the variance of the length of the verbalizations increases from 0.68 to 1.44 after translation (see Table 2), and surpasses most of listed monolingual RC datasets (SemEval, NYT and SciERC), making it harder to unify the length; (2) manually adjusting the translated prompts requires manual effort per target language, making it much more expensive than adjusting only English verbalizations. Therefore, we use an encoder-decoder PLM for prompting (Han et al., 2022; Song et al., 2022).

Training objective For an encoder-decoder PLM $\mathcal{M}$, given the prompt input $T(\boldsymbol{x})$ and the target sequence $\phi(r)$ (i.e. label verbalization), we denote the output sequence as y. The probability of an exact-match decoding is calculated as follows:

$$\prod_{t=1}^{|\phi(r)|} P_\theta\left(y_t = \phi_t(r) | y_{<t}, T(\boldsymbol{x})\right), \tag{5}$$

where y_t, $\phi_t(r)$ denote the t-th token of y and $\phi(r)$, respectively. $y_{<t}$ denotes the decoded sequence on the left. θ represents the set of all the learnable parameters, including those of the PLM $\theta_{\mathcal{M}}$, and those of the soft tokens θ_{sp} in case of variant "soft prompt". Hence, the final objective over the training set $\mathcal{X}$ is to minimize the negative log-likelihood:

$$\begin{aligned}\underset{\theta}{\operatorname{argmin}} & -\frac{1}{|\mathcal{X}|}\sum_{\boldsymbol{x}\in\mathcal{X}}\sum_{t=1}^{|\phi(r)|} \\ & \log P_\theta\left(y_t = \phi_t(r) | y_{<t}, T(\boldsymbol{x})\right).\end{aligned} \tag{6}$$

Inference We collect the output logits of the decoder, $\mathsf{L} \in \mathbb{R}^{|V|\times L}$, where $|V|$ is the vocabulary size of $\mathcal{M}$, and L is the maximum decode length. For each relation $r \in \mathcal{R}$, its score is given by (Han et al., 2022):

$$\text{score}_\theta(r) := \frac{1}{|\phi(r)|}\sum_{t=1}^{|\phi(r)|} P_\theta(y_t = \phi_t(r)), \tag{7}$$

Lang.	Fine-tuning data			Pre-train tokens	
	#Class	#Train(K)	Max.	mT5(B)	XLM-R(B)
AR	9	9.3	74	57	2.9
DE	22	51.5	84	347	10.3
EN	36	267.6	110	2733	55.6
ES	21	11.1	70	433	9.4
FA	8	2.6	93	52	13.3
FR	22	60.9	83	318	9.8
IT	22	74.0	86	162	5.0
KO	28	18.7	95	26	5.6
NL	22	38.9	76	73	5.0
PL	21	16.8	86	130	6.5
PT	22	43.3	82	146	8.4
RU	8	6.4	69	713	23.4
SV	22	4.5	84	45	0.08
UK	7	1.0	65	41	0.006

Table 3: Statistics of the 14 languages in the SMiLER dataset, including the number of classes, the number of training examples (in thousands), and the maximum text length over train and test splits. Appended to the table are the sizes (in billion tokens) of pre-training corpora of the referred languages for mT5 and XLM-R, respectively.

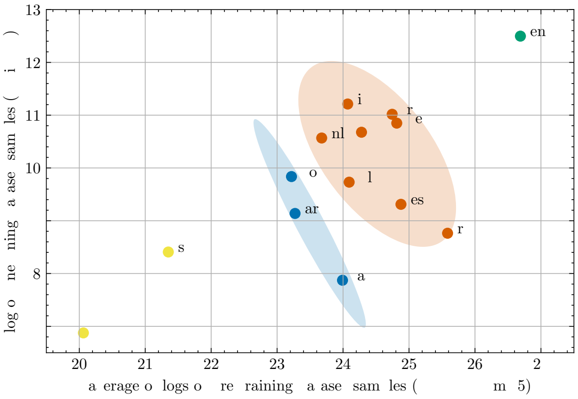

Figure 2: Pre-training and fine-tuning dataset size by language. Four languages groups are distinguishable: English (green) has by far the largest dataset, many other European languages (orange) have large datasets for pre-training and fine-tuning. The three non-European languages (blue) have either less pre-training or fine-tuning data and lowest resource are Swedish and Ukrainian (yellow).

where we compute P by looking up in the t-th column of L and applying softmax at each time step t. We aggregate P by addition to encourage partial matches as well, instead of enforcing exact matches. The score is normalized by the length of verbalization in order to avoid predictions favoring longer relations. Finally, we select the relation with the highest score as prediction.

4 Experiments

We implement our experiments using the Hugging Face Transformers library (Wolf et al., 2020), Hydra (Yadan, 2019) and PyTorch (Paszke et al., 2019).[2] We use micro-F1 as the evaluation metric, as the SMiLER paper (Seganti et al., 2021) suggests. To measure the overall performance over multiple languages, we report the macro average across languages, following Zhao and Schütze (2021) and Lin et al. (2022). We also group the languages by their available resources in both pre-training and fine-tuning datasets for additional aggregate results. Details of the dataset, the models, and the experimental setups are as follows. Further experimental details are listed in Appendix A.

4.1 Dataset

We conduct an experimental evaluation of our multilingual prompt methods on the SMiLER (Seganti et al., 2021) dataset, which contains 1.1M annotated texts across 14 languages.[3] Table 3 lists the main statistics of the different languages in the SMiLER dataset. Note that languages have varying number of relations, mostly related to how many samples are present. We do not evaluate other datasets because the only prior multilingual RC dataset that fits our task, RELX (Köksal and Özgür, 2020), contains only 502 parallel examples in 5 languages.

Grouping of the languages We visualize the languages in Figure 2 based on the sizes of RC training data, but include the pre-training data as well, to give a more comprehensive overview of the availability of resources for each language. We divide the 14 languages into 4 groups, according to the detectable clusters in Figure 2 and language origins.

4.2 Model

For prompting, we use mT5$_{\text{BASE}}$ (Xue et al., 2021), an encoder-decoder PLM that supports 101 languages, including all languages in SMiLER. mT5$_{\text{BASE}}$ (Xue et al., 2021) has 220M parameters.

4.3 Baselines

EN(B) (Seganti et al., 2021) EN(B) is the baseline proposed together with the SMiLER dataset.

[2] We make our code publicly available at `https://github.com/DFKI-NLP/meffi-prompt` for better reproducibility.

[3] Note that SMiLER contains 3 versions of the English split: *en* (268K training examples), *en-small* (36K) and *en-full* (744K). We use the *en* version by default, unless specified otherwise.

They fine-tune $\text{BERT}_{\text{BASE}}$ on the English training split and report the micro-F1 on the English test split. $\text{BERT}_{\text{BASE}}$ has 110M parameters.

XLM-$\mathbf{R}_{\mathbf{EM}}$ To provide a fine-tuning baseline, we re-implement BERT_{EM} (Baldini Soares et al., 2019) with the ENTITY START variant.[4] In this method, the top-layer representations at the starts of the two entities are concatenated for linear classification. To adapt BERT_{EM} to multilingual tasks, we change the PLM from BERT to a multilingual autoencoder, XLM-R_{BASE} (Conneau et al., 2020), and refer to this model as XLM-R_{EM}. XLM-R_{BASE} has 125M parameters.

Null prompts (Logan IV et al., 2022) To better verify the effectiveness of our method, we implement null prompts as a strong task-agnostic prompt baseline. Null prompts involve minimal prompt engineering by directly asking the LM about the relation, without giving any task instruction (see Table 1). Logan IV et al. (2022) show that null prompts surprisingly achieve on-par performance with handcrafted prompts on many tasks. For best comparability, we use the same PLM mT5_{BASE}.

4.4 Fully Supervised Setup

We evaluate the performance of XLM-R_{EM}, null prompts, and our method on each of the 14 languages, after training on the full train split from that language. The prompt input and target of null prompts and our prompts are listed in Table 1.

We employ the randomly generated seed 319 for all the evaluated methods. For XLM-R_{EM}, we follow Baldini Soares et al. (2019) and set the batch size to be 64, the optimizer to be Adam with the learning rate 3×10^{-5} and the number of epochs to be 5. For null prompts and ours, we use AdamW as the optimizer with the learning rate 3×10^{-5}, as Zhang et al. (2022) suggest for most of the sequence-to-sequence tasks, the number of epochs to 5, and batch size to 16. The maximum sequence length is 256 for all methods.

4.5 Few-shot Setup

Few-shot learning is normally cast as a K-shot problem, where K labelled examples per class are available. We follow Chen et al. (2022) and Han et al. (2021), and evaluate on 8, 16 and 32 shots.

The few-shot training set $\mathcal{D}_{train}$ is generated by randomly sampling K instances per relation from the training split. The test set $\mathcal{D}_{test}$ is the original test split from that language. We follow Gao et al. (2021) and sample another K-shot set from the English train split as validation set $\mathcal{D}_{val}$. We tune hyperparameters on $\mathcal{D}_{val}$ for the English task, and apply these to all languages.

We evaluate the same methods as in the fully supervised scenarios, but repeat 5 runs as suggested in Gao et al. (2021), and report the mean and standard deviation of micro-F1. We use a fixed set of random seeds {13, 36, 121, 223, 319} for data generation and training across the 5 runs. For XLM-R_{EM}, we use the same hyperparameters as Baldini Soares et al. (2019), a batch size of 256, and a learning rate of 1×10^{-4}. For null prompts and our prompts, we set the learning rate to 3×10^{-4}, batch size to 16, and the number of epochs to 20.

4.6 Zero-shot Setup

We consider two scenarios for zero-shot multilingual relation classification.

Zero-shot in-context learning Following Kojima et al. (2022), we investigate whether PLMs are also decent zero-shot reasoners for RC. This scenario does not require any samples or training. We test the out-of-the-box performance of the PLM by directly prompting it with $\boldsymbol{x}_{prompt}$. Zero-shot in-context learning does not specify further hyperparameters since it is training-free.

Zero-shot cross-lingual transfer In this scenario, following Krishnan et al. (2021), we fine-tune the model with in-language prompting on the English train split, and then conduct zero-shot in-context tests with this fine-tuned model on other languages using code-switch prompting. Through this setting, we want to verify if task-specific pretraining in a high-resource language such as English helps in other languages. In zero-shot cross-lingual transfer, we use the same hyperparameters and random seed to fine-tune on the English task.

5 Results and Discussion

We first present the results in fully supervised, few-shot and zero-shot scenarios, and then discuss the main findings for answering the research questions in Section 1.

5.1 Fully Supervised Results

Table 4 presents the experimental results in the fully supervised scenario, for different methods, languages, and language groups. We see that all

[4] We also open-source our implementation of XLM-R_{EM} at https://github.com/DFKI-NLP/mtb-bert-em.

Method	AR	DE	EN	ES	FA	FR	IT	KO	NL	PL	PT	RU	SV	UK	EN	H	M	L	$\overline{X}$
EN(B)	-	-	94.9	-	-	-	-	-	-	-	-	-	-	-	94.9	-	-	-	-
XLM-R$_{\text{EM}}$	**98.4**	**95.7**	95.9	27.9	0.0	82.6	**98.9**	64.6	92.2	**97.4**	**97.4**	**96.9**	2.2	5.1	95.9	86.1	54.3	3.7	68.2
null prompts	85.5	81.6	84.7	59.8	71.2	82.6	84.2	63.3	71.4	49.4	12.9	84.9	48.9	46.2	84.7	65.8	73.3	47.6	66.2
CS	95.1	95.4	96.0	74.7	69.2	**97.2**	98.3	82.1	**96.9**	94.8	95.3	87.6	48.9	46.2	96.0	**92.5**	82.1	47.6	84.1
SP	95.1	88.5	**96.1**	**81.1**	65.4	97.0	97.1	83.1	59.9	95.6	96.9	87.3	**63.0**	51.3	**96.1**	87.9	81.2	57.2	82.7
IL	94.1	94.0	96.0	70.5	**73.1**	97.2	97.0	**83.2**	93.5	93.0	85.2	83.3	58.7	**71.8**	96.0	89.2	**83.5**	**65.2**	**85.0**

Table 4: Fully-supervised results in micro-F1 (%) on the SMiLER dataset. The evaluated methods are the proposed baseline EN(B) (Seganti et al., 2021), XLM-R$_{\text{EM}}$, null prompts, and ours. EN, H, M, L: macro average across the languages within the respective group. $\overline{X}$: macro average across all 14 languages. Our variants outperform all baselines along all groups averages, XLM-R$_{\text{EM}}$ has good results for many high-resource languages. Overall, in-language prompting performs best, especially for lower-resource languages.

the three variants of our method beat the fine-tuning baseline XLM-R$_{\text{EM}}$ and the prompting baseline null prompts, according to the macro-averaged performance across 14 languages. In-language prompting delivers the most promising result, achieving an average F_1 of 85.0, which is higher than XLM-R$_{\text{EM}}$ (68.2) and null prompts (66.2). The other two variants, code-switch prompting with and w/o soft tokens, achieve F_1 scores of 84.1 and 82.7, respectively, only 0.9 and 2.3 lower than in-language. All three prompt variants are hence effective in fully supervised scenarios.

On a per-group basis, we find that the lower-resourced a language is, the greater an advantage prompting enjoys against fine-tuning. In particular, in-language prompts shows better robustness compared to XLM-R$_{\text{EM}}$ in low-resource languages. They both yield 95.9-96.0 F_1 scores for English, but XLM-R$_{\text{EM}}$ decreases to 54.3 and 3.7 F_1 in Group-M and -L, while in-language prompting still delivers 83.5 and 65.2 F_1.

5.2 Few-shot Results

Table 5 presents the per-group results in few-shot experiments. All the methods benefit from larger K. Similarly, in-language prompting still turns out to be the best contender, performing 1st in 8- and 32-shot, and the 2nd in 16-shot. We see that in-language outperforms XLM-R$_{\text{EM}}$ in all K-shots, while code-switch achieves comparable or even lower F_1 to XLM-R$_{\text{EM}}$ for $K = 8$, suggesting that the choice of prompt affects the few-shot performance greatly, thus needs careful consideration.

On a per-group basis, we find that in-language prompting outperforms other methods for middle- and low-resourced languages. Similar observations can also be drawn from fully supervised results. We conclude that, with sufficient supervision, in-language is the optimal variant to prompt rather than code-switch. We hypothesize it is due to the pre-training scenario, where the PLM rarely sees code-mixed text (Santy et al., 2021).

Shots	Method	EN	H	M	L	$\overline{X}$
8	XLM-R$_{\text{EM}}$	31.8	**43.0**	27.5	6.6	33.7
	null prompts	37.4	27.6	26.6	37.4	29.5
	CS	42.2	30.6	27.8	38.4	32.0
	SP	**45.4**	27.8	17.9	33.6	27.4
	IL	42.2	40.5	**38.3**	**43.4**	**40.6**
16	XLM-R$_{\text{EM}}$	**56.4**	**56.9**	34.1	10.4	45.3
	null prompts	42.1	31.6	34.3	49.7	35.5
	CS	50.5	50.1	41.9	53.9	**48.9**
	SP	53.7	46.7	38.4	49.0	45.8
	IL	50.5	45.2	**42.1**	**54.6**	46.3
32	XLM-R$_{\text{EM}}$	73.2	62.4	44.4	6.5	51.3
	null prompts	56.0	36.4	47.7	53.9	42.7
	CS	**80.9**	57.0	**65.1**	59.4	60.8
	SP	61.2	53.5	46.3	63.1	53.9
	IL	**80.9**	**63.6**	64.2	**67.4**	**65.5**

Table 5: Few-shot results by group in micro-F1 (%) on the SMiLER (Seganti et al., 2021) dataset averaged over five runs. We macro-average results for each language group (see Figure 2) and over all languages ($\overline{X}$). In-language prompting performs best in most settings and language groups. Our variants are especially strong for medium- and lower-resource language groups. See Table 7 in Appendix C for detailed results with mean and std. for each language.

5.3 Zero-shot Results

Table 6 presents the per-language results in zero-shot scenarios. We consider the random baseline for comparison (Zhao and Schütze, 2021; Winata et al., 2021). We notice that performance of the random baseline varies a lot across languages, since the languages have different number of classes in the dataset (cf. Table 3), with English being the hardest task.

For zero-shot in-context, code-switch prompting always outperforms the random baseline by a large

		EN	AR	DE	ES	FA	FR	IT	KO	NL	PL	PT	RU	SV	UK
Random		2.8	11.1	4.6	4.8	12.5	4.6	4.6	3.6	4.6	4.8	4.6	12.5	4.6	14.3
Zero-Shot In-Context Learning															
SVO	CS	**5.5**	**69.9**	**10.4**	12.7	38.5	**13.3**	11.2	**10.0**	**12.4**	**14.0**	8.1	52.3	27.2	**51.3**
	IL		2.2	5.2	1.8	5.3	9.2	1.3	3.6	7.6	9.0	1.7	7.1	5.4	25.6
SOV	CS	4.8	68.4	10.0	**13.2**	36.9	12.3	**12.6**	5.0	11.8	13.4	**10.3**	**52.6**	**29.4**	**51.3**
	IL		3.8	5.0	3.6	**59.8**	7.7	1.3	3.1	10.0	7.9	1.4	6.0	4.5	25.6
Zero-Shot Cross-Lingual Transfer															
EN (268K)		-	94.0	94.9	91.7	91.1	96.0	97.5	78.2	97.5	93.3	95.2	93.8	97.8	94.7
EN-small (36K)		-	45.9	64.7	73.1	70.3	82.2	77.5	30.8	79.9	59.0	67.3	76.1	77.2	54.1

Table 6: Zero-shot results in micro-F1 (%) on the SMiLER dataset. "SVO" and "SOV": word order of prompting. Overall, Code-switch prompting performs the best in the zero-shot in-context scenario. In cross-lingual transfer experiments, English-task training greatly improves the performance on all the other 13 languages.

margin, in both word orders, while in-language prompting performs worse than the random baseline in 6 languages. Code-switch prompting outperforms in-language prompting across all the 13 non-English languages, using SVO-template. We assume that, without in-language training, the PLM understands the task best when prompted in English. The impressive performance of code-switch shows the PLM is able to transfer its pre-trained knowledge in English to other languages. We also find that the performance is also highly indicated by the number of classes, with worst F_1 scores achieved in EN, KO and PT (36, 28 and 22 classes), and best scores in AR, RU and UK (9, 8 and 7 classes). In addition, we observe that word order does not play a significant role for most languages, except for FA, which is an SOV-language and has 54.5 F_1 gain from in-language prompting with an SOV-template.

For zero-shot cross-lingual transfer, we see that non-English tasks benefit from English in-domain prompt-based fine-tuning, and the F_1 gain improves with the English data size. For 5 languages (ES, FA, NL, SV, and UK), zero-shot transfer after training on 268k English examples delivers even better results than in-language fully supervised training (cf. Table 4). Sanh et al. (2022) show that including RC-specific prompt input in English during pre-training can help in other languages.

5.4 Discussion

Based on the results above, we answer the research questions from Section 1.

RQ1. Which is the most effective way to prompt? In the fully-supervised and few-shot scenario, in-language prompting displays the best results. This appears to stem from a solid performance across all languages in both settings. Its worst performance is 31.8 F_1 for Polish 8-shot (see Table 7 in Appendix C). All other methods have results lower than 15.0 F_1 for some language. This indicates that with little supervision mT5 is able to perform the task when prompted in the language of the original text. However, zero-shot results strongly prefer code-switch prompting. It could follow that, without fine-tuning, the model's understanding of this task is much better in English.

RQ2. How well does our method perform in different data regimes and languages? Averaged over all languages, all our variants outperform the baselines, except for 8-shot. For some high-resource languages, XLM-R_{EM} is able to outperform our method. On the other hand, for low-resource languages null prompts are a better baseline which we consistently outperform. This could indicate that prompting the underlying mT5 model is better suited for multilingual RC on SMiLER. Overall, the results suggest that minimal translation can be very helpful for multilingual relation classification.

6 Related Work

Multilingual relation classification Previous work in multilingual RC has primarily focused on traditional methods rather than prompting PLMs. Faruqui and Kumar (2015) machine-translate non-English full text to English to deal with multilinguality. Akbik et al. (2016) employ a shared semantic role labeler to get language-agnostic abstraction and apply rule-based methods to classify the unified abstractions. Lin et al. (2017) employ convolutional networks to extract relation embeddings from texts, and propose cross-lingual attention between relation embeddings to model cross-lingual information consistency. Sanh et al. (2019) leverage the embeddings from BiLSTM, which is trained with a set of selected semantic tasks to help (multilingual) relation extraction. Köksal and Özgür (2020) fine-tune (multilingual) BERT, classifying the embedding at [CLS]. To take entity-related embeddings into consideration as well, Nag et al. (2021) add an extra summarization layer on top of a multilingual BERT to collect and pool the embeddings at both [CLS] and entity starts.

Multilingual prompting Multilingual prompting is a new yet fast-growing topic. Winata et al. (2021) reduce handcrafting efforts by reformulating general classification tasks into binary classification with answers restricted to true or false for all languages. Huang et al. (2022) propose a unified multilingual prompt by introducing a so-called "two-tower" encoder, with the template tower producing language-agnostic prompt representation, and the context tower encoding text information. Fu et al. (2022) manually translate prompts and suggest multilingual multitask training to boost the performance for a target downstream task.

7 Conclusion

In this paper, we present a first, simple yet efficient and effective prompt method for multilingual relation classification, by translating only the relation labels. Our prompting outperforms fine-tuning and null prompts in fully supervised and few-shot experiments. With supervised data, in-language prompting enjoys the best performance, while in the zero-shot scenarios prompting in English is preferable. We attribute the good performance of our method to its well-suitedness for RC, with the derivation of *entity*$_1$-*relation*-*entity*$_2$ prompts from relation triples. We would like to see our method extended to similar tasks, such as semantic role labeling, with a structure between concepts that can be described in natural language.

Limitations

We acknowledge the main limitation of this work is that we only experiment on one dataset with 14 languages. Multilingual RC datasets prior to SMiLER are limited in the coverage of languages or in the size of unique training examples. It would be interesting to see how our method performs on other multilingual RC datasets, especially for underrepresented languages (Winata et al., 2022).

We restrict the target language to be supported by the underlying PLM. The popular multilingual PLMs, mT5 and mBART, include 101 and 25 languages during pre-training. We rely on these PLMs and fail to study true low-resource languages that are not represented in such PLMs (Aji et al., 2022).

It is noticeable that in the fully supervised scenario, for 7 out of the 14 languages, at least one method achieves over 0.95 micro-F_1 score. We hypothesize that is due to high homogeneity in and between the train and test split. If so, the dataset itself might not be challenging, which could indicate that the results are mostly measuring how well the model is able to fit a few indicators (quickly).

Like most other prompt methods, ours requires the label names to be natural language which are indicative of the class. Therefore, our method would suffer from labels being non-descriptive.

Ethics Statement

We use automated machine translation by Google Translate and DeepL for our method. These MT systems contain biases regarding, e.g., gender ("has-author": "hat Autor") where gender-neutral English nouns are translated to gendered nouns in target languages.

In this work we evaluate SMiLER (Seganti et al., 2021), which is crawled from Wikipedia. In the paper, they have not stated measures that prevent collecting sensitive text. Therefore, we do not rule out the possible risk of sensitive content in the data.

The PLMs involved in this paper are BERT$_{\text{BASE}}$ for EN(B), XLM-R$_{\text{BASE}}$ for XLM-R$_{\text{EM}}$, and mT5$_{\text{BASE}}$ for null prompts and ours. BERT$_{\text{BASE}}$ is pre-trained on the BooksCorpus (Zhu et al., 2015) and English Wikipedia. XLM-R$_{\text{EM}}$ is pre-trained on a CommonCrawl corpus. mT5$_{\text{BASE}}$ is pre-trained on mC4, a filtered CommonCrawl corpus. All our published models may have inherited biases from these corpora.

Acknowledgments

We would like to thank Nils Feldhus and the anonymous reviewers for their valuable comments and feedback on the paper. This work has been supported by the German Federal Ministry for Economic Affairs and Climate Action as part of the project PLASS (01MD19003E), and by the German Federal Ministry of Education and Research as part of the projects CORA4NLP (01IW20010) and BBDC2 (01IS18025E).

References

Alham Fikri Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Timothy Baldwin, Jey Han Lau, and Sebastian Ruder. 2022. One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7226–7249, Dublin, Ireland. Association for Computational Linguistics.

Alan Akbik, Laura Chiticariu, Marina Danilevsky, Yonas Kbrom, Yunyao Li, and Huaiyu Zhu. 2016. Multilingual information extraction with PolyglotIE. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 268–272, Osaka, Japan. The COLING 2016 Organizing Committee.

Christoph Alt, Marc Hübner, and Leonhard Hennig. 2019. Fine-tuning pre-trained transformer language models to distantly supervised relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1388–1398, Florence, Italy. Association for Computational Linguistics.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy. Association for Computational Linguistics.

Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 2778–2788, New York, NY, USA. Association for Computing Machinery.

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.

Leyang Cui, Yu Wu, Jian Liu, Sen Yang, and Yue Zhang. 2021. Template-based named entity recognition using BART. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1835–1845, Online. 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.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: general language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 320–335. Association for Computational Linguistics.

Manaal Faruqui and Shankar Kumar. 2015. Multilingual open relation extraction using cross-lingual projection. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1351–1356, Denver, Colorado. Association for Computational Linguistics.

Jinlan Fu, See-Kiong Ng, and Pengfei Liu. 2022. Polyglot prompt: Multilingual multitask promptraining. *CoRR*, abs/2204.14264.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. 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 3816–3830, Online. Association for Computational Linguistics.

Yuxian Gu, Xu Han, Zhiyuan Liu, and Minlie Huang. 2022. PPT: Pre-trained prompt tuning for few-shot learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8410–8423, Dublin, Ireland. Association for Computational Linguistics.

Jiale Han, Shuai Zhao, Bo Cheng, Shengkun Ma, and Wei Lu. 2022. Generative prompt tuning for relation classification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, Online and Abu Dhabi, the United Arab Emirates. Association for Computational Linguistics.

Xu Han, Weilin Zhao, Ning Ding, Zhiyuan Liu, and Maosong Sun. 2021. PTR: prompt tuning with rules for text classification. *CoRR*, abs/2105.11259.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38, Uppsala, Sweden. 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. *CoRR*, abs/2202.11451.

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

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *ICML 2022 Workshop on Knowledge Retrieval and Language Models*.

Abdullatif Köksal and Arzucan Özgür. 2020. The RELX dataset and matching the multilingual blanks for cross-lingual relation classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 340–350, Online. Association for Computational Linguistics.

Keshav Kolluru, Muqeeth Mohammed, Shubham Mittal, Soumen Chakrabarti, and Mausam . 2022. Alignment-augmented consistent translation for multilingual open information extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2502–2517, Dublin, Ireland. Association for Computational Linguistics.

Jitin Krishnan, Antonios Anastasopoulos, Hemant Purohit, and Huzefa Rangwala. 2021. Multilingual code-switching for zero-shot cross-lingual intent prediction and slot filling. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 211–223, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Teven Le Scao and Alexander Rush. 2021. How many data points is a prompt worth? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2627–2636, Online. Association for Computational Linguistics.

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

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

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

Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2017. Neural relation extraction with multi-lingual attention. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 34–43, Vancouver, Canada. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2022. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys (CSUR)*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Robert L. Logan IV, Ivana Balazevic, Eric Wallace, Fabio Petroni, Sameer Singh, and Sebastian Riedel. 2022. Cutting down on prompts and parameters: Simple few-shot learning with language models. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2824–2835. Association for Computational Linguistics.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language*

Processing, pages 3219–3232, Brussels, Belgium. Association for Computational Linguistics.

Shengfei Lyu and Huanhuan Chen. 2021. Relation classification with entity type restriction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 390–395, Online. Association for Computational Linguistics.

Arijit Nag, Bidisha Samanta, Animesh Mukherjee, Niloy Ganguly, and Soumen Chakrabarti. 2021. A data bootstrapping recipe for low-resource multilingual relation classification. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 575–587, Online. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473, Hong Kong, China. 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.

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

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 148–163. Springer.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Victor Sanh, Thomas Wolf, and Sebastian Ruder. 2019. A hierarchical multi-task approach for learning embeddings from semantic tasks. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Sebastin Santy, Anirudh Srinivasan, and Monojit Choudhury. 2021. BERTologiCoMix: How does code-mixing interact with multilingual BERT? In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 111–121, Kyiv, Ukraine. Association for Computational Linguistics.

Alessandro Seganti, Klaudia Firląg, Helena Skowronska, Michał Satława, and Piotr Andruszkiewicz. 2021. Multilingual entity and relation extraction dataset and model. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1946–1955, Online. Association for Computational Linguistics.

Haoyu Song, Li Dong, Weinan Zhang, Ting Liu, and Furu Wei. 2022. CLIP models are few-shot learners: Empirical studies on VQA and visual entailment. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6088–6100, Dublin, Ireland. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Pascale Fung, Timothy Baldwin, Jey Han Lau, Rico Sennrich, and Sebastian Ruder. 2022. Nusax: Multilingual parallel sentiment dataset for 10 indonesian local languages. *CoRR*, abs/2205.15960.

Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. 2021. Language models are few-shot multilingual learners. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 1–15, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

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.

Omry Yadan. 2019. Hydra - a framework for elegantly configuring complex applications. Github.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Xiangrong Zeng, Daojian Zeng, Shizhu He, Kang Liu, and Jun Zhao. 2018. Extracting relational facts by an end-to-end neural model with copy mechanism. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 506–514, Melbourne, Australia. Association for Computational Linguistics.

Cheng Zhang, Hao Zhang, and Jie Wang. 2022. Downstream transformer generation of question-answer pairs with preprocessing and postprocessing pipelines. *CoRR*, abs/2205.07387.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark. Association for Computational Linguistics.

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

Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 19–27. IEEE Computer Society.

Bowei Zou, Zengzhuang Xu, Yu Hong, and Guodong Zhou. 2018. Adversarial feature adaptation for cross-lingual relation classification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 437–448, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

A Experimental Details

A.1 Hyperparameter Search

We investigated the following possible hyperparameters for few-shot settings. For fully-supervised, we take hyperparameters from literature (see Section 4.4).

Number of epochs: $[10, 20]$; Learning rate: $[1 \times 10^{-5}, 3 \times 10^{-5}, 1 \times 10^{-4}, 3 \times 10^{-4}]$. Batch size: $[16, 64, 256]$, not tuned but selected based on available GPU VRAM.

We manually tune these hyperparameters, based on the micro-F_1 score on the validation set.

A.2 Computing Infrastructure

Fully supervised experiments are conducted on a single A100-80GB GPU. Few-shot and zero-shot experiments are conducted on a single A100 GPU.

A.3 Average Running Time

Fully supervised It takes 5 hours to train for 1 run with mT5$_{\text{BASE}}$ and a prompt method (null prompts, CS, SP and IL) on either English, or all other languages in total. With XLM-R$_{\text{EM}}$ the running time is 3 hours.

Few-shot It takes 20 (8-shot), 26 (16-shot), and 36 minutes (32-shot) for 1 run with mT5$_{\text{BASE}}$ and a prompt method over all languages. With XLM-R$_{\text{EM}}$ the running time is 8 minutes.

Zero-shot For zero-shot in-context experiments, it takes 6 minutes with mT5$_{\text{BASE}}$ and a prompt method over all languages. For zero-shot cross-lingual transfer, the running time equals English training time (5 hours) plus inference-only time (6 minutes).

B Verbalizers for SMiLER

- **EN** "birth-place": "birth place", "eats": "eats", "event-year": "event year", "first-product": "first product", "from-country": "from country", "has-author": "has author", "has-child": "has child", "has-edu": "has education", "has-genre": "has genre", "has-height": "has height", "has-highest-mountain": "has highest mountain", "has-length": "has length", "has-lifespan": "has lifespan", "has-nationality": "has nationality", "has-occupation": "has occupation", "has-parent": "has parent", "has-population": "has population", "has-sibling": "has sibling", "has-spouse": "has spouse", "has-tourist-attraction": "has tourist attraction", "has-type": "has type", "has-weight": "has weight", "headquarters": "headquarters", "invented-by": "invented by", "invented-when": "invented when", "is-member-of": "is member of", "is-where": "located in", "loc-leader": "location leader", "movie-has-director": "movie has director", "no_relation": "no relation", "org-has-founder": "organization has founder", "org-has-member": "organization has member", "org-leader": "organization leader", "post-code": "post code", "starring": "starring", "won-award": "won award";

- **AR** "event-year": "سنة الحدث", "has-edu": "لديه تعليم", "has-genre": "له النوع", "has-occupation": "لديه احتلال", "has-population": "عدد السكان", "has-type": "لديه نوع", "is-member-of": "عضو في", "no_relation": "لا علاقة", "won-award": "فاز بالجائزة";

- **DE** "birth-place": "Geburtsort", "event-year": "Veranstaltungsjahr", "from-country": "vom Land", "has-author": "hat Autor", "has-child": "hat Kind", "has-edu": "hat Bildung", "has-genre": "hat Genre", "has-occupation": "hat Beruf", "has-parent": "hat Elternteil", "has-population": "hat Bevölkerung", "has-spouse": "hat Ehepartner", "has-type": "hat Typ", "headquarters": "Hauptsitz", "is-member-of": "ist Mitglied von", "is-where": "gelegen in", "loc-leader": "Standortleiter", "movie-has-director": "Film hat Regisseur", "no_relation": "keine Beziehung", "org-has-founder": "Organisation hat Gründer", "org-has-member": "Organisation hat Mitglied", "org-leader": "Organisationsleiter", "won-award": "gewann eine Auszeichnung";

- **ES** "birth-place": "lugar de nacimiento", "event-year": "año del evento", "from-country": "del país", "has-author": "tiene autor", "has-child": "tiene hijo", "has-edu": "tiene educación", "has-genre": "tiene género", "has-occupation": "tiene ocupación", "has-parent": "tiene padre", "has-population": "tiene población", "has-spouse": "tiene cónyuge", "has-type": "tiene tipo", "headquarters": "sede central", "is-member-of": "es miembro de", "is-where": "situado en", "loc-leader": "líder de ubicación", "movie-has-director": "película cuenta con el director", "no_relation": "sin relación", "org-has-founder": "organización cuenta con el fundador", "org-has-member": "organización tiene miembro", "won-award": "ganó el premio";

- **FA** "event-year": "سال رویداد", "has-edu": "تحصیلات دارد", "has-genre": "ژانر دارد", "has-occupation": "شغل دارد", "has-population": "جمعیت دارد", "has-type": "نوع دارد", "is-member-of": "عضو است", "no_relation": "هیچ رابطه ای";

- **FR** "birth-place": "lieu de naissance", "event-year": "année de l'événement", "from-country": "du pays", "has-author": "a un auteur", "has-child": "a un enfant", "has-edu": "a une éducation", "has-genre": "a un genre", "has-occupation": "a une profession", "has-parent": "a un parent", "has-population": "a de la population", "has-spouse": "a un conjoint", "has-type": "a le type", "headquarters": "siège social", "is-member-of": "est membre de", "is-where": "situé à", "loc-leader": "guide d'emplacement", "movie-has-director": "le film a un réalisateur", "no_relation": "aucune relation", "org-has-founder": "l'organisation a un fondateur", "org-has-member": "l'organisation a un membre", "org-leader": "chef d'organisation", "won-award": "a remporté le prix";

- **IT** "birth-place": "luogo di nascita", "event-year": "anno dell'evento", "from-country": "dal paese", "has-author": "ha autore", "has-child": "ha un figlio", "has-edu": "ha un'educazione", "has-genre": "ha genere", "has-occupation": "ha occupazione", "has-parent": "ha un genitore", "has-population": "ha una popolazione", "has-spouse": "ha un coniuge", "has-type": "ha il tipo", "headquarters": "sede centrale", "is-member-of": "è membro di", "is-where": "situato in", "loc-leader": "leader della posizione", "movie-has-director": "il film ha direttore", "no_relation": "nessuna relazione", "org-has-founder": "l'organizzazione ha fondatore", "org-has-member": "l'organizzazione ha un membro", "org-leader": "leader dell'organizzazione", "won-award": "ha vinto un premio";

- **KO** "birth-place": "출생지", "event-year": "이벤트 연도", "first-product": " 첫 번째 제품", "from-country": "나라에서", "has-author": "저자가 있다", "has-child": "아이가 있다", "has-edu": "교육이 있다", "has-genre": "장르가 있다", "has-highest-mountain": "가장 높은 산이 있다", "has-nationality": "국적이 있다", "has-occupation": "직업이 있다", "has-parent": "부모가 있다", "has-population": "인구가 있다", "has-sibling": "형제가 있다", "has-spouse": "배우자가 있다", "has-tourist-attraction": "관광명소가 있다", "has-type": "유형이 있습니다", "headquarters": "본부", "invented-by": "에 의해 발명", "invented-when": "언제 발명", "is-member-of": "의 회원입니다", "is-where": "어디에", "movie-has-director": "영화에 감독이 있다", "no_relation": "관계가 없다", "org-has-founder": "조직에는 설립자가 있습니다", "org-has-member": "조직에 구성원이 있습니다", "org-leader": "조직 리더", "won-award": "수상";

- **NL** "birth-place": "geboorteplaats", "event-year": "evenementenjaar", "from-country": "van het land", "has-author": "heeft auteur", "has-child": "heeft kind", "has-edu": "heeft onderwijs", "has-genre": "heeft genre", "has-occupation": "heeft beroep", "has-parent": "heeft ouder", "has-population": "heeft bevolking", "has-spouse": "heeft echtgenoot", "has-type": "heeft type", "headquarters": "hoofdkantoor", "is-member-of": "is lid van", "is-where": "gevestigd in", "loc-leader": "locatieleider", "movie-has-director": "film had regisseur", "no_relation": "geen relatie", "org-has-founder": "organisatie heeft oprichter", "org-has-member": "organisatie heeft lid", "org-leader": "organisatieleider", "won-award": "won prijs";

- **PL** "birth-place": "miejsce urodzenia", "event-year": "rok imprezy", "from-country": "z kraju", "has-author": "ma autor", "has-child": "ma dziecko", "has-edu": "ma wykształcenie", "has-genre": "ma gatunek", "has-occupation": "ma zawód", "has-parent": "ma rodzica", "has-population": "ma ludność", "has-spouse": "ma współmałżonka", "has-type": "ma typ", "headquarters": "siedziba główna", "is-member-of": "jest członkiem", "is-where": "mieszczący się w", "loc-leader": "lider lokalizacji", "movie-has-director": "film ma reżysera", "org-has-founder": "organizacja ma założyciela", "org-has-member": "organizacja ma członków", "org-leader": "lider organizacji", "won-award": "otrzymał nagrodę";

- **PT** "birth-place": "local de nascimento", "event-year": "ano do evento", "from-country": "do país", "has-author": "tem autor", "has-child": "tem filho", "has-edu": "tem educação", "has-genre": "tem género", "has-occupation": "tem ocupação", "has-parent": "tem pai", "has-population": "tem população", "has-spouse": "tem cônjuge", "has-type": "tem tipo", "headquarters": "sede", "is-member-of": "é membro de", "is-where": "localizado em", "loc-leader": "loc leader", "movie-has-director": "filme tem realizador", "no_relation": "sem relação", "org-has-founder": "organização tem fundador", "org-has-member": "organização tem membro", "org-leader": "líder da organização", "won-award": "ganhou prémio";

- **RU** "event-year": "год события", "has-edu": "имеет образование", "has-genre":

"имеет жанр", "has-occupation": "имеет профессию", "has-population": "имеет население", "has-type": "имеет тип", "is-member-of": "является членом", "no_relation": "без связи";

- **SV** "birth-place": "födelseort", "event-year": "År för evenemanget", "from-country": "från ett land", "has-author": "har en författare", "has-child": "har chili", "has-edu": "har utbildning", "has-genre": "har en genre", "has-occupation": "har ockuperat", "has-parent": "har en förälder", "has-population": "har en befolkning", "has-spouse": "har make eller maka", "has-type": "har typ", "headquarters": "huvudkontor", "is-member-of": "är medlem i", "is-where": "som ligger i", "loc-leader": "platsansvarig", "movie-has-director": "filmen har regissör", "no_relation": "ingen relation", "org-has-founder": "organisationen har en grundare", "org-has-member": "organisationen har en medlem", "org-leader": "ledare för organisationen", "won-award": "vann ett pris";
- **UK** "event-year": "рік події", "has-edu": "має освіту", "has-genre": "має жанр", "has-occupation": "має заняття", "has-population": "має населення", "has-type": "має тип", "no_relation": "ніякого відношення".

C Detailed Few-shot Results

Shots	Method	AR	DE	EN	ES	FA	FR	IT	KO	NL	PL	PT	RU	SV	UK	$\overline{X}$
8	XLM-R$_{\text{EM}}$	**58.8**±20.2	**49.2**±7.2	31.8±11.3	12.8±6.4	7.3±4.6	30.6±4.0	**52.3**±5.0	16.5±7.9	38.1±10.3	**46.6**±10.0	**53.8**±3.8	**60.7**±7.2	1.3±0.8	12.0±7.0	33.7
	null prompts	17.2±10.6	28.1±16.6	37.4±10.1	10.4±7.9	25.8±10.1	14.6±10.4	28.0±22.3	**36.9**±14.1	44.4±8.5	29.7±15.2	26.1±17.0	39.2±12.3	47.4±14.5	27.4±15.6	29.5
	CS	19.6±10.2	11.1±17.2	42.2±17.5	26.2±21.5	45.0±12.0	36.3±17.4	42.3±4.9	18.7±17.4	28.0±15.6	26.7±16.6	27.0±14.1	47.5±15.4	**48.9**±16.6	28.0±17.5	32.0
	SP	14.2±5.5	29.1±18.6	**45.4**±7.9	32.5±12.1	18.8±12.0	20.1±11.0	26.8±19.1	20.6±19.4	31.8±14.7	26.3±16.2	26.0±15.2	29.6±16.6	36.4±25.3	30.8±28.8	27.4
	IL	33.4±25.4	39.0±19.3	42.2±17.5	**37.9**±15.3	**46.0**±28.5	**39.1**±15.9	35.1±19.0	35.5±19.9	**52.5**±4.2	31.8±12,9	32.8±19.3	55.9±14.7	34.1±26.9	**52.8**±14.2	**40.6**
16	XLM-R$_{\text{EM}}$	**67.7**±17.5	44.3±23.1	**56.4**±4.2	19.6±6.5	7.8±9.4	47.5±8.4	**76.1**±4.3	26.7±5.0	64.7±2.8	**62.8**±5.6	**69.1**±2.8	**70.9**±8.6	1.3±0.4	19.5±12.2	45.3
	null prompts	34.5±18.4	18.1±20.4	42.1±15.5	20.5±12.0	43.2±14.9	28.7±22.0	38.0±18.9	25.3±19.0	37.5±14.3	37.8±8.8	17.8±16.1	54.3±20.3	**56.6**±23.5	42.8±20.6	35.5
	CS	36.6±18.1	**62.5**±11.0	50.5±32.3	26.1±21.5	**49.7**±11.1	47.3±30.3	53.5±27.7	**39.3**±9.4	**71.2**±9.0	33.5±25.0	45.3±19.1	61.2±26.1	49.4±24.7	**58.4**±22.4	**48.9**
	SP	38.6±17.6	40.2±29.4	53.7±25.2	**52.0**±13.8	37.9±14.0	**51.3**±27.3	46.6±24.3	38.6±17.6	40.2±29.4	53.7±25.2	52.0±13.8	37.9±14.0	51.3±27.3	46.6±24.3	45.8
	IL	47.0±32.3	62.5±11.0	50.5±32,3	31.1±22.2	45.6±23.1	21.7±17.5	32.8±18.1	33.7±20.2	39.2±11.4	58.5±18.9	50.2±19.4	65.4±6.5	51.1±22.9	58.2±20.9	46.3
32	XLM-R$_{\text{EM}}$	**81.6**±9.4	59.9±29.8	73.2±4.4	21.4±3.1	12.7±6.3	58.8±10.1	**81.0**±2.6	38.8±3.3	**74.5**±2.8	**77.7**±1.6	63.2±26.2	62.5±12.8	1.3±1.3	11.7±5.9	51.3
	null prompts	45.4±20.0	26.0±24.3	56.0±13.4	14.3±15.1	67.4±6.3	48.6±16.8	42.8±21.0	30.2±29.8	54.9±25.6	40.7±21.7	15.1±16.9	48.4±33.7	49.7±30.7	**58.1**±27.4	42.7
	CS	62.0±26.7	**72.1**±15.0	**80.9**±4.3	40.6±30.3	61.0±28.2	51.4±22.1	50.4±37.9	**72.2**±6.9	71.4±23.5	39.0±30.0	**73.3**±8.0	57.7±20.3	67.6±12.0	51.3±23.7	60.8
	SP	50.3±19.7	35.5±33.8	61.2±29.3	**60.8**±26.7	59.0±27.1	**74.2**±12.5	34.7±36.2	29.6±34.7	42.7±33.5	67.4±12.3	47.4±28.0	**65.1**±19.0	**69.2**±20.1	57.0±32.1	53.9
	IL	65.5±22.0	61.9±29.5	80.9±4.3	53.1±28.5	**76.4**±9.9	62.0±29.1	71.7±26.6	50.8±24.9	71.3±12.5	65.2±25.8	59.5±28.5	63.8±27.4	63.6±26.1	71.1±17.4	**65.5**

Table 7: Few-shot results in micro-F1 (%) on the SMiLER dataset. We evaluate XLM-R$_{\text{EM}}$, null prompts, and our prompt variants. For each result, the mean and standard deviation of 5 runs are reported. $\overline{X}$: macro average across 14 languages. The standard deviations are quite large which indicates that multiple runs are needed and results are seed dependent. In-language prompting provides the most consistent results, with Polish 8-shot as lowest score (31.8 F_1). Other methods all have results below 15.0 F_1.