Learning what to share between tasks has been a topic of great importance recently, as strategic sharing of knowledge has been shown to improve downstream task performance. This is particularly important for multilingual applications, as most languages in the world are under-resourced. Here, we consider the setting of training models on multiple different languages at the same time, when little or no data is available for languages other than English. We show that this challenging setup can be approached using meta-learning, where, in addition to training a source language model, another model learns to select which training instances are the most beneficial to the first. We experiment using standard supervised, zero-shot cross-lingual, as well as few-shot cross-lingual settings for different natural language understanding tasks (natural language inference, question answering). Our extensive experimental setup demonstrates the consistent effectiveness of meta-learning for a total of 15 languages. We improve upon the state-of-the-art for zero-shot and few-shot NLI (on MultiNLI and XNLI) and QA (on the MLQA dataset). We perform an extensive error analysis, which reveals that cross-lingual trends can partly be explained by typological commonalities between languages.

1 Introduction

There are more than 7000 languages spoken in the world, over 90 of which have more than 10 million native speakers each (Eberhard et al., 2019). Despite this, very few languages have proper linguistic resources when it comes to natural language understanding tasks. Although there is a growing awareness in the field, as evidenced by the release of datasets such as XNLI (Conneau et al., 2018), most NLP research still only considers English (Bender, 2019). While one solution to this issue is to collect annotated data for all languages, this process is both too time-consuming and expensive to be feasible. Additionally, it is not trivial to train a model for a task in a particular language (e.g., English) and apply it directly to another language where only limited training data is available (i.e., low-resource languages). Therefore, it is essential to investigate strategies that allow one to use the large amount of training data available for English for the benefit of other languages.

Meta-learning has recently been shown to be beneficial for several machine learning tasks (Koch et al., 2015; Vinyals et al., 2016; Santoro et al., 2016; Finn et al., 2017; Ravi and Larochelle, 2017; Nichol et al., 2018). For NLP, recent work has also shown the benefits of this sharing between tasks and domains (Dou et al., 2019; Gu et al., 2018; Qian and Yu, 2019). Although cross-lingual transfer with meta-learning has been investigated for machine translation (Gu et al., 2018), this paper – to best of our knowledge – is the first attempt to study meta-learning for cross-lingual natural language understanding. Our contributions are as follows:

- We propose X-MAML, a cross-lingual meta-learning architecture, and study it for two different natural language understanding tasks (i.e., Natural Language Inference, Question Answering);
- We test it for cross-domain, cross-lingual, standard supervised, few-shot as well as zero-shot learning, across a total of 15 languages;
- We observe consistent improvements over strong models including Multilingual BERT (Devlin et al., 2018) and XLM-RoBERTa (Conneau et al., 2019);
- We perform an extensive error analysis, which reveals that cross-lingual trends can partly be explained by typological commonalities between languages.
2 Meta-Learning

Meta-learning tries to tackle the problem of fast adaptation to a handful of new training data instances. It discovers the structure among multiple tasks such that learning new tasks can be done quickly. This is done by repeatedly simulating the learning process on low-resource tasks using many high-resource ones (Gu et al., 2018). There are several ways of performing meta-learning: (i) metric-based (Koch et al., 2015; Vinyals et al., 2016); (ii) model-based (Santoro et al., 2016); and (iii) optimisation-based (Finn et al., 2017; Ravi and Larochelle, 2017; Nichol et al., 2018). Metric-based methods aim to learn similarities between feature representations of instances from different training sets given a similarity metric. For model-based architectures, the focus has been on adapting models that learn fast (e.g., memory networks) for meta-learning (Santoro et al., 2016). In this work, we focus on optimisation-based methods due to their superiority in several tasks (e.g., computer vision (Finn et al., 2017)) over the above-mentioned meta-learning architectures. These optimisation-based methods are able to find good initialisation parameter values and adapt to new tasks quickly. To the best of our knowledge, we are the first to exploit the usage of Model Agnostic Meta-Learning (MAML) which uses gradient descent and achieves a good generalisation for a variety of tasks (Finn et al., 2017). MAML is able to quickly adapt to new target tasks by using only a few instances at test time, assuming that these new target tasks are drawn from the same distribution.

Formally, MAML assumes that there is a distribution $p(T)$ of tasks $\{T_1, T_2, ..., T_k\}$. The parameters $\theta$ of model $M$ for a particular task $T_i$, sampled from the distribution $p(T)$, are updated to $\theta_i$. In particular, the parameters $\theta$ is updated using one or a few iterations of gradient descent steps on the training examples (i.e., $D_i^{train}$) of task $T_i$. For example, for one gradient update,

$$\theta_i' = \theta - \alpha \nabla_{\theta} L_{T_i}(M_{\theta})$$

where $\alpha$ is the step size, the $M_{\theta}$ is the learned model from the neural network and $L_{T_i}$ is the loss on the specific task $T_i$. The parameters of the model $\theta$ are trained to optimise the performance of $M_{\theta}$ on the unseen test examples (i.e., $D_i^{test}$) across tasks $p(T)$. The meta-learning objective is:

$$\min_{\theta} \sum_{T_i \sim p(T)} L_{T_i}(M_{\theta}) = \sum_{T_i \sim p(T)} L_{T_i}(M_{\theta - \alpha \nabla_{\theta} L_{T_i}(M_{\theta})} - \alpha \nabla_{\theta} L_{T_i}(M_{\theta}))$$

The MAML algorithm aims to optimise the model parameters via a few number of gradient steps on a new task, which we refer to as the meta-update. The meta-update across all involved tasks is performed for the $\theta$ parameters of the model using stochastic gradient descent (SGD) as:

$$\theta \leftarrow \theta - \beta \nabla_{\theta} \sum_{T_i \sim p(T)} L_{T_i}(M_{\theta})$$

where $\beta$ is the meta-update step size.

3 Cross-Lingual Meta-Learning

The underlying idea of using MAML in NLP tasks (Gu et al., 2018; Dou et al., 2019; Qian and Yu, 2019) is to employ a set of high-resource auxiliary tasks / languages to find an optimal initialisation from which learning a target task / language can be done using only a small number of training instances. In a cross-lingual setting (i.e., XNLI, MLQA), where only an English dataset is available as a high-resource language, and a small number of instances are available for other languages, the training procedure for MAML requires some non-trivial changes. For this purpose, we introduce a cross-lingual meta-learning framework (X-MAML), which uses the following training steps:

1. Pre-training on the high-resource language $h$ (i.e., English): Given all the training samples in the high-resource language $h$, we first train the model $M$ on $h$ to initialise the model parameters $\theta$.
2. Meta-learning using low-resource languages: This step consists of choosing one or more auxiliary languages from the low-resource set. Using the development set of each auxiliary language, we construct a randomly sampled batch of tasks $T_i$. Then, we update the model parameters using $K$ data points of $T_i$ ($D_i^{train}$) by one gradient descent step (see Eq. (1)). After this step, we can calculate the loss value using $Q$ examples ($D_i^{test}$) in each task. It should be noted that the $K$ data points used for training ($D_i^{train}$) are different from the $Q$ data points used for testing ($D_i^{test}$). We sum up the loss values from all tasks to minimise
Perform either (i) zero-shot or (ii) few-shot learning on $\mathcal{A}$.

A more formal description of the proposed model X-MAML is given in Algorithm 1.

**Algorithm 1: X-MAML**

**Input:** high-resource language $\mathcal{L}$, set of low-resource languages $\mathcal{L} \setminus \mathcal{A}$, Model $M$, step size $\alpha$ and learning rate $\beta$

1. Pre-train $M$ on $\mathcal{L}$ and provide initial model parameters $\theta$.
2. Select one or more languages from $\mathcal{L} \setminus \mathcal{A}$ as a set of auxiliary languages ($\mathcal{A}$).
3. While not done do
   4. For $l \in \mathcal{A}$ do
      5. Sample batch of tasks $\mathcal{T}$ using the development set of the auxiliary language $l$.
      6. For each $\mathcal{T}$ do
         7. Sample $K$ data-points to form $D_{i}^{\text{train}} = \{(X^k, Y^k)\}_{k=1}^K \in \mathcal{T}$.
         8. Sample $Q$ data-points to form $D_{i}^{\text{test}} = \{(X^q, Y^q)\}_{q=1}^Q \in \mathcal{T}$ for meta-update.
         9. Compute $\nabla_{\theta} L_{\mathcal{T}}(\theta_0)$ on $D_{i}^{\text{train}}$.
        10. Compute adapted parameters with gradient descent: $\theta' = \theta - \alpha \nabla_{\theta} L_{\mathcal{T}}(\theta_0)$.
        11. Compute $L_{\mathcal{T}}(\theta_{i'})$ using $D_{i}^{\text{test}}$.
      12. Update $\theta \leftarrow \theta - \beta \nabla_{\theta} \sum_{i} L_{\mathcal{T}}(\theta_{i'})$.
   13. Perform either (i) zero-shot or (ii) few-shot learning on $\{\mathcal{L} \setminus \mathcal{A}\}$ using meta-learned parameters $\theta$.

The Cross-Lingual Natural Language Inference (XNLI) dataset (Conneau et al., 2018) consists of 5000 test and 2500 development hypothesis-premise pairs with their textual entailment labels for English. Translations of these pairs are provided in 14 languages: French (fr), Spanish (es), German (de), Greek (el), Bulgarian (bg), Russian (ru), Turkish (tr), Arabic (ar), Vietnamese (vi), Thai (th), Chinese (zh), Hindi (hi), Swahili (sw) and Urdu (ur). XNLI provides a multilingual benchmark to evaluate how to perform inference in low-resource languages, in which only training data for the high-resource language English is available from MultiNLI. This allows us to study the impact of meta-learning with one low-resource language to serve as an auxiliary language, and evaluate the resulting NLI model on the target languages provided in the XNLI test set.

**Natural Language Inference (NLI):** NLI is the task of predicting whether a hypothesis sentence is true (entailment), false (contradiction), or undetermined (neutral) given a premise sentence. The Multi-Genre Natural Language Inference (MultiNLI) dataset has 433k sentence pairs annotated with textual entailment information (Williams et al., 2018). It covers a range of different genres of spoken and written text and thus supports cross-genre evaluation. The NLI premise sentences are provided in 10 different genres: facetoface, telephone, verbatim, state, government, fiction, letters, nineeleven, travel and oup. All of the genres appear in the test and development sets, but only five are included in the training set. To verify our learning routine more generally, we define $\mathcal{T}$ as an NLI task in each genre. We exploit MAML, in its original setting, to investigate whether meta-learning encourages the model to learn a good initialisation for all target genres, which can then be fine-tuned with limited supervision for each genre’s development instances (2000 examples) to achieve a good performance on its test set.

The Cross-Lingual Natural Language Inference (XNLI) dataset (Conneau et al., 2018) consists of 5000 test and 2500 development hypothesis-premise pairs with their textual entailment labels for English. Translations of these pairs are provided in 14 languages: French (fr), Spanish (es), German (de), Greek (el), Bulgarian (bg), Russian (ru), Turkish (tr), Arabic (ar), Vietnamese (vi), Thai (th), Chinese (zh), Hindi (hi), Swahili (sw) and Urdu (ur). XNLI provides a multilingual benchmark to evaluate how to perform inference in low-resource languages, in which only training data for the high-resource language English is available from MultiNLI. This allows us to study the impact of meta-learning with one low-resource language to serve as an auxiliary language, and evaluate the resulting NLI model on the target languages provided in the XNLI test set.

**Question Answering (QA):** Given a context and a question, the task in QA is to identify the span in the context which answers the question. Lewis et al. (2019) introduce a Multilingual Question Answering dataset (MLQA) that contains QA instances in 7 languages: English (en), Arabic (ar), German (de), Spanish (es), Hindi (hi), Vietnamese (vi) and Simplified Chinese (zh). It includes over 12k QA instances in English and 5k for every other language, with each QA instance being available in 4 languages (on average). This dataset has been used in many recent studies on cross-lingual transfer learning (e.g., Hu et al. (2020); Liang et al. (2020)). In our experiments, we investigate meta-learning
for QA with one or two auxiliary languages.

4 Experiments

We want to investigate how meta-learning can be used for cross-lingual sharing. We experiment with different architectures in order to verify that our method generalises across them, further detailed below. We report results for few-shot, zero-shot cross-domain and cross-lingual learning.\(^1\)

NLI: We experiment with two different settings. (i) For MultiNLI, a cross-genre dataset, we employ the Enhanced Sequential Inference Model (ESIM) (Chen et al., 2016), which is commonly used for textual entailment problems. ESIM uses LSTMs with attention to create a rich representation, capturing the relationship between premise and hypothesis sentences. (ii) For XNLI, a cross-lingual dataset, we use the PyTorch version of BERT\(^2\) (Devlin et al., 2018) as the underlying model \(M\). However, since our proposed meta-learning method is model-agnostic, it can easily be extended to any other architecture. Note that for Setting (i), we apply MAML, whereas for Setting (ii), we apply X-MAML on the original English BERT model (En-BERT) and on Multilingual BERT (Multi-BERT) models. As the first training step (i.e., pre-training on a high-resource language, see Step 1 in Section 3), we first initialise the model parameters \(\theta\) learned by MAML. We then continue to fine-tune the model using the development set of MultiNLI and report the accuracy on the English test set of XNLI. We proportionally select sub-samples \(x = [1\%, 2\%, 3\%, 5\%, 10\%, 20\%, 50\%, 100\%]\) from the training data (with random sampling). The results obtained by training on the corresponding proportions \((x\%)\) of the MultiNLI dataset using ESIM (as the learner model \(M\)) are shown in Table 1. We observe that for both settings (i.e., MAML on the training (5 tasks) and on development set (10 tasks)), the performance of all models (including baselines) improves as more instances become available. However, the effectiveness of MAML is larger when only limited training data is available (improving by 12\% in accuracy when 2\% of the data is available on the development set).

QA: For question answering, we use different base models \(M\) for X-MAML, namely XLM (Conneau and Lample, 2019) and XLM-RoBERTa (XLM-R) (Conneau et al., 2019), both state-of-the-art models. XLM uses a similar pre-training objective as Multi-BERT with a larger model, a larger shared vocabulary, and leverages both monolingual and parallel data. XLM-R is a RoBERTa version of XLM and is trained on a much larger multilingual corpus (i.e., Common Crawl), achieving state-of-the-art performance on most cross-lingual benchmarks (Hu et al., 2020). We employ the XLM-15 (Masked Language Model + Translation Language Model, 15 languages), XLM-R\(_{base}\) and XLM-R\(_{large}\) models released by the authors. The SQuAD v1.1 training data is used in the pre-training step of X-MAML (see Step 1 in Section 3). We use the cross-lingual development and test splits provided in the MLQA dataset for the meta-learning and evaluation steps, respectively.

Baselines: We create: (i) zero-shot baselines: directly evaluate the model on the test set of the target languages; (ii) few-shot baselines: fine-tune the model on the development set, then evaluate on the test set of the low-resource languages.

4.1 Few-Shot Cross-Domain NLI

We train ESIM on the MultiNLI training set to provide initial model parameters \(\theta\) (see Step 1 in Section 3). We evaluate the pre-trained model on the English test set of XNLI (since the MultiNLI test set is not publicly available) as a baseline. Since MultiNLI is already split into genres, we use each genre as a task within MAML. We then include either the training set (5 genres) or the development set (10 genres) during meta-learning (similar to Step 2 in X-MAML).

In the last phase (similar to Step 3 in X-MAML), we first initialise the model parameters with those learned by MAML. We then continue to fine-tune the model using the development set of MultiNLI and report the accuracy on the English test set of XNLI. We proportionally select sub-samples \(x = [1\%, 2\%, 3\%, 5\%, 10\%, 20\%, 50\%, 100\%]\) from the training data (with random sampling). The results obtained by training on the corresponding proportions \((x\%)\) of the MultiNLI dataset using ESIM (as the learner model \(M\)) are shown in Table 1. We observe that for both settings (i.e., MAML on the training (5 tasks) and on the development set (10 tasks)), the performance of all models (including baselines) improves as more instances become available. However, the effectiveness of MAML is larger when only limited training data is available (improving by 12\% in accuracy when 2\% of the data is available on the development set).

4.2 Zero- and Few-Shot Cross-Lingual NLI

Zero-Shot Learning: In this set of experiments, we employ the proposed framework (i.e., X-MAML) within a zero-shot setup, in which we do not fine-tune after the meta-learning step. We report the impact of meta-learning for each target language as a difference in accuracy with and without meta-learning on top of the baseline model (Multi-BERT) on the test set (Fig. 1). Each column corresponds to the performance of Multi-BERT af-
Baseline meta-learning with a single auxiliary language, and evaluation on the target language of the XNLI test set. Overall, we observe that our zero-shot approach with X-MAML outperforms the baseline model without MAML and results reported by Devlin et al. (2018). This way, we improve the state-of-the-art performance for zero-shot cross-lingual NLI (in several languages for up to +3.6% in accuracy, e.g., Hindi (hi) as target and Urdu (ur) as auxiliary language). For the exact accuracy scores, we refer to Table 5 in the Appendix. We hypothesise that the degree of typological commonalities among the languages has an effect (i.e., positive or negative) on the performance of X-MAML. It can be observed that the proposed learning approach provides positive impacts across most of the target languages. However, including Swahili (sw) as an auxiliary language in X-MAML is not beneficial for the performance on the other target languages. It is worth noting that we experimented by just training the model using an auxiliary language, instead of performing meta-learning (step 2). From this experiment, we observe that meta-learning has a strongly positive effect on predictive performance (see also Fig. 2 in the Appendix).

In Table 2, we include the original baseline performances reported in Devlin et al. (2018) and Wu and Dredze (2019). We report the average and maximum performance by using one auxiliary language for each target language. We also report the performance of X-MAML by also using Hindi (which is the most effective auxiliary language for the zero-shot setting, as shown in Fig. 1). We suspect that this is because of the typological similarities between Hindi (hi) and other languages. Furthermore, by using two auxiliary languages in X-MAML results to the largest benefit in our zero-shot experiments.

Few-Shot Learning: For few-shot learning, meta-learning in X-MAML (Step 3) is performed by fine-tuning on the development set (2.5k instances) of target languages, and then evaluating on the test set. Detailed ablation results are presented in the Appendix (Table 6 and Fig. 4). In Table 2, we compare X-MAML results with one or two auxiliary languages to the external and internal baselines. We also showcase the performance using specifically Swahili (sw), the overall most effective auxiliary language for meta-learning with Multi-BERT in the few-shot learning setting. In addition, we report
### Zero-shot cross-lingual transfer

<table>
<thead>
<tr>
<th></th>
<th>en</th>
<th>fr</th>
<th>es</th>
<th>de</th>
<th>el</th>
<th>bg</th>
<th>ru</th>
<th>tr</th>
<th>ar</th>
<th>vi</th>
<th>zh</th>
<th>hi</th>
<th>sw</th>
<th>ur</th>
<th>avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Devlin et al. (2018)</td>
<td>81.4</td>
<td>-</td>
<td>74.3</td>
<td>70.5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>58.35</td>
</tr>
<tr>
<td>Wu and Dredze (2019)</td>
<td>82.1</td>
<td>73.8</td>
<td>74.3</td>
<td>71.1</td>
<td>66.4</td>
<td>68.9</td>
<td>69.0</td>
<td>61.6</td>
<td>64.9</td>
<td>69.5</td>
<td>55.8</td>
<td>69.3</td>
<td>60.0</td>
<td>50.4</td>
<td>58.0</td>
</tr>
<tr>
<td>Multi-BERT (Our baseline)</td>
<td>81.36</td>
<td>73.45</td>
<td>73.85</td>
<td>69.74</td>
<td>65.73</td>
<td>67.82</td>
<td>67.94</td>
<td>59.04</td>
<td>64.63</td>
<td>70.12</td>
<td>52.46</td>
<td>68.90</td>
<td>58.56</td>
<td>47.58</td>
<td>58.70</td>
</tr>
</tbody>
</table>

**X-MAML (One aux. lang.):**

- **AVG:**
  - 81.69 73.86 74.43 71.00 67.16 68.39 68.90 60.41 64.33 70.95 54.08 70.09 60.51 47.97 59.94
  - 82.09 74.42 75.07 71.83 67.95 69.45 70.19 61.20 66.05 71.82 55.39 71.11 62.20 49.76 61.51

**X-MAML (Two aux. lang.):**

- (hi,de) (hi,ar) (fr,de) (bg,zh) (ru,ru) (hi,ru) (de,bg) (ur,sw) (el,tr) (en,ru) (hi,de)

| (l₁, l₂) → X | 82.59 75.69 75.97 73.45 69.16 71.42 71.44 62.57 67.19 73.27 62.63 65.73 73.13 63.53 70.42 62.93 |

**Few-Shot learning:**

- Multi-BERT (Our baseline) | 81.94 | 75.39 | 75.79 | 73.25 | 69.54 | 71.60 | 70.84 | 64.85 | 67.37 | 73.23 | 61.18 | 73.93 | 64.37 | 57.82 | 63.71 |

**X-MAML (One aux. lang.):**

- 82.22 75.24 76.06 73.34 69.97 71.80 72.28 64.76 67.82 73.41 65.74 70.62 64.97 65.23 68.31
- 82.39 75.32 76.18 73.46 70.03 71.94 71.45 64.92 67.95 73.52 67.41 74.21 64.97 58.23 70.01

**X-MAML (Two aux. lang.):**

- (ru,ru) (ru,hi) (ru,th) (el,hi) (sv,vi) (az,az) (de,de) (cz,cz) (en,ru)

| (l₁, l₂) → X | 82.71 75.97 76.51 74.07 70.66 72.77 72.12 65.69 68.4 73.87 62.5 74.85 65.75 59.94 64.59 |

**Machine translate at test (TRANSLATE-TEST):**

- Devlin et al. (2018) | 81.4 | - | 74.9 | 74.4 | - | - | - | 70.4 | - | 70.1 | - | - | 62.1 | - |

**Machine translate at training (TRANSLATE-TRAIN):**

- Wu and Dredze (2019) | 82.1 | 76.9 | 78.5 | 74.8 | 72.1 | 75.4 | 74.3 | 70.6 | 70.8 | 67.8 | 63.2 | 76.2 | 65.3 | 60.6 | 71.6 |

### Table 2: Accuracy results on the XNLI test set for zero- and few-shot X-MAML. Columns indicate the target languages. The models of Devlin et al. (2018) and Wu and Dredze (2019) are also Multi-BERT models. For our Multi-BERT baseline model for (i) zero-shot learning, we evaluate the pre-trained model on the test set of the target language; and for (ii) few-shot learning, we fine-tune the model on the development set and evaluate on the test set of the target language. The avg column indicates row-wise average accuracy. We also report the average (AVG) and maximum (MAX) performance by using one auxiliary language for each target language. (l₁, l₂) are the most beneficial auxiliary languages for X-MAML in improving the test accuracy of each target language X. In TRANSLATE-TEST (Devlin et al., 2018), the target language test data is translated to English and then the model is fine-tuned on English. In TRANSLATE-TRAIN (Wu and Dredze, 2019), the English training data is translated to the target language and the model is fine-tuned using the translated data.

Results from Devlin et al. (2018) that use machine translation at test time (TRANSLATE-TEST) and results from Wu and Dredze (2019) that use machine translation at training time (TRANSLATE-TRAIN). Note that, using X-MAML, we are able to avoid the machine translation step (TRANSLATE-TEST) from the target language into English. The results also indicate that X-MAML boosts Multi-BERT performance on XNLI. It is worthwhile mentioning that Multi-BERT in the TRANSLATE-TRAIN setup outperforms few-shot X-MAML, however, we only use 2k development examples from the target languages, whereas in the aforementioned work, 433k translated sentences are used for fine-tuning.

### 4.3 Zero-Shot Cross-Lingual QA

We use a similar approach for cross-lingual QA on the MLQA dataset. Zero-shot results on MLQA are shown in Table 3. We compare our results to those reported in two benchmark papers, Hu et al. (2020) and Liang et al. (2020). We also report our own baselines for the task. The baselines are provided by training each base model on the SQuAD v1.1 train set (see Step 1 in Section 3) and evaluating on the test set of MLQA. All target languages benefit from meta-learning with at least one of the auxiliary languages. Using two auxiliary languages in X-MAML further improves results. Overall, zero-shot learning models with X-MAML outperform both internal and external baselines. The improvement is +1.04%, +0.89% and +1.47% in average F₁ score compared to XLM-15, XLM-R<sub>base</sub> and XLM-R<sub>large</sub>, respectively.

We also evaluate on the less widely used cross-lingual QA dataset X-WikiRE (Abdou et al., 2019), for which we observe similar result trends, and 0.55% improvement in terms of average F₁ score on zero-shot QA. More details can be found in the appendix (Section A.2).
Cross-Lingual NLU:

Cross-lingual learning has a fairly short history in NLP, and has mainly been restricted to traditional NLP tasks, such as PoS tagging and parsing. In contrast to these tasks, which have seen much cross-lingual attention (Plank et al., 2016; Bjerva, 2017; de Lhoneux et al., 2018), there has seen much cross-lingual attention (Plank et al., 2016; Bjerva, 2017; de Lhoneux et al., 2018), there has been relatively little work on cross-lingual NLU, partly due to lack of benchmark datasets. Existing work has mainly been focused on NLI (Conneau et al., 2018; Agic and Schluter, 2018), and to a lesser degree on RE (Verga et al., 2016; Faruqui and Kumar, 2015) and QA (Lewis et al., 2019; Abdou et al., 2019). Previous research generally reports that cross-lingual learning is challenging and that it is hard to beat a machine translation baseline (e.g., Conneau et al. (2018)). Such a baseline is (for instance) suggested by Faruqui and Kumar (2015), where the text in the target language is automatically translated to English. We achieve competitive performance compared to a machine translation baseline (for XNLI, and propose a method that requires no training instances for the target task in the target language. Furthermore, our method is model agnostic, and can be used to extend any pre-existing model.

6 Discussion and Analysis

Cross-Lingual Transfer: Somewhat surprisingly, we find that cross-lingual transfer with meta-
learning yields improved results even when languages strongly differ from one another. For instance, for zero-shot meta-learning on XNLI, we observe gains for almost all auxiliary languages, with the exception of Swahili (sw). This indicates that the meta-parameters learned with X-MAML are sufficiently language agnostic, as we otherwise would not expect to see any benefits in transferring from, e.g., Russian (ru) to Hindi (hi) (one of the strongest results in Fig. 1). This is dependent on having access to a pre-trained multilingual model such as BERT, however, using monolingual BERT (En-BERT) yields overwhelmingly positive gains in some target/auxiliary settings (see additional results in Fig. 3 in the Appendix). For few-shot learning, our findings are similar, as almost all combinations of auxiliary and target languages lead to improvements when using Multi-BERT (Fig. 4 in the Appendix). However, when we only have access to a handful of training instances as in few-shot learning, even the English BERT model mostly leads to improvements in this setting (see additional results in Fig. 5 in the Appendix).

**Typological Correlations:** In order to better explain our results for cross-lingual zero-shot and few-shot learning, we investigate typological features, and their overlap between target and auxiliary languages. We evaluate on the World Atlas of Language Structure (WALS, Dryer and Haspelmath (2013)), which is the largest openly available typological database. It comprises approximately 200 linguistic features with annotations for more than 2500 languages, which have been made by expert typologists through study of grammars and field work. We draw inspiration from previous work (Bjerva and Augenstein, 2018a,b) which attempts to predict typological features based on language representations learned under various NLP tasks. Similarly, we experiment with two conditions: (i) we attempt to predict typological features based on the mutual gain/loss in performance using X-MAML; (ii) we investigate whether sharing between two typologically similar languages is beneficial for performance using X-MAML. We train a simple logistic regression classifier per condition above, for each WALS feature. In the first condition (i), the task is to predict the exact WALS feature value of a language, given the change in accuracy in combination with other languages. In the second condition (ii), the task is to predict whether a main and auxiliary language have the same WALS feature value, given the change in accuracy when the two languages are used in X-MAML. We compare with two simple baselines, one based on always predicting the most frequent feature value in the training set, and one based on predicting feature values with respect to the distribution of feature values in the training set. We then investigate whether any features could be consistently predicted above baseline levels, given different test-training splits. We apply a simple paired t-test to compare our models predictions to the baselines. As we are running a large number of tests (one per WALS feature), we apply Bonferroni correction, changing our cut-off p-value from $p = 0.05$ to $p = 0.00025$.

We first investigate few-shot X-MAML, when using Multi-BERT, as reported in Table 6 (Appendix). We find that languages sharing the feature value for WALS feature $67A$ The Future Tense are beneficial to each other. This feature encodes whether or not a language has an inflectional marking of future tense, and can be considered to be a morphosyntactic feature. We next look at zero-shot X-MAML with Multi-BERT, as reported in Table 5 (Appendix). For this case, we find that languages sharing a feature value for the WALS feature $25A$ Locus of Marking: Whole-language Typology typically help each other. This feature describes whether the morphosyntactic marking in a language is on the syntactic heads or dependents of a phrase. For example en, de, ru, and zh are ‘dependent-marking’ in this feature. And if we look at the results in Fig. 1, they have the largest mutual gains from each other during the zero-shot X-MAML. In both cases, we thus find that languages with similar morphosyntactic properties can be beneficial to one another when using X-MAML.

7 Conclusion

In this work, we show that meta-learning can be used to effectively leverage training data from an auxiliary language for zero-shot and few-shot cross-lingual transfer. We evaluated this on two challenging NLU tasks (NLI and QA), and on a total of 15 languages. We are able to improve the performance of state-of-the-art baseline models for (i) zero-shot XNLI, and (ii) zero-shot QA on the MLQA dataset. Furthermore, we show in a typological analysis that languages which share certain morphosyntactic features tend to benefit from this type of transfer. Future studies will extend this work to other cross-lingual NLP tasks and more languages.
8 Acknowledgements

This research has received funding from the Swedish Research Council under grant agreement No 2019-04129, as well as the Research Foundation - Flanders (FWO). This work was also funded by UiO: Energy to support international mobility. We are grateful to the Nordic Language Processing Laboratory (NLPL) for providing access to its supercluster infrastructure.

References


Emily M Bender. 2019. The #BenderRule: On Naming the Languages We Study and Why It Matters.


Miryam de Lhoneux, Johannes Bjerva, Isabelle Au-
Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian
Alex Nichol, Joshua Achiam, and John Schulman.
Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu,
Omer Levy, Minjoon Seo, Eunsol Choi, and Luke
Gregory Koch, Richard Zemel, and Ruslan Salakhutdi-
Diederik P. Kingma and Jimmy Ba. 2014. Adam: A
Souvik Kundu and Hwee Tou Ng. 2018. A nil-aware
Junjie Hu, Sebastian Ruder, Aditya Siddhant, Gra-
Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan
Yao, Zhiyuan Liu, and Maosong Sun. 2018. Fewrel:
A large-scale supervised few-shot relation classifica-
tion dataset with state-of-the-art evaluation. In Pro-
cedings of the 2018 Conference on Empirical Meth-
dods in Natural Language Processing, pages 4803–
4809.
Diederik P. Kingma and Jimmy Ba. 2014. Adam: A
method for stochastic optimization.
Gregory Koch, Richard Zemel, and Ruslan Salakhutdi-
ov. 2015. Siamese neural networks for one-shot im-
age recognition. In ICML deep learning workshop.
Souvik Kundu and Hwee Tou Ng. 2018. A nil-aware
answer extraction framework for question answering. In Proceedings of the 2018 Conference on Em-
pirical Methods in Natural Language Processing, pages 4243–4252, Brussels, Belgium. Association for Compu-
tational Linguistics.
Omer Levy, Minjoon Seo, Eunsol Choi, and Luke
Zettlemoyer. 2017. Zero-shot relation extraction via
reading comprehension. In Proceedings of the 21st
Conference on Computational Natural Language
Learning (CoNLL 2017), pages 333–342, Vancouver,
Canada. Association for Computational Linguistics.
Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian
Riedel, and Holger Schwenk. 2019. Mlqa: Evalu-
ating cross-lingual extractive question answering.
Mirjam de Lhoneux, Johannes Bjerva, Isabelle Au-
genstein, and Anders Søgaard. 2018. Parameter
sharing between dependency parsers for related lan-
guages. In Proceedings of the 2018 Conference on
Empirical Methods in Natural Language Processing,
pages 4992–4997, Brussels, Belgium. 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,
Bruce Zhang, Rahul Agrawal, Edward Cui, Sining
Wei, Taroon Bharti, Jiun-Hung Chen, Winnie Wu,
Xglue: A new benchmark dataset for cross-lingual
pre-training, understanding and generation. ArXiv,
Alex Nichol, Joshua Achiam, and John Schulman.
Barbara Plank, Anders Søgaard, and Yoav Goldberg.
2016. Multilingual part-of-speech tagging with bidi-
rectional long short-term memory models and auxil-
iary loss. In Proceedings of the 54th Annual Meet-
ing of the Association for Computational Linguistics
(Volume 2: Short Papers), volume 2, pages 412–418.
Kun Qian and Zhou Yu. 2019. Domain adaptive dia-
olog generation via meta learning. In Proceedings of
the 57th Annual Meeting of the Association for Com-
putational Linguistics, pages 2639–2649, Florence,
Italy. Association for Computational Linguistics.
Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev,
and Percy Liang. 2016. Squad: 100, 000+ ques-
tions for machine comprehension of text. CoRR,
abs/1606.05250.
Sachin Ravi and Hugo Larochelle. 2017. Optimization
as a model for few-shot learning. In International
Conference on Learning Representations.
Adam Santoro, Sergey Bartunov, Matthew Botvinick,
Daan Wierstra, and Timothy Lillicrap. 2016. Meta-
learning with memory-augmented neural networks.
In International conference on machine learning,
pages 1842–1850.
Patrick Verga, David Belanger, Emma Strubell, Ben-
jamin Roth, and Andrew McCallum. 2016. Multilin-
gual Relation Extraction using Compositional Uni-
versal Schema. In Proceedings of the 2016 Confer-
ce of the North American Chapter of the Associa-
tion for Computational Linguistics: Human Lan-
guage Technologies, pages 886–896. Association for
Computational Linguistics.
Oriol Vinyals, Charles Blundell, Timothy Lillicrap,
one shot learning. In Advances in neural informa-
tion processing systems, pages 3630–3638.
Alex Wang, Amanpreet Singh, Julian Michael, Felix
GLUE: A multi-task benchmark and analysis plat-
form for natural language understanding. CoRR,
abs/1804.07461.
Adina Williams, Nikita Nangia, and Samuel Bowman.
2018. A broad-coverage challenge corpus for sen-
tence understanding through inference. In Proceed-
ings of the 2018 Conference of the North American
Chapter of the Association for Computational Lin-
guistics: Human Language Technologies, Volume 1
(Long Papers), pages 1112–1122, New Orleans,
Louisiana. Association for Computational Linguis-
tics.
Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas:
The surprising cross-lingual effectiveness of BERT.
CoRR, abs/1904.09077.
Mo Yu, Xiaoxiao Guo, Jinfeng Yi, Shiyu Chang, Saloni
Potdar, Yu Cheng, Gerald Tesauro, Haoyu Wang,
and Bowen Zhou. 2018. Diverse few-shot text clas-
sification with multiple metrics. In Proceedings of
A Appendices

A.1 Experimental Setup

We implement X-MAML using the *higher* library.\(^4\) We use the Adam optimiser (Kingma and Ba, 2014) with a batch size of 32 for both zero-shot and few-shot learning. We fix the step size $\alpha$ and learning rate $\beta$ to $1 \times 10^{-4}$ and $1 \times 10^{-5}$, respectively. We experimented using $[10, 20, 30, 50, 100, 200, 300]$ meta-learning iterations in X-MAML. However, 100 iterations led to the best results in our experiments. The sample sizes $K$ and $Q$ in X-MAML are equal to 16 for each dataset. We report results for each experiment by averaging the performance over ten different runs.

A.2 X-MAML using X-WikiRE dataset

**X-WikiRE:** Levy et al. (2017) frame the Relation Extraction (RE) task as a QA problem using pre-defined natural language question templates. For example, a relation type such as author is transformed to at least one language question template (e.g., *who is the author of x?*, where x is an entity). Building on the work of Levy et al. (2017), Abdou et al. (2019) introduce a new dataset (called X-WikiRE) for multilingual QA-based relation extraction in five languages (i.e., English, French, Spanish, Italian and German). Each instance in the dataset includes a question, a context and an answer. The question is a transformation of a target relation and the context may contain the answer. If the answer is not present, it is marked as NIL. In the QA task, we evaluate the performance of our method on the UnENT setting of the X-WikiRE dataset, where the goal is to generalise to unseen entities. For the evaluation, we use F$_1$ scores (for questions with valid answers) similar to Kundu and Ng (2018).

**QA experiments:** We use the Nil-Aware Answer Extraction Framework (NAMANDA, Kundu and Ng (2018))\(^5\) as the base model $M$ in X-MAML for our QA experiments. NAMANDA encodes the question and context sequences to compute a similarity matrix. It creates evidence vectors through joint encoding of question and context and applies multi-factor self-attentive encoding. Finally, the evidence vectors are decomposed to output either the answer to the question or NIL. We set the parameters to the default values (as in the original work) for the training and evaluation phases. The NAMANDA model $M$ is pre-trained on the full English training set (1M instances - see Step 1 in our training algorithm). The model $M$ is further used by our meta-learning step to adapt the pre-trained QA model. We then evaluate how well the English model has been adapted by each of the auxiliary language through X-MAML via performing either few- or zero-shot learning. In few-shot X-MAML, the meta-learned $M$ in fine-tuned on the development set (1k instances) of other languages (i.e., fr, es, it and de). For both few- and zero-shot learning, we evaluate on the 10k test set of each of the target languages. Following the work of Abdou et al. (2019), the Multi-BERT model is used to jointly encode text for different languages in the QA model.

**Zero- and Few-Shot Cross-Lingual QA:** We use a similar approach for cross-lingual QA on the X-WikiRE dataset. Table 4 shows the results of both zero- and few-shot X-MAML for the UnENT part (i.e., generalise to unseen entities) of the X-WikiRE dataset. We compare our results for the UnENT scenario on the X-WikiRE dataset to those reported in the original paper. All of the target languages benefit from at least one of the auxiliary languages by adapting the model using X-MAML, highlighting the benefits of this method. We were not able to directly reproduce the result for the zero-shot scenario of the original paper, thus we also report our own baseline for the task. We find that: (i) our zero-shot results with X-MAML improve on those without meta-learning (i.e., baselines); and (ii) we outperform Abdou et al. (2019) for the UnENT scenario of zero-shot cross-lingual QA. Furthermore, for the few-shot scenario, adapting the QA model using few-shot X-MAML with only 1k development data outperforms their cross-lingual transfer model where Abdou et al. (2019) use 10k in the fine-tuning phase.

\(^4\)https://github.com/facebookresearch/higher
\(^5\)https://github.com/nusnlp/namanda
Table 4: \(F_1\) scores (average over 10 runs) for the test set of the UnENT part of the X-WikiRE dataset using zero- and few-shot X-MAML. Baseline for (i) zero-shot learning: we evaluate the pre-trained NAMANDA model on the test set of the target language indicated in each row; and for (ii) few-shot learning: we fine-tune the model on the development set and evaluate on the test set of the target language. We report results with few-shot X-MAML with only 1k instances from the development set.

```
<table>
<thead>
<tr>
<th>Auxiliary language</th>
<th>Baseline</th>
<th>Abdou et al. (2019)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BERT</td>
<td>fastText</td>
</tr>
<tr>
<td>zero-shot</td>
<td></td>
<td></td>
</tr>
<tr>
<td>es</td>
<td>49.01</td>
<td>50.11</td>
</tr>
<tr>
<td>fr</td>
<td>52.20</td>
<td>50.59</td>
</tr>
<tr>
<td>it</td>
<td>50.53</td>
<td>50.65</td>
</tr>
<tr>
<td>de</td>
<td>49.92</td>
<td>48.78</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>few-shot</td>
<td></td>
<td></td>
</tr>
<tr>
<td>es</td>
<td>78.09</td>
<td>78.33</td>
</tr>
<tr>
<td>fr</td>
<td>80.68</td>
<td>80.81</td>
</tr>
<tr>
<td>it</td>
<td>82.04</td>
<td>81.76</td>
</tr>
<tr>
<td>de</td>
<td>78.29</td>
<td>78.48</td>
</tr>
</tbody>
</table>
```

Figure 2: Differences in performance in terms of accuracy scores on the test set for the zero-shot case using training (without meta-learning) on XNLI with the Multi-BERT model. Rows correspond to target and columns to auxiliary languages. Numbers on the off-diagonal indicate performance differences between training on the auxiliary languages (without meta-learning) and the baseline model in the same row. The coloring scheme indicates the differences in performance (e.g., blue for large improvement).
Figure 3: Differences in performance in terms of accuracy scores on the test set for zero-shot X-MAML on XNLI using the En-BERT (English) model. Rows correspond to target and columns to auxiliary languages used in X-MAML. Numbers on the off-diagonal indicate performance differences between X-MAML and the baseline model in the same row. The coloring scheme indicates the differences in performance (e.g., blue for large improvement).
Table 5: The performance in terms of average test accuracy for the zero-shot setting over 10 runs of X-MAML on the XNLI dataset using Multi-BERT (multilingual BERT), as base model. Each column corresponds to the performance of the Multi-BERT system after meta-learning with a single auxiliary language, and evaluation on the target language of the XNLI test set. The auxiliary language is not included during the evaluation phase. Results of the Multi-BERT model without X-MAML (baseline) are also reported.

| Target Language | ar  | bg  | de  | el  | en  | es  | fr  | hi  | ru  | sw  | th  | tr  | ur  | vi  | zh  | baseline |
|-----------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-------|
| ar              | -   | 65.76 | 65.48 | 66.05 | 64.41 | 65.27 | 65.24 | 65.86 | 65.31 | 63.66 | 65.25 | 65.58 | 65.56 | 65.84 | 65.32 | 64.63 |
| bg              | 68.36 | -   | 68.79 | 68.39 | 68.95 | 68.45 | 68.80 | 68.86 | 69.41 | 66.10 | 67.62 | 67.95 | 68.63 | 68.67 | 69.45 | 67.82 |
| de              | 70.88 | 71.46 | -   | 71.26 | 71.09 | 71.12 | 71.11 | 71.59 | 71.83 | 68.65 | 70.29 | 70.37 | 71.42 | 71.15 | 71.83 | 69.74 |
| el              | 67.53 | 67.58 | 67.25 | -   | 66.11 | 67.13 | 67.39 | 67.95 | 67.71 | 65.11 | 67.12 | 67.15 | 67.69 | 67.19 | 67.53 | 65.73 |
| en              | 81.68 | 81.79 | 82.02 | 81.77 | -   | 81.88 | 81.91 | 81.88 | 82.03 | 80.44 | 81.18 | 81.43 | 81.80 | 81.73 | 82.09 | 81.36 |
| es              | 74.48 | 74.51 | 74.63 | 74.58 | 74.41 | -   | 74.95 | 74.81 | 74.63 | 72.66 | 73.91 | 74.12 | 74.51 | 74.71 | 75.07 | 73.85 |
| fr              | 74.13 | 74.02 | 74.22 | 74.11 | 73.75 | 74.18 | -   | 74.17 | 74.34 | 71.87 | 73.04 | 73.41 | 74.15 | 74.21 | 74.42 | 73.45 |
| hi              | 60.75 | 61.59 | 60.84 | 60.61 | 59.31 | 60.18 | 60.66 | -   | 61.75 | 57.10 | 59.39 | 60.47 | 62.20 | 60.76 | 61.56 | 58.56 |
| ru              | 68.78 | 69.47 | 69.47 | 68.93 | 68.64 | 68.89 | 69.25 | 69.44 | -   | 66.11 | 68.18 | 68.72 | 69.52 | 69.02 | 70.19 | 67.94 |
| sw              | 48.71 | 48.53 | 47.36 | 49.13 | 46.70 | 48.43 | 47.81 | 47.11 | 47.28 | -   | 49.20 | 49.76 | 46.61 | 48.43 | 46.50 | 47.58 |
| th              | 54.65 | 55.39 | 53.80 | 54.98 | 51.14 | 54.09 | 54.15 | 55.26 | 53.82 | 52.90 | -   | 55.24 | 53.79 | 54.99 | 52.85 | 52.46 |
| tr              | 60.94 | 61.20 | 60.22 | 61.09 | 58.66 | 60.60 | 60.32 | 60.93 | 60.29 | 59.98 | 60.53 | -   | 60.82 | 60.68 | 59.47 | 59.04 |
| ur              | 60.30 | 60.87 | 60.34 | 60.20 | 58.82 | 59.81 | 60.12 | 61.51 | 61.02 | 56.37 | 59.38 | 60.02 | -   | 59.87 | 60.46 | 58.70 |
| vi              | 71.27 | 71.56 | 71.32 | 70.35 | 71.22 | 71.42 | 71.57 | 71.73 | 68.11 | 69.87 | 70.53 | 71.43 | -   | 71.82 | 70.12 |
| zh              | 70.24 | 70.68 | 70.65 | 70.12 | 69.91 | 70.29 | 70.47 | 71.11 | 67.47 | 69.33 | 69.50 | 70.29 | 70.54 | -   | 68.90 |
Figure 5: Differences in performance in terms of accuracy scores on the test set for few-shot X-MAML on XNLI using the En-BERT (English) model. Rows correspond to target and columns to auxiliary languages used in X-MAML. Numbers on the off-diagonal indicate performance differences between X-MAML and the baseline model in the same row. The coloring scheme indicates the differences in performance (e.g., blue for large improvement).

<table>
<thead>
<tr>
<th>Auxiliary Language</th>
<th>ar</th>
<th>bg</th>
<th>de</th>
<th>el</th>
<th>en</th>
<th>es</th>
<th>fr</th>
<th>hi</th>
<th>ru</th>
<th>sw</th>
<th>th</th>
<th>tr</th>
<th>ur</th>
<th>vi</th>
<th>zh</th>
</tr>
</thead>
<tbody>
<tr>
<td>ar</td>
<td>0.06</td>
<td>0.64</td>
<td>0.18</td>
<td>-0.19</td>
<td>0.68</td>
<td>0.6</td>
<td>0.09</td>
<td>0.06</td>
<td>0.31</td>
<td>-0.25</td>
<td>0.53</td>
<td>-0.33</td>
<td>0.25</td>
<td>0.08</td>
<td></td>
</tr>
<tr>
<td>bg</td>
<td>0.24</td>
<td>1.3</td>
<td>-0.09</td>
<td>-1.2</td>
<td>1.3</td>
<td>1.5</td>
<td>-0.17</td>
<td>0.93</td>
<td>0.22</td>
<td>-0.42</td>
<td>0.89</td>
<td>-0.13</td>
<td>0.51</td>
<td>0.17</td>
<td></td>
</tr>
<tr>
<td>de</td>
<td>0.94</td>
<td>0.88</td>
<td>0.94</td>
<td>1</td>
<td>1.8</td>
<td>2.1</td>
<td>0.93</td>
<td>0.84</td>
<td>1.3</td>
<td>0.95</td>
<td>1.5</td>
<td>0.58</td>
<td>1.4</td>
<td>1.2</td>
<td></td>
</tr>
<tr>
<td>el</td>
<td>-3.5</td>
<td>-3.4</td>
<td>-2.2</td>
<td>-4</td>
<td>-2.4</td>
<td>-1.7</td>
<td>-3.3</td>
<td>-3.4</td>
<td>-2.7</td>
<td>-3.5</td>
<td>-2.6</td>
<td>-3.5</td>
<td>-2.7</td>
<td>-3.2</td>
<td></td>
</tr>
<tr>
<td>en</td>
<td>-0.39</td>
<td>-0.39</td>
<td>-0.38</td>
<td>-0.33</td>
<td>-0.37</td>
<td>-0.33</td>
<td>-0.36</td>
<td>-0.37</td>
<td>-0.44</td>
<td>-0.34</td>
<td>-0.38</td>
<td>-0.39</td>
<td>-0.43</td>
<td></td>
<td></td>
</tr>
<tr>
<td>es</td>
<td>0.3</td>
<td>0.27</td>
<td>0.37</td>
<td>0.24</td>
<td>0.19</td>
<td>0.35</td>
<td>0.28</td>
<td>0.29</td>
<td>0.05</td>
<td>0.01</td>
<td>0.28</td>
<td>0.28</td>
<td>0.32</td>
<td>0.23</td>
<td></td>
</tr>
<tr>
<td>fr</td>
<td>0.73</td>
<td>0.67</td>
<td>0.88</td>
<td>0.68</td>
<td>0.73</td>
<td>1.1</td>
<td>0.69</td>
<td>0.7</td>
<td>0.71</td>
<td>0.59</td>
<td>0.82</td>
<td>0.68</td>
<td>0.77</td>
<td>0.81</td>
<td></td>
</tr>
<tr>
<td>hi</td>
<td>-1.4</td>
<td>-1.5</td>
<td>-0.82</td>
<td>-1.8</td>
<td>-2.7</td>
<td>-0.92</td>
<td>-0.61</td>
<td>-1.2</td>
<td>-1.6</td>
<td>-2.5</td>
<td>-1.1</td>
<td>-1.7</td>
<td>-1.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ru</td>
<td>-0.22</td>
<td>0.04</td>
<td>0.34</td>
<td>-0.23</td>
<td>-0.63</td>
<td>0.38</td>
<td>0.4</td>
<td>-0.5</td>
<td>-0.26</td>
<td>-0.61</td>
<td>-0.05</td>
<td>-0.41</td>
<td>-0.18</td>
<td>-0.29</td>
<td></td>
</tr>
<tr>
<td>sw</td>
<td>-0.4</td>
<td>-0.61</td>
<td>0.42</td>
<td>-0.57</td>
<td>-1.1</td>
<td>0.28</td>
<td>0.64</td>
<td>-0.68</td>
<td>-0.44</td>
<td>-0.68</td>
<td>-0.28</td>
<td>-0.53</td>
<td>0.13</td>
<td>-0.45</td>
<td></td>
</tr>
<tr>
<td>th</td>
<td>0.25</td>
<td>0.21</td>
<td>0.41</td>
<td>0.28</td>
<td>0.4</td>
<td>0.45</td>
<td>0.4</td>
<td>0.31</td>
<td>0.25</td>
<td>0.47</td>
<td>0.47</td>
<td>0.27</td>
<td>0.44</td>
<td>0.34</td>
<td></td>
</tr>
<tr>
<td>tr</td>
<td>2.7</td>
<td>3.6</td>
<td>3.6</td>
<td>4.4</td>
<td>3.7</td>
<td>2.2</td>
<td>4.2</td>
<td>4.5</td>
<td>3.4</td>
<td>3.9</td>
<td>2.8</td>
<td>3.2</td>
<td>3.8</td>
<td>3.7</td>
<td>3.6</td>
</tr>
<tr>
<td>ur</td>
<td>-0.71</td>
<td>-0.58</td>
<td>-0.29</td>
<td>-0.74</td>
<td>-0.67</td>
<td>-0.3</td>
<td>-0.16</td>
<td>-0.76</td>
<td>-0.64</td>
<td>-0.72</td>
<td>-0.81</td>
<td>-0.84</td>
<td>-0.78</td>
<td>-0.88</td>
<td></td>
</tr>
<tr>
<td>vi</td>
<td>-0.32</td>
<td>-0.41</td>
<td>0.01</td>
<td>-0.33</td>
<td>-0.71</td>
<td>0.02</td>
<td>0.08</td>
<td>-0.44</td>
<td>-0.43</td>
<td>0.01</td>
<td>-0.58</td>
<td>0.03</td>
<td>-0.56</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>zh</td>
<td>74.04</td>
<td>73.97</td>
<td>74.02</td>
<td>74.02</td>
<td>74.01</td>
<td>74.02</td>
<td>74.10</td>
<td>74.11</td>
<td>73.99</td>
<td>74.01</td>
<td>74.06</td>
<td>73.95</td>
<td>-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6: The performance in terms of average test accuracy for the few-shot setting over 10 runs of X-MAML on the XNLI dataset using Multi-BERT (multilingual BERT), as base model. Each column corresponds to the performance of the Multi-BERT system after meta-learning with a single auxiliary language, and evaluation on the target language of the XNLI test set. The auxiliary language is not included during the evaluation phase. Results of the Multi-BERT model without X-MAML (baseline) are also reported.