X-METRA-ADA: Cross-lingual Meta-Transfer Learning Adaptation to Natural Language Understanding and Question Answering

Meryem M'hamdi[1]*, **Doo Soon Kim**[2], **Franck Dernoncourt**[2],
Trung Bui[2], **Xiang Ren**[1] and **Jonathan May**[1]
[1]Information Sciences Institute, University of Southern California
{meryem, xiangren, jonmay}@isi.edu
[2]Adobe Research
{dkim, franck.dernoncourt, bui}@adobe.com

Abstract

Multilingual models, such as M-BERT and XLM-R, have gained increasing popularity, due to their zero-shot cross-lingual transfer learning capabilities. However, their generalization ability is still inconsistent for typologically diverse languages and across different benchmarks. Recently, meta-learning has garnered attention as a promising technique for enhancing transfer learning under low-resource scenarios: particularly for cross-lingual transfer in Natural Language Understanding (NLU).

In this work, we propose **X-METRA-ADA**, a **cross**-lingual **ME**ta-**TRA**nsfer learning **ADA**ptation approach for NLU. Our approach adapts MAML, an optimization-based meta-learning approach, to learn to adapt to new languages. We extensively evaluate our framework on two challenging cross-lingual NLU tasks: multilingual task-oriented dialog and typologically diverse question answering. We show that our approach outperforms naive fine-tuning, reaching competitive performance on both tasks for most languages. Our analysis reveals that X-METRA-ADA can leverage limited data for faster adaptation.

1 Introduction

Cross-lingual transfer learning is a technique used to adapt a model trained on a downstream task in a source language to directly generalize to the task in new languages. It aims to come up with common cross-lingual representations and leverages them to bridge the divide between resources to make any NLP application scale to multiple languages. This is particularly useful for data-scarce scenarios, as it reduces the need for API calls implied by machine translation or costly task-specific annotation for new languages.

*Work was started while the first author was a research intern at Adobe.

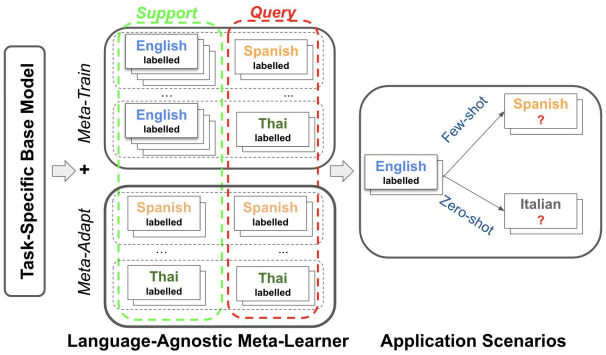


Figure 1: An overview of the X-METRA-ADA framework: we use English as the source and Spanish as the target language. The meta-train stage transfers from the source to the target languages, while the meta-adaptation further adapts the model to the target language. The application is **few-shot** if the test language is seen in any stage of X-METRA-ADA; or **zero-shot** if the test language is unseen.

Transformer-based contextualized embeddings and their multilingual counterparts such as M-BERT (Devlin et al., 2019) have become popular as off-the-shelf representations for cross-lingual transfer learning. While these multilingual representations exhibit some cross-lingual capability even for languages with low lexical overlap with English, the transfer quality is reduced for languages that exhibit different typological characteristics (Pires et al., 2019).

The generalization of such representations has been extensively evaluated on traditional tasks such as Part-of-Speech (POS) tagging, Named Entity Recognition (NER) and Cross-lingual Document Classification (CLDC) (Ahmad et al., 2019; Wu and Dredze, 2019; Bari et al., 2020a; Schwenk and Li, 2018), with ever-growing open community annotation efforts like Universal Dependencies (Nivre et al., 2020) and CoNLL shared tasks (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003). On the other hand, cross-lingual Natural Language Understanding (NLU)

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3617–3632
June 6–11, 2021.

tasks have gained less attention, with smaller benchmark datasets that cover a handful of languages and don't truly model linguistic variety (Conneau et al., 2018; Artetxe et al., 2020). Natural Language Understanding tasks are critical for dialog systems, as they make up an integral part of the dialog pipeline. Understanding and improving the mechanism behind cross-lingual transfer for natural language understanding in dialog systems require evaluations on more challenging and typologically diverse benchmarks.

Numerous approaches have attempted to build stronger cross-lingual representations on top of those multilingual models; however, most require parallel corpora (Wang et al., 2019; Lample and Conneau, 2019) and are biased towards high-resource and balanced setups. This fuels the need for a method that doesn't require explicit cross-lingual alignment for faster adaptation to low-resource setups.

Meta-learning, a method for "learning to learn", has found favor especially among the computer vision and speech recognition communities (Nichol et al., 2018; Triantafillou et al., 2020; Winata et al., 2020). Meta-learning has been used for machine translation (Gu et al., 2018), few-shot relation classification (Gao et al., 2019), and on a variety of GLUE tasks (Dou et al., 2019). Recently, Nooralahzadeh et al. (2020) apply the MAML (Finn et al., 2017) algorithm to cross-lingual transfer learning for XNLI (Conneau et al., 2018) and MLQA (Lewis et al., 2020), NLU tasks that are naturally biased towards machine translation-based solutions. Nooralahzadeh et al. are able to show improvement over strong multilingual models, including M-BERT. However, they mainly show the effects of meta-learning as a first step in a framework that relies on supervised fine-tuning, making it difficult to properly compare and contrast both approaches.

We study cross-lingual meta-transfer learning from a different perspective. We distinguish between meta-learning and fine-tuning and design systematic experiments to analyze the added value of meta-learning compared to naive fine-tuning. We also build our analysis in terms of more typologically diverse cross-lingual NLU tasks: Multilingual Task-Oriented Dialogue System (MTOD) (Schuster et al., 2019) and Typologically Diverse Question Answering (TyDiQA) (Clark et al., 2020). While XNLI is a classification task, MTOD is a joint classification and sequence labelling task and is more typologically diverse. TyDiQA is not a classification task, but we show how meta-learning can be applied usefully to it. We also show greater performance improvements from meta-learning than fine-tuning on transfer between typologically diverse languages.

To the best of our knowledge, we are the first to conduct an extensive analysis applied to MTOD and TyDiQA to evaluate the quality of cross-lingual meta-transfer. Our contributions are three-fold:

- Proposing X-METRA-ADA,[1] a language-agnostic meta-learning framework (Figure 1), and extensively evaluating it.
- Applying X-METRA-ADA to two challenging cross-lingual and typologically diverse task-oriented dialog and QA tasks, which includes recipes for constructing appropriate meta-tasks (Section 2.3).
- Analyzing the importance of different components in cross-lingual transfer and the scalability of our approach across different k-shot and down-sampling configurations (Section 4.2).

2 Methodology

We make use of optimization-based meta-learning on top of pre-trained models with two levels of adaptation to reduce the risk of over-fitting to the target language: (i) **meta-training** from the source language to the target language(s) (ii) **meta-adaptation** on the same target language(s) for more language-specific adaptation (Figure 1).

We apply our approach to two cross-lingual downstream tasks: MTOD (Section 2.1) and TyDiQA (Section 2.2). We start by describing the base architectures for both tasks, before explaining how they are incorporated into our meta-learning pipeline. Applying meta-learning to a task requires the construction of multiple 'pseudo-tasks', which are instantiated as pairs of datasets. We describe this construction for our downstream tasks in Section 2.3. Finally, we present our X-METRA-ADA algorithm (Section 2.4).

2.1 Multilingual Task-Oriented Dialog (MTOD)

Similar to the architecture in Castellucci et al. (2019), we model MTOD's intent classification and slot filling subtasks jointly. For that purpose, we

[1] We release our code at: github.com/meryemmhamdi1/meta_cross_nlu_qa.

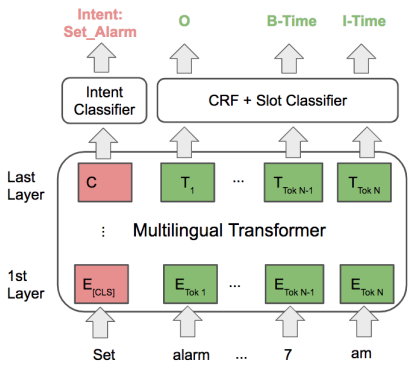


Figure 2: Architecture of Base MTOD.

use a joint text classification and sequence labeling framework with feature representation based on Transformer (Vaswani et al., 2017). More specifically, given a multilingual pre-trained model, we use it to initialize the word-piece embeddings layer. Then, we add on top of it a text classifier to predict the intent from the $[CLS]$ token representation and a sequence labeling layer in the form of a linear layer to predict the slot spans (in BIO annotation), as shown in Figure 2. We optimize parameters using the sum of both intent and CRF based slot losses.

2.2 Typologically Diverse Question Answering (TyDiQA)

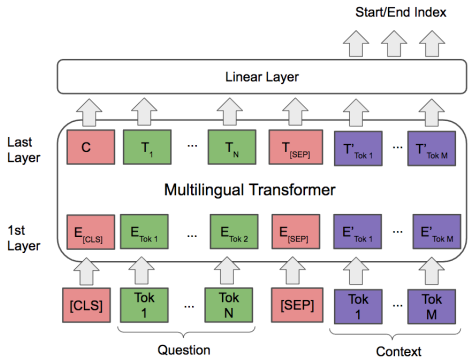


Figure 3: Question Answering Base Model.

Inspired by Hu et al. (2020), we apply to TyDiQA the same architecture as the original BERT fine-tuning procedure for question answering on SQuAD (Devlin et al., 2019). Specifically, the input question (after prepending it with a $[CLS]$ token) and the context are concatenated as a single packed sequence separated by a $[SEP]$ token. Then, the embeddings of the context are fed to a linear layer plus a softmax to compute the probability that each token is the START or END of the answer. The whole architecture is fine-tuned by optimizing for the joint loss over the START and END predictions. Any START and END positions that are outside of the scope of the context end up being truncated because of Transformer-based embeddings length limitations and are ignored during training. Figure 3 illustrates the architecture.

2.3 Psuedo-task Datasets

Meta-learning is distinguished from fine-tuning in that the former seeks an initialization point that is maximally useful to multiple downstream learning tasks, while the latter seeks to directly optimize a downstream 'child' task from the initialization point of a 'parent' task. To apply meta-learning to data scenarios that more closely fit fine-tuning, we construct multiple 'pseudo-tasks' by subsampling from parent and child task datasets. A pseudo-task is defined as a tuple $T = (S, Q)$, where each of S and Q are labeled samples. In the inner loops of meta-learning, the loss on Q from a model trained on S is used to adapt the initialization point (where Q and S are referred to as the *query* and *support* in meta-learning literature). Pseudo-tasks are constructed in such a way as to make them balanced and non-overlapping. We describe our approach for each task below.

2.3.1 MTOD Pseudo-task Construction

MTOD labeled data consists of a sentence from a dialogue along with a sentence-level intent label and subsequence slot labels. From the available data, we draw a number of task sets $\mathcal{T}$; each $T = (S, Q) \in \mathcal{T}$ consists of k intent and slot-labeled items per intent class in S and q items per class in Q. Although carefully arranged to have the same number of items per class per task in each of the support and the query sets, the same task splits are used for slot prediction as well. During meta-training and meta-adaptation, task batches are sampled randomly from $\mathcal{T}$.

2.3.2 QA Pseudo-task Construction

Unlike MTOD, QA is not a standard classification task with fixed classes; thus, it is not directly amenable to class distribution balancing across pseudo-task query and support sets. To construct pseudo-tasks for QA from the available (question, context, answer) span triplet data, we use the following procedure: We draw a task $T = (S, Q)$, by first randomly drawing q triplets, forming Q. For each triplet t in Q, we draw the k/q *most similar* triplets to t from the remaining available data,

thus forming S.[2] For two triplets t_1, t_2 we define similarity as $\cos(f(t_1), f(t_2))$, where $f(.)$ is a representation of the concatenation of the triplet elements delimited by a space; we use a cross-lingual extension to SBERT's pre-trained model (Reimers and Gurevych, 2019, 2020).

2.3.3 Cross-lingual extension

In the original MAML (Finn et al., 2017), in every iteration we sample a task set $\mathcal{T}$ from a single distribution $\mathcal{D}$, and the support and query sets in a single task T would be drawn from a common space. We distinguish between the distributions $\mathcal{D}_{\text{meta-train}}$ and $\mathcal{D}_{\text{meta-adapt}}$, which correspond to the two levels of adaptation introduced in Section 2 and explained below in Section 2.4.

To enable cross-lingual transfer, we draw data for the support set of tasks in $\mathcal{D}_{\text{meta-train}}$ from task data in the high-resource base language (English, in our experiments). For the query set in $\mathcal{D}_{\text{meta-train}}$ and for both support and query sets in $\mathcal{D}_{\text{meta-adapt}}$, we sample from task data in the language to be evaluated.

2.4 X-METRA-ADA Algorithm

Following the notation described in the above sections, we present our algorithm X-METRA-ADA, our adaptation of MAML to cross-lingual transfer learning in two stages. In each stage we use the procedure outlined in Algorithm 1. We start by sampling a batch of tasks from distribution $\mathcal{D}$. For every task $T_j = (S_j, Q_j)$, we update θ_j over n steps using batches drawn from S_j. At the end of this inner loop, we compute the gradients with respect to the loss of θ_j on Q_j. At the end of all tasks of each batch, we sum over all pre-computed gradients and update θ, thus completing one outer loop. The difference between meta-train and meta-adapt stages comes down to the parameters and hyperparameters passed into Algorithm 1.

- **Meta-train**: This stage is similar to classical MAML. Task sets are sampled from $\mathcal{D}_{\text{meta-train}}$, which uses high-resource (typically English) data in support sets and low-resource data in the query sets. The input model θ_B is typically a pre-trained multilingual downstream base model, and we use hyperparameters $n = 5, \alpha = 1\text{e}{-3}$ and $\beta = 1\text{e}{-2}$ for MTOD and $\alpha = \beta = 3\text{e}{-5}$ for QA.

[2]Thus k is constrained to be a multiple of q.

Algorithm 1 X-METRA-ADA

Require: Task set distribution $\mathcal{D}$, pre-trained learner B with parameters θ_B, meta-learner M with parameters (θ, α, β, n)

Initialize $\theta \leftarrow \theta_B$
while not done **do**
  Sample batch of tasks $\mathcal{T} = \{T_1, T_2, \dots T_b\} \sim \mathcal{D}$
  for all $T_j = (S_j, Q_j)$ in $\mathcal{T}$ **do**
    Initialize $\theta_j \leftarrow \theta$
    for $t = 1 \dots n$ **do**
      Evaluate $\partial B_{\theta_j} / \partial \theta_j = \nabla_{\theta_j} \mathcal{L}_{T_j}^{S_j}(B_{\theta_j})$
      Update $\theta_j = \theta_j - \alpha \partial B_{\theta_j} / \partial \theta_j$
    end for
    Evaluate query loss $\mathcal{L}_{T_j}^{Q_j}(B_{\theta_j})$ and save it for outer loop
  end for
  Update $\theta \leftarrow \theta - \beta \nabla_\theta \sum_{j=1}^{b} \mathcal{L}_{T_j}^{Q_j}(B_{\theta_j})$
end while

- **Meta-adapt**: During this stage, we ensure the model knows how to learn from examples within the target language under a low-resource regime. Task sets are sampled from $\mathcal{D}_{\text{meta-adapt}}$, which uses low-resource data in both support and query sets. The input model is the optimization resulting from meta-train, and we use hyperparameters $n = 5, \alpha = 1\text{e}{-3}$ and $\beta = 1\text{e}{-2}$ for MTOD and $\alpha = \beta = 3\text{e}{-5}$ for QA.

3 Experimental Setup

3.1 Datasets

For dialogue intent prediction, we use the Multilingual Task-Oriented Dialogue (MTOD) (Schuster et al., 2019) dataset. MTOD covers 3 languages (English, Spanish, and Thai), 3 domains (alarm, reminder, and weather), 12 intent types, and 11 slot types.[3] We train models with the English training data ($Train$) but for the other languages we use the provided development sets (Dev) to further our goals to analyze methods of few-shot transfer. We evaluate on the provided test sets. Moreover, we evaluate on an in-house dataset of 7 languages.[4]

For QA, we use the Typologically Diverse QA (TyDiQA-GoldP) (Clark et al., 2020) dataset. TyDiQA is a typologically diverse question answering dataset covering 11 languages. Like Hu et al. (2020), we use a simplified version of the primary task. Specifically, we discard questions that don't have an answer and use only the gold passage as context, keeping only the short answer and its spans. This makes the task similar to XQuAD and MLQA,

[3]We follow the same pre-processing and evaluation as Liu et al. (2020).

[4]More details are included in the Appendix C.

although unlike these tasks, the questions are written without looking at the answers and without machine translation. As with MTOD, we use the English training data as $Train$. Since development sets are not specified for MTOD, we instead reserve 10% of the training data in each of the other languages as Dev. We report on the provided test sets. Statistics of datasets for both tasks can be found in Appendix A.

3.2 Evaluation

In order to fairly and consistently evaluate our approach to few-shot transfer learning via meta-learning and to ablate components of the method, we design a series of experiments based on both internal and external baselines. Our internal baselines ablate the effect of the X-METRA-ADA algorithm vs. conventional fine-tuning from a model trained on a high-resource language by keeping the data sets used for training constant. As our specific data conditions are not reproduced in any externally reported results on these tasks, we instead compare to other reported results using English-only or entirely zero-shot training data.

Internal Evaluation We design the following fine-tuning/few-shot schemes:

- *PRE*: An initial model is fine-tuned on the $Train$ split of English only and then evaluated on new languages with no further tuning or adaptation. This strawman baseline has exposure to English task data only.
- *MONO*: An initial model is fine-tuned on the Dev split of the target language. This baseline serves as a comparison for standard fine-tuning (FT, below), which shows the value of combining MONO and PRE.
- *FT*: We fine-tune the PRE model on the Dev split of the target language. This is a standard transfer learning approach that combines PRE and MONO.
- *FT w/EN*: Like FT, except both the Dev split of the target language and the $Train$ split of English are used for fine-tuning. This is used for dataset equivalence with X-METRA-ADA (below).
- *X-METRA*: We use the PRE model as θ_B for meta-train, the $Train$ split from English to form support sets in $D_{meta-train}$, and all of the Dev split of the target language to form query sets in $D_{meta-train}$.
- *X-METRA-ADA*: We use the PRE model as θ_B for meta-train, the $Train$ split from English to form support sets in $\mathcal{D}_{\text{meta-train}}$. For MTOD, we use 75% of the Dev split of the target language to form query sets in $\mathcal{D}_{\text{meta-train}}$. We use the remaining 25% of the Dev split of the target language for both the support and query sets of $\mathcal{D}_{\text{meta-adapt}}$. For QA, we use ratios of 60% for $\mathcal{D}_{\text{meta-train}}$ and 40% for $\mathcal{D}_{\text{meta-adapt}}$.

All models are ultimately fine-tuned versions of BERT and all have access to the same task training data relevant for their variant. That is, X-METRA-ADA and PRE both see the same English $Train$ data and MONO, FT, and X-METRA-ADA see the same target language Dev data. However, since X-METRA-ADA uses both $Train$ and Dev to improve upon PRE, and FT only uses Dev, we make an apples-to-apples comparison, data-wise, by including FT w/EN experiments as well.

External Baselines We focus mainly on transfer learning baselines from contextualized embeddings for a coherent external comparison; supervised experiments on target language data such as those reported in Schuster et al. (2019) are inappropriate for comparison because they use much more in-language labeled data to train. The experiments we compare to are zero-shot in the sense that they are not trained directly on the language-specific task data. However, most of these external baselines involve some strong cross-lingual supervision either through cross-lingual alignment or mixed-language training. We also include machine translation baselines, which are often competitive and hard to beat. Our work, by contrast, uses no parallel language data or resources beyond pretrained multilingual language models, labeled English data, and few-shot labeled target language data. To the best of our knowledge, we are the first to explore cross-lingual meta-transfer learning for those benchmarks, so we only report on our X-METRA-ADA approach in addition to those baselines.

For MTOD, then, we focus on the following external baselines:

- *Cross-lingual alignment-based approaches:* We use MCoVe, a multilingual version of contextualized word vectors with an autoencoder objective as reported by Schuster et al. (2019) in addition to M-BERT (Liu et al., 2020). We also include XLM trained on Translation Language Modeling (TLM) + Masked Language Modeling (MLM)

(Lample and Conneau, 2019) as enhanced by Transformer and mixed-training as reported by Liu et al. (2020).

- *Mixed-language training approaches*: We use M-BERT + Transformer + mixed training using data from the dialogue domain: from (a) human-based word selection (MLT_H) and (b) attention-based word selection (MLT_A), both are reported by Liu et al. (2020).
- *Translation-based approaches*: We use the zero-shot version of MMTE, the massively multilingual translation encoder by Siddhant et al. (2020) fine-tuned on intent classification. We also include Translate Train (TTrain) (Schuster et al., 2019), which translates English training data into target languages to train on them in addition to the target language training data.

For TyDiQA-GoldP, out of the already mentioned baselines, we use M-BERT, XLM, MMTE, and TTrain (which unlike (Schuster et al., 2019) only translates English to the target language to train on it without data augmentation). In addition to that we also include XLM-R as reported by Hu et al. (2020).

3.3 Implementation Details

We use M-BERT (bert-base-multilingual-cased)[5] with 12 layers as initial models for MTOD and TyDiQA-GoldP in our internal evaluation. We use xlm-r-distilroberta-base-paraphrase-v1[6] model for computing similarities when constructing the QA meta-dataset (Section 2.3.2).

Our implementation of X-METRA-ADA from scratch uses learn2learn (Arnold et al., 2020) for differentiation and update rules in the inner loop.[7] We use the first-order approximation option in learn2learn for updating the outer loop, also introduced in Finn et al. (2017). For each model, we run for 3 to 4 different random initializations (for some experiments like PRE for TyDiQA-GoldP we use only 2 seeds respectively) and report the average and standard deviation of the best model for the few-shot language for each run. We use training loss convergence as a criteria for stopping. For the FT and MONO baselines, we don't have the luxury of *Dev* performance, since those baselines use the *Dev* dataset for training.[8] The *Dev* set is chosen to simulate a low-resource setup. More details on the hyperparameters used can be found in Appendix B.

4 Results and Discussion

4.1 Zero-shot and Few-shot Cross-Lingual NLU and QA

Model	Spanish		Thai	
	Intent Acc	Slot F1	Intent Acc	Slot F1
External Baselines				
MCoVe†	53.9	19.3	70.7	35.6
M-BERT‡	73.7	51.7	28.1	10.6
MLT_H‡	82.9	**74.9**	53.8	26.1
MLT_A‡	87.9	73.9	73.5	27.1
XLM‡	87.5	68.5	72.6	27.9
MMTE+	**93.6**	-	89.6	-
TTrain‡	85.4	72.9	**95.9**	55.4
Zero-shot Learning				
PRE	70.2	38.2	45.4	12.5
Few-shot Learning				
MONO	82.4 ±6.0	43.9 ±1.5	79.1 ±4.7	54.1 ±3.9
FT	90.7 ±0.3	***67.6*** ±1.3	78.9 ±0.2	66.0 ±2.1
FT w/EN	88.7 ±0.4	67.4 ±1.4	73.7 ±0.1	66.0 ±1.6
X-METRA	89.6 ±1.3	63.6 ±0.5	80.2 ±1.2	**70.4** ±1.2
X-METRA-ADA	***92.9*** ±0.6	60.9 ±1.9	***86.3*** ±1.7	69.6 ±1.9

Table 1: Performance evaluation on MTOD between meta-learning approaches, fine-tuning internal baselines and external baselines. All our internal experiments use $k = q = 6$. Zero-shot learning experiments that train only on English are distinguished from few-shot learning, which include a fair internal comparison. Models in bold indicate our own internal models. **MONO**, **FT**, **FT w/EN**, **X-METRA**, and **X-METRA-ADA** models include results for each test language when training on that language. **FT w/EN** trains jointly on English and only the target language. We highlight the best scores in bold and underline the second best for each language and sub-task. The rest are reported from † (Schuster et al., 2019), ‡ (Liu et al., 2020), and + (Siddhant et al., 2020).

Table 1 shows the results for cross-lingual transfer learning on MTOD comparing different baselines.[9] In general, PRE model performs worse than other baselines. It performs less than the simplest baseline, MCoVe, when transferring to Thai with a decrease of 25.3% and 23.1% and an average cross-lingual relative loss of 4.5% and 2.1% for intent classification and slot filling respectively.

[5] github.com/huggingface/transformers version 3.4.0 pre-trained on 104 languages, including all languages evaluated on in this paper.

[6] github.com/UKPLab/sentence-transformers which uses XLM-R as the base model.

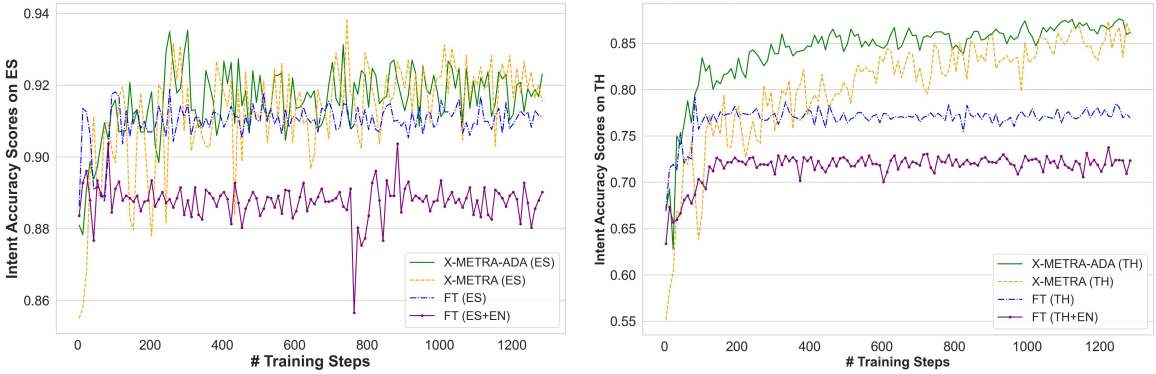
[7] github.com/learnables/learn2learn.

[8] All experiments are run using Pytorch version 1.6.0, 1 GeForce RTX P8 GPU of 11MB of memory CUDA version 10.1. The runtime depends on the size of the dev data but most MTOD models take around 3 hours to converge and TyDiQA models take a maximum of 10 hours training (including evaluation at checkpoints).

[9] More results on our in-house NLU dataset can be found in Appendix C.

Model	Test on Arabic	Bengali	Finnish	Indonesian	Russian	Swahili	Telugu
External Baselines							
M-BERT†	62.2	49.3	59.7	64.8	60.0	57.5	49.6
XLM†	59.4	27.2	58.2	62.5	49.2	39.4	15.5
XLM-R†	67.6	64.0	70.5	77.4	67.0	66.1	70.1
MMTE†	63.1	55.8	53.9	60.9	58.9	63.1	54.2
TTrain†	61.5	31.9	62.6	68.6	53.1	61.9	27.4
Zero-shot Learning							
PRE	62.4 ±2.2	32.9 ±1.4	57.7 ±4.4	67.8 ±3.8	58.2 ±3.7	55.5 ±2.9	33.0 ±5.9
Few-shot Learning							
MONO	74.0 ±1.1	38.9 ±0.8	63.3 ±1.5	67.1 ±1.9	54.4 ±1.3	60.3 ±1.2	61.4 ±1.0
FT	77.0 ±0.3	51.0 ±2.7	70.9 ±0.4	77.0 ±0.4	64.8 ±0.4	70.2 ±1.7	65.4 ±0.6
X-METRA	**78.5** ±0.6	53.2 ±0.5	72.7 ±0.4	**77.7** ±0.2	66.1 ±0.1	**71.7** ±0.2	66.6 ±0.4
X-METRA-ADA	76.6 ±0.1	**57.8** ±0.6	**73.0** ±0.3	77.3 ±0.1	**66.9** ±0.1	70.3 ±0.2	**72.8** ±0.1

Table 2: F1 comparison on TyDiQA-GoldP between different meta-learning approaches, fine tuning and external baselines. We highlight the best scores in bold and underline the second best for each language. Our own models are in bold, whereas the rest are reported from † (Hu et al., 2020). This is using $k = q = 6$.



(a) Intent Accuracy on Spanish

(b) Intent Accuracy on Thai

Figure 4: Ablation of the role of adaptation in X-METRA-ADA compared to X-METRA (X-METRA-ADA with the meta-training stage only). X-METRA-ADA converges faster than X-METRA which in turn is better than FT for both languages. More plots can be found in Appendix E.

This suggests that zero-shot fine-tuning M-BERT on English only is over-fitting on English and its similar languages. Using MLT_A which adds more dialogue-specific mixed training helps reduce that gap for Thai on intent accuracy mainly, but not with the same degree on slot filling.

The results confirm the positive effects of cross-lingual fine-tuning; although PRE is not a very effective cross-lingual learner, fine-tuning with in-language data on top of PRE (i.e. FT) adds value over the MONO baseline. Adding English data to fine-tuning (FT w/EN) is slightly harmful. However, the meta-learning approach appears to make the most effective use of this data in almost all cases (Spanish slot filling is an exception). We perform a pairwise two-sample t-test (assuming unequal variance) and find the results of X-METRA-ADA compared to FT on intent classification to be statistically significant with p-values of 1.5% and 2.4% for Spanish and Thai respectively, rejecting the null hypothesis with 95% confidence.

X-METRA-ADA outperforms all previous external baselines and fine-tuning models for both Spanish and Thai (except for slot filling on Spanish). We achieve the best overall performance with an average cross-lingual cross-task increase of 3.2% over the FT baseline, 6.9% over FT w/EN, and 12.6% over MONO. Among all models, MONO has the least stability as suggested by higher average standard deviation. There is a tendency for X-METRA-ADA to work better for languages like Thai compared to Spanish as Thai is a truly low-resource language. This suggests that pre-training on English only learns an unsuitable initialization, impeding its generalization to other languages. As expected, fine-tuning on small amounts of the Dev data does not help the model generalize to new languages. MONO baselines exhibit less stability than

X-METRA-ADA. On the other hand, X-METRA-ADA learns a more stable and successful adaptation to that language even on top of a model pre-trained on English with less over-fitting.

Table 2 shows a comparison of methods for TyDiQA-GoldP across seven language, evaluating using F1.[10] The benefits of fine-tuning and improvements from X-METRA-ADA observed in Table 1 are confirmed. We also compare X-METRA-ADA to X-METRA, which is equivalent to X-METRA-ADA without the meta-adaptation phase. On average, X-METRA increases by 10.8% and 1.5% over the best external and fine-tuning baseline respectively, whereas MONO results lag behind. X-METRA-ADA outperforms X-METRA on average and is especially helpful on languages like Bengali and Telugu. We compare X-METRA and X-METRA-ADA in more depth in Section 4.2. Meta-learning significantly and consistently outperforms fine-tuning.

In Appendix D, we report zero-shot results for QA and notice improvements using X-METRA-ADA over FT for some languages. However, we cannot claim that there is a direct correlation between the degree to which the language is low-resource and the gain in performance of X-METRA-ADA over fine-tuning. Other factors like similarities of grammatical and morphological structure, and shared vocabulary in addition to consistency of annotation may play a role in the observed cross-lingual benefits. Studying such correlations is beyond the scope of this paper.

4.2 More Analysis

Meta-Adaptation Role The learning curves in Figure 4 compare X-METRA-ADA, X-METRA (i.e. meta-training but no meta-adaptation), and fine-tuning, both with English and with target language data only, for both Spanish and Thai intent detection in MTOD. In general, including English data in with in-language fine-tuning data lags behind language-specific training for all models, languages, and sub-tasks. With the exception of slot filling on Spanish, there is a clear gap between naive fine-tuning and meta-learning, with a gain in the favor of X-METRA-ADA especially for Thai. Naive fine-tuning, X-METRA, and X-METRA-ADA all start from the same checkpoint fine-tuned on English. All model variants are sampled from the same data. For Spanish, continuing to use English in naive fine-tuning to Spanish reaches better performance than both variants of meta-learning for Slot filling on Spanish (see Appendix E). This could be due to the typological similarity of Spanish and English, which makes optimization fairly easy for naive fine-tuning compared to Thai, which is both typologically distant and low-resource.

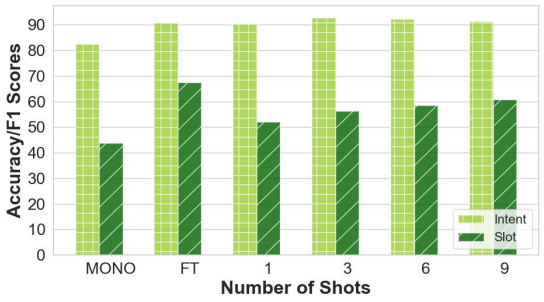


Figure 5: MTOD intent classification and slot filling on Spanish with different shots. The number of shots is the same for both support and query sets (i.e. $k = q$).

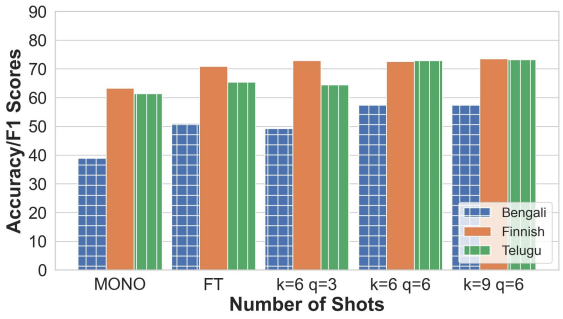


Figure 6: TyDiQA-GoldP F1 score analysis of different shots for both the support and query.

K-Shot Analysis We perform a k-shot analysis by treating the number of instances seen per class (i.e. 'shots') as a hyper-parameter to determine at which level few-shot meta-learning starts to outperform the fine-tuning and monolingual baselines. As shown in Figure 5, it seems that while even one shot for X-METRA-ADA is better than fine-tuning on intent classification, $k = q = 9$ shot and $k = q = 6$ shot are at the same level of stability with very slightly better results for 6 shot showing that more shots beyond this level will not improve the performance. While 1 shot performance is slightly below our monolingual baseline, it starts approaching the same level of performance as 3 shot upon convergence.

Figure 6 shows an analysis over both k and q shots for TyDiQA-GoldP. In general, increasing q

[10] Full results using Exact Match scores too can be found in Appendix D.

helps more than increasing k. The gap is bigger between $k = 6$ $q = 3$ and $k = 6$ $q = 6$ especially for languages like Bengali and Telugu. We can also see that $k = 6$ $q = 3$ is at the same level of performance to FT for those languages.

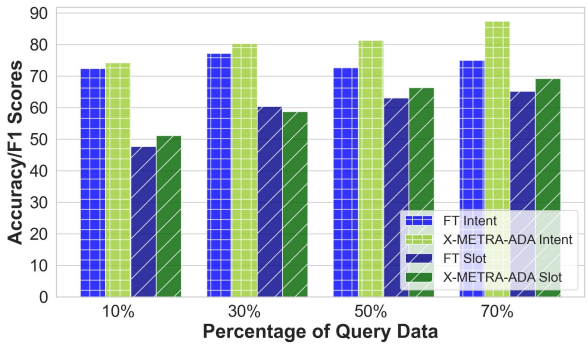


Figure 7: Downsampling analysis for Thai MTOD with different percentages of query data.

Downsampling Analysis We perform a downsampling analysis, where we gradually decrease the proportion of the overall set from which the target language is sampled used for few-shot learning in X-METRA-ADA and FT. Figure 7 shows a comparison between intent accuracies and slot F1 scores between the main models X-METRA-ADA and FT on Thai. We notice that as the percentage of query data increases, the gap between X-METRA-ADA and FT increases slightly, whereas the gain effect on slots is steadier. This suggests that X-METRA-ADA is at the same level of effectiveness even for lower percentages.

5 Related Work

Cross-lingual transfer learning Recent efforts apply cross-lingual transfer to downstream applications such as information retrieval (Jiang et al., 2020); information extraction (M'hamdi et al., 2019, Bari et al., 2020b), and chatbot applications (Lin et al., 2020, Abbet et al., 2018). Upadhyay et al. (2018) and Schuster et al. (2019) propose the first real attempts at cross-lingual task-oriented dialog using transfer learning. Although they show that cross-lingual joint training outperforms monolingual training, their zero-shot model lags behind machine translation for other languages.

To circumvent imperfect alignments in the cross-lingual representations, Liu et al. (2019) propose a latent variable model combined with cross-lingual refinement with a small bilingual dictionary related to the dialogue domain. Liu et al. (2020) enhance Transformer-based embeddings with mixed language training to learn inter-lingual semantics across languages. However, although these approaches show promising zero-shot performance for Spanish, their learned refined alignments are not good enough to surpass machine translation baselines on Thai.

More recently, Hu et al. (2020) and Liang et al. (2020) introduce XTREME and XGLUE benchmarks for the large-scale evaluation of cross-lingual capabilities of pre-trained models across a diverse set of understanding and generation tasks. In addition to M-BERT, they analyze models like XLM (Lample and Conneau, 2019) and Unicoder (Huang et al., 2019). Although the latter two models slightly outperform M-BERT, they need a large amount of parallel data to be pre-trained. It is also not clear the extent to which massive cross-lingual supervision helps to bridge the gap to linguistically distant languages.

Meta-learning for NLP Previous work in meta-learning for NLP is focused on the application of first-order MAML (Finn et al., 2017). Earlier work by Gu et al. (2018) extends MAML to improve low-resource languages for neural machine translation. Dou et al. (2019) apply MAML to NLU tasks in the GLUE benchmark. They show that meta-learning is a better alternative to multi-task learning, but they only validate their approach on English. Wu et al. (2020) also use MAML for cross-lingual NER with a slight enhancement to the loss function. More recently, Nooralahzadeh et al. (2020) also directly leverage MAML on top of M-BERT and XLM-R for zero-shot and few-shot XNLI and MLQA datasets. Although their attempt shows that cross-lingual transfer using MAML outperforms other baselines, the degree of typological commonalities among languages plays a significant role in that effect. In addition to that, their approach is an oversimplification of the n-way k-shot setup, with a one-fit-all sampling of data points for support and query and additional supervised fine-tuning.

6 Conclusion

In this paper, we adapt a meta-learning approach for cross-lingual transfer learning in Natural Language Understanding tasks. Our experiments cover two challenging cross-lingual benchmarks: task-oriented dialog and natural questions including an extensive set of low-resource and typologically diverse languages. X-METRA-ADA reaches better convergence stability on top of fine-tuning, reaching a new state of the art for most languages.

7 Acknowledgments

This work was started while the first author was a research intern at Adobe Research (Summer 2020). This material is partially based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via Contract No. 2019-19051600007. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein. We thank the anonymous reviewers for their detailed comments.

References

Christian Abbet, Meryem M'hamdi, Athanasios Giannakopoulos, Robert West, Andreea Hossmann, Michael Baeriswyl, and Claudiu Musat. 2018. Churn intent detection in multilingual chatbot conversations and social media. In *Proceedings of the 22nd Conference on Computational Natural Language Learning, CoNLL 2018, Brussels, Belgium, October 31 - November 1, 2018*, pages 161–170. Association for Computational Linguistics.

Wasi Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard Hovy, Kai-Wei Chang, and Nanyun Peng. 2019. On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2440–2452, Minneapolis, Minnesota. Association for Computational Linguistics.

Sébastien M. R. Arnold, Praateek Mahajan, Debajyoti Datta, Ian Bunner, and Konstantinos Saitas Zarkias. 2020. learn2learn: A library for meta-learning research.

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

M. Saiful Bari, Shafiq R. Joty, and Prathyusha Jwalapuram. 2020a. Zero-resource cross-lingual named entity recognition. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7415–7423. AAAI Press.

M. Saiful Bari, Shafiq R. Joty, and Prathyusha Jwalapuram. 2020b. Zero-resource cross-lingual named entity recognition. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7415–7423. AAAI Press.

Giuseppe Castellucci, Valentina Bellomaria, Andrea Favalli, and Raniero Romagnoli. 2019. Multilingual intent detection and slot filling in a joint bert-based model. *CoRR*, abs/1907.02884.

Jon Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Zi-Yi Dou, Keyi Yu, and Antonios Anastasopoulos. 2019. Investigating meta-learning algorithms for low-resource natural language understanding tasks. 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 1192–1197, Hong Kong, China. Association for Computational Linguistics.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135. PMLR.

Tianyu Gao, Xu Han, Hao Zhu, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2019. Fewrel 2.0:

Towards more challenging few-shot relation classification. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 6249–6254. Association for Computational Linguistics.

Jiatao Gu, Yong Wang, Yun Chen, Victor O. K. Li, and Kyunghyun Cho. 2018. Meta-learning for low-resource neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3622–3631, Brussels, Belgium. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. 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 2485–2494, Hong Kong, China. Association for Computational Linguistics.

Zhuolin Jiang, Amro El-Jaroudi, William Hartmann, Damianos G. Karakos, and Lingjun Zhao. 2020. Cross-lingual information retrieval with BERT. In *Proceedings of the workshop on Cross-Language Search and Summarization of Text and Speech, CLSSTS@LREC 2020, Marseille, France, May 2020*, pages 26–31. European Language Resources Association.

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

Patrick S. H. Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7315–7330. Association for Computational Linguistics.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. Xglue: A new benchmark dataset for cross-lingual pre-training, understanding and generation. *arXiv*, abs/2004.01401.

Zhaojiang Lin, Zihan Liu, Genta Indra Winata, Samuel Cahyawijaya, Andrea Madotto, Yejin Bang, Etsuko Ishii, and Pascale Fung. 2020. Xpersona: Evaluating multilingual personalized chatbot. *CoRR*, abs/2003.07568.

Zihan Liu, Jamin Shin, Yan Xu, Genta Indra Winata, Peng Xu, Andrea Madotto, and Pascale Fung. 2019. Zero-shot cross-lingual dialogue systems with transferable latent variables. 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 1297–1303, Hong Kong, China. Association for Computational Linguistics.

Zihan Liu, Genta Indra Winata, Zhaojiang Lin, Peng Xu, and Pascale Fung. 2020. Attention-informed mixed-language training for zero-shot cross-lingual task-oriented dialogue systems. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8433–8440. AAAI Press.

Meryem M'hamdi, Marjorie Freedman, and Jonathan May. 2019. Contextualized cross-lingual event trigger extraction with minimal resources. In *Proceedings of the 23rd Conference on Computational Natural Language Learning, CoNLL 2019, Hong Kong, China, November 3-4, 2019*, pages 656–665. Association for Computational Linguistics.

Alex Nichol, Joshua Achiam, and John Schulman. 2018. On first-order meta-learning algorithms. *CoRR*, abs/1803.02999.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Farhad Nooralahzadeh, Giannis Bekoulis, Johannes Bjerva, and Isabelle Augenstein. 2020. Zero-shot cross-lingual transfer with meta learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 4547–4562. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Asso-*

ciation for Computational Linguistics, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 4512–4525. Association for Computational Linguistics.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. 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 3795–3805, Minneapolis, Minnesota. Association for Computational Linguistics.

Holger Schwenk and Xian Li. 2018. A corpus for multilingual document classification in eight languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Aditya Siddhant, Melvin Johnson, Henry Tsai, Naveen Ari, Jason Riesa, Ankur Bapna, Orhan Firat, and Karthik Raman. 2020. Evaluating the cross-lingual effectiveness of massively multilingual neural machine translation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8854–8861. AAAI Press.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

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.

Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Utku Evci, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, and Hugo Larochelle. 2020. Meta-dataset: A dataset of datasets for learning to learn from few examples. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

S. Upadhyay, M. Faruqui, G. Tür, H. Dilek, and L. Heck. 2018. (almost) zero-shot cross-lingual spoken language understanding. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6034–6038.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Yuxuan Wang, Wanxiang Che, Jiang Guo, Yijia Liu, and Ting Liu. 2019. Cross-lingual BERT transformation for zero-shot dependency parsing. 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 5721–5727, Hong Kong, China. Association for Computational Linguistics.

Genta Indra Winata, Samuel Cahyawijaya, Zhaojiang Lin, Zihan Liu, Peng Xu, and Pascale Fung. 2020. Meta-transfer learning for code-switched speech recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 3770–3776. Association for Computational Linguistics.

Qianhui Wu, Zijia Lin, Guoxin Wang, Hui Chen, Börje F. Karlsson, Biqing Huang, and Chin-Yew Lin. 2020. Enhanced meta-learning for cross-lingual named entity recognition with minimal resources. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9274–9281. AAAI Press.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

A Dataset Statistics

Tables 3 and 4 show the statistics of MTOD and TyDiQA respectively per language and split.

Lang	ISO	Train	Dev	Test
English	EN	30,521	4,181	8,621
Spanish	ES	3,617	1,983	3,043
Thai	TH	2,156	1,235	1,692

Table 3: Statistics of MTOD dataset (Schuster et al., 2019) per language and split.

Lang	ISO	Train	Dev	Test
English	EN	3,326	370	440
Arabic	AR	13,324	1,481	921
Bengali	BN	2,151	239	113
Finnish	FI	6,169	686	782
Indonesian	ID	5,131	571	565
Russian	RU	5,841	649	812
Swahili	SW	2,479	276	499
Telugu	TE	5,006	557	669

Table 4: Statistics of TyDiQA-GoldP dataset per language and split. Korean is excluded due to some encoding issues.

B Hyperparameters

For MTOD, we fine-tune PRE on English training data. We use a batch size of 32, a dropout rate of 0.3, AdamW with a learning rate of $4e{-}5$, and ϵ of $1e{-}8$. We train for around 2000 steps. Beyond that point more training does not reveal necessary, so we perform early stopping at that point. For MONO, using a smaller learning rate of $4e{-}5$ helped achieve a good convergence for that model. For all FT experiments, we use the same learning rate of $1e{-}3$, which gave a better convergence.

For QA, we use a batch size of 4, doc stride of 128, a fixed maximum sequence length of 384, and a maximum length of questions of 30 words. We use AdamW optimizer throughout all experiments, which uses weight decay of $1e{-}3$, learning rate of $3e{-}5$, and a scheduler of 4 warm-up steps.[11] We fine-tune PRE for 2 epochs and observe no more gains in performance. For all MONO and FT experiments, we use the same learning rate of $3e{-}5$. This is the same optimizer and learning rate used for the outer loops in meta-learning as well.

For X-METRA-ADA and X-METRA, we sample 2500 tasks in total for both MTOD and QA. For each task, we randomly sample $k = q = 6$ examples from each intent class to form the support and query sets respectively (we consider all classes not only the intersection across languages). For QA, we use only one support example per query class and 6 query examples as classes. For the inner loop, we use learn2learn pre-built optimizer. For the outer loop, we use a standard Adam optimizer. In splitting the few-shot set, we use 75% for the meta-training and 25% for the meta-adaptation for MTOD. For QA, we use 60% of the query for meta-train and the remaining for meta-adaptation.

[11] Those hyperparameters are chosen based on Hu et al. (2020).

C Results on in-House Intent Classification Dataset

We perform an extensive evaluation including other languages for intent classification. We use an in-house dataset covering 6 target languages in addition to English. Statistics of train/dev/test splits are shown in Table 5. Table 6 shows a better performance in the favor of X-METRA with an average cross-lingual gain of 13.5% in accuracy over PRE. We notice that few-shot learning on the language of interest leads to the best performance, as indicated by higher numbers on the diagonal in the confusion matrix. Evaluation on more languages shows some complicity trends between languages from the same family. In addition to that, we notice that languages like Japanese and Korean help each other where few-shot on one helps zero-shot on the other by a margin of 15.6 and 5.6 on Korean and Japanese respectively.

Lang	Train	Dev	Test
English	5,438	1,814	1,814
German	1,570	526	526
French	1,082	362	362
Italian	1,082	362	362
Portuguese	1,150	386	386
Japanese	1,070	358	358
Korean	938	314	314

Table 5: Statistics of In-House multilingual intent classification Dataset.

D Full results for QA

Tables 7 and 8 show the full results for F1 and Exact Match (EM) metrics for QA respectively.

Type	Model	Test on					
		DE	FR	IT	PT	JA	KR
PRE	EN	19.1	30.0	30.1	26.1	14.6	5.1
X-METRA	DE	**34.3**	33.3	30.0	30.2	13.5	8.9
	FR	19.2	**34.1**	29.9	29.1	5.8	9.0
	IT	18.3	32.2	**44.4**	30.2	6.7	10.2
	PT	19.1	27.7	30.1	**31.4**	5.8	9.0
	JA	24.1	25.7	33.2	26.1	**30.9**	20.7
	KR	24.4	25.6	34.4	25.0	20.2	**30.7**

Table 6: X-METRA results on an In-House multilingual intent data. Bold results highlight best results for each test language.

Model	Test on AR	BN	FI	ID	RU	SW	TE
MONO							
AR	74.0 ±1.1	30.1 ±2.4	50.0 ±0.8	59.5 ±1.3	48.4 ±0.8	50.8 ±1.7	24.1 ±2.7
BN	32.2 ±2.6	38.9 ±0.8	33.9 ±1.4	36.3 ±1.5	31.8 ±1.4	37.2 ±1.8	34.7 ±4.2
FI	54.2 ±2.5	30.7 ±1.3	63.3 ±1.5	52.5 ±1.7	43.0 ±2.1	48.6 ±1.7	28.7 ±2.8
ID	58.0 ±1.8	31.8 ±0.5	48.2 ±2.0	67.1 ±1.9	45.1 ±1.8	50.3 ±1.8	29.4 ±2.7
RU	50.9 ±2.3	34.5 ±2.1	45.2 ±4.2	52.0 ±4.0	54.4 ±1.3	47.1 ±2.1	30.7 ±2.5
SW	35.8 ±1.5	27.6 ±1.5	33.6 ±2.1	37.4 ±1.9	25.7 ±1.7	60.3 ±1.2	13.2 ±2.3
TE	34.0 ±0.9	38.0 ±2.2	39.5 ±0.6	35.3 ±1.1	35.9 ±1.1	43.5 ±1.0	61.4 ±1.0
FT							
AR	77.0 ±0.3	36.8 ±2.9	58.8 ±0.6	67.0 ±2.7	60.9 ±0.8	52.4 ±3.6	32.0 ±1.0
BN	60.7 ±0.4	51.0 ±2.7	59.2 ±0.6	67.1 ±1.6	59.2 ±0.3	56.2 ±0.8	43.7 ±0.9
FI	60.3 ±1.9	36.7 ±1.3	70.9 ±0.4	65.7 ±1.4	62.1 ±0.5	50.9 ±1.3	36.4 ±3.6
ID	65.7 ±1.4	37.0 ±1.1	60.8 ±0.2	77.0 ±0.4	61.1 ±0.5	56.8 ±1.0	36.7 ±0.4
RU	60.9 ±2.5	37.2 ±2.0	59.0 ±2.1	66.8 ±1.3	64.8 ±0.4	55.2 ±1.8	36.8 ±1.3
SW	57.4 ±0.5	35.2 ±1.5	56.2 ±1.0	65.4 ±1.8	58.8 ±0.8	70.2 ±1.7	33.1 ±2.8
TE	54.0 ±3.2	39.1 ±2.1	54.8 ±2.3	63.5 ±2.6	58.1 ±0.9	56.9 ±1.8	65.4 ±0.6
X-METRA							
AR	**78.4** ±0.6	33.0 ±0.8	58.2 ±0.2	66.4 ±1.4	59.9 ±0.1	53.2 ±3.8	31.4 ±3.0
BN	56.9 ±3.2	53.2 ±0.5	56.7 ±1.4	67.4 ±1.2	56.7 ±1.3	56.0 ±0.9	41.7 ±0.6
FI	58.9 ±0.6	33.6 ±1.1	72.8 ±0.3	61.9 ±2.0	60.7 ±0.9	46.5 ±1.2	36.6 ±1.7
ID	65.8 ±0.3	35.0 ±2.2	60.5 ±0.9	**77.7** ±0.2	60.4 ±1.3	57.4 ±1.1	35.3 ±0.3
RU	60.3 ±1.6	37.2 ±0.7	59.1 ±0.3	66.8 ±0.8	**66.2** ±0.1	53.7 ±0.8	33.2 ±3.1
SW	58.5 ±0.0	36.9 ±1.2	56.0 ±0.2	64.8 ±0.7	58.4 ±0.4	**71.9** ±0.2	33.7 ±1.5
TE	56.0 ±3.0	38.8 ±0.1	53.6 ±1.7	61.1 ±1.9	58.6 ±0.6	55.8 ±0.2	66.4 ±0.5
X-METRA-ADA							
AR	76.6 ±0.1	49.6 ±1.3	63.4 ±0.4	70.9 ±0.1	60.1 ±1.0	56.8 ±0.4	42.4 ±2.5
BN	59.4 ±0.3	**57.8** ±0.6	59.2 ±0.2	63.1 ±0.2	56.5 ±0.2	56.1 ±0.3	44.1 ±0.4
FI	62.8 ±1.3	50.8 ±1.3	**73.0** ±0.3	65.5 ±1.2	60.1 ±0.4	54.9 ±0.3	42.5 ±0.5
ID	66.7 ±0.3	49.9 ±0.5	62.6 ±0.7	77.3 ±0.1	58.3 ±0.9	58.1 ±0.6	42.6 ±0.4
RU	62.2 ±0.7	47.6 ±1.6	63.1 ±0.2	63.4 ±0.9	**66.9** ±0.1	56.0 ±1.1	43.3 ±1.2
SW	59.1 ±0.7	49.1 ±1.1	58.1 ±0.2	62.1 ±1.0	54.6 ±0.6	70.3 ±0.2	43.2 ±0.7
TE	58.2 ±2.8	52.1 ±1.7	61.5 ±1.0	62.0 ±0.5	58.2 ±0.5	59.7 ±1.4	**72.8** ±0.1

Table 7: Full F1 Results on TyDiQA-GoldP between external, pre-training, monolingual and fine-tuning baselines on one hand and X-METRA and X-METRA-ADA on the other hand.

E More Ablation

Figure 8 compares between the learning curves for language-specific and joint training with respect to slot filling for both Spanish and Thai.

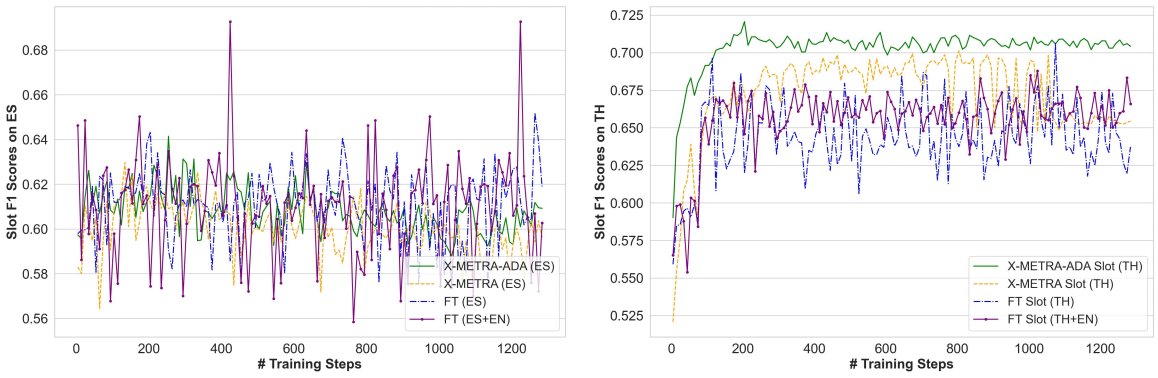
F More Analysis

More Downsampling Analysis Figure 9 shows a downsampling analysis on Spanish. Due to the typological similarity between Spanish and English, even lower percentages starting from 50% of the query reach a maximal performance for both intent classification and slot filling.

BERTology Analysis We analyze the degree of contribution of M-BERT layers by freezing each pair of layers separately. Our analysis is not conclusive as the performance doesn't change significantly between layers. We then proceed to freeze all layers of M-BERT to discover that linear layers are more important in refining the cross-lingual alignment to the target language as shown by the narrow gap between freezing vs non-freezing BERT layers in Figure 10. This can be explained by the challenge of fine-tuning M-BERT alone with many layers and higher dimensionality for such a low-resource setting.

Model	Test on AR	BN	FI	ID	RU	SW	TE
MONO							
AR	57.5 ±1.5	19.7 ±2.9	35.1 ±1.0	44.2 ±1.3	25.2 ±0.9	33.8 ±1.4	14.9 ±1.7
BN	17.1 ±1.4	24.5 ±2.9	17.5 ±0.4	20.8 ±2.0	14.4 ±0.5	20.5 ±1.4	19.9 ±5.0
FI	33.7 ±4.0	15.6 ±1.6	49.8 ±1.3	35.3 ±2.3	21.4 ±1.4	26.1 ±9.9	16.5 ±3.9
ID	39.7 ±1.4	18.6 ±1.3	32.7 ±1.9	54.9 ±0.1	23.8 ±0.6	34.4 ±1.2	16.9 ±4.9
RU	30.8 ±1.9	26.3 ±4.9	29.7 ±2.4	34.9 ±4.0	37.9 ±1.6	30.7 ±3.1	19.9 ±1.9
SW	16.0 ±1.3	16.5 ±1.5	15.6 ±1.0	21.1 ±1.3	10.5 ±1.3	48.6 ±1.2	5.3 ±1.7
TE	18.8 ±2.0	26.3 ±1.5	23.8 ±2.6	21.6 ±2.5	20.4 ±1.2	26.7 ±1.7	46.3 ±1.1
FT							
AR	61.3 ±1.0	26.5 ±4.4	43.1 ±1.0	52.2 ±2.0	37.9 ±2.5	35.6 ±3.3	21.0 ±3.0
BN	42.2 ±0.9	38.0 ±4.4	44.8 ±1.2	51.5 ±2.2	36.8 ±1.6	37.2 ±1.7	27.3 ±0.2
FI	43.2 ±1.8	23.6 ±1.1	56.5 ±0.6	50.8 ±2.1	40.5 ±0.8	33.5 ±1.2	20.7 ±3.3
ID	49.4 ±1.6	23.3 ±2.4	46.4 ±0.4	63.8 ±0.5	40.5 ±0.1	38.1 ±2.1	24.1 ±0.5
RU	42.6 ±2.6	24.8 ±3.3	43.5 ±2.0	52.4 ±2.3	46.5 ±0.4	37.6 ±1.5	24.5 ±1.3
SW	38.9 ±0.6	23.0 ±1.4	40.1 ±1.4	50.0 ±1.7	38.0 ±0.8	59.0 ±3.1	23.5 ±1.4
TE	36.1 ±2.2	30.0 ±2.3	40.0 ±2.5	49.4 ±2.1	38.6 ±0.9	39.0 ±1.7	49.2 ±0.5
X-METRA							
AR	**63.3** ±0.8	21.2 ±1.9	42.6 ±1.0	51.8 ±1.2	34.9 ±1.1	36.0 ±3.5	20.9 ±1.7
BN	29.2 ±16.5	39.0 ±1.9	41.9 ±1.6	51.1 ±1.7	34.1 ±0.4	37.1 ±1.4	25.6 ±0.2
FI	42.0 ±1.0	20.4 ±0.7	**59.1** ±1.1	46.0 ±2.7	36.8 ±1.3	30.9 ±0.6	22.5 ±0.9
ID	54.8 ±7.9	20.1 ±1.5	46.1 ±1.2	**65.2** ±0.5	38.5 ±1.9	39.6 ±0.8	23.1 ±1.4
RU	42.9 ±1.3	26.5 ±1.2	43.0 ±0.6	53.0 ±0.1	**48.9** ±0.4	35.3 ±1.0	21.6 ±2.4
SW	39.9 ±0.4	26.0 ±1.1	40.0 ±0.7	50.3 ±0.4	38.0 ±0.9	**61.4** ±0.4	23.9 ±0.7
TE	38.0 ±3.9	28.3 ±0.0	37.0 ±2.3	47.6 ±3.4	36.3 ±0.5	36.9 ±1.2	49.7 ±0.5
X-METRA-ADA							
AR	55.0 ±0.3	36.0 ±3.0	43.8 ±0.5	55.2 ±0.5	35.4 ±2.6	40.0 ±0.2	31.9 ±2.2
BN	38.4 ±0.3	**41.0** ±0.8	43.5 ±0.4	46.7 ±0.1	32.4 ±0.4	37.9 ±0.4	33.8 ±0.7
FI	40.9 ±1.1	34.2 ±1.1	57.9 ±1.0	49.0 ±1.4	35.3 ±0.3	38.0 ±0.7	30.0 ±1.0
ID	45.4 ±0.4	33.9 ±1.1	47.6 ±0.4	63.4 ±0.4	36.3 ±0.9	43.4 ±0.8	31.9 ±0.2
RU	39.4 ±0.1	34.8 ±1.5	45.1 ±0.5	48.6 ±0.9	47.5 ±0.3	39.3 ±1.3	33.8 ±1.3
SW	36.7 ±0.6	36.3 ±1.4	42.5 ±0.5	45.8 ±1.4	32.4 ±0.7	59.6 ±0.5	33.8 ±1.0
TE	37.9 ±1.9	38.1 ±2.6	44.9 ±1.4	48.0 ±0.3	38.8 ±0.4	43.5 ±1.6	**56.4** ±0.4

Table 8: Full EM Results on TyDiQA-GoldP between external, pre-training, monolingual and fine-tuning baselines on one hand, X-METRA and X-METRA-ADA on the other hand.



(a) Slot F1 on Spanish

(b) Slot F1 on Thai

Figure 8: Ablation Study on the role of the adaptation in X-METRA-ADA compared to X-METRA (MAML with only the meta-training stage) for different languages, language-specific vs joint training. All models are compared to their fine-tuning counterparts.

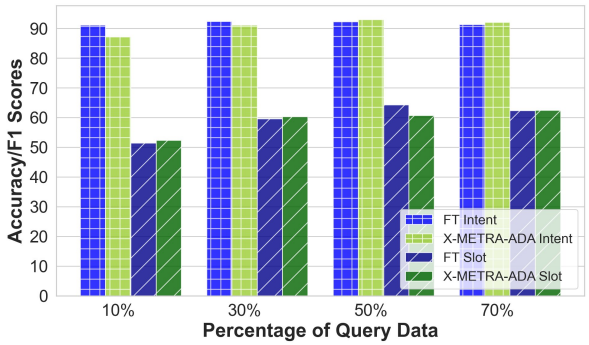


Figure 9: Downsampling Analysis for Few-shot on Spanish with Different Percentages of Query data.

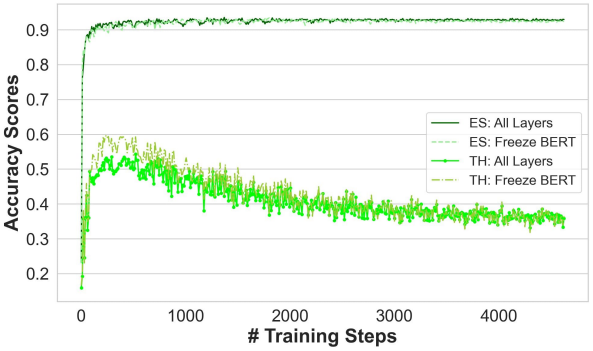


Figure 10: The effect of freezing BERT layers of X-METRA-ADA during few-shot on intent classification.