Zero-Shot Cross-Lingual Transfer with Meta Learning

Farhad Nooralahzadeh
University of Oslo
farhad.nooralahzadeh@uzh.ch

Giannis Bekoulis
Vrije Universiteit Brussel - imec
gbekouli@etrovub.be

Johannes Bjerva
University of Copenhagen / Aalborg University
jbjerva@cs.aau.dk

Isabelle Augenstein
University of Copenhagen
augenstein@di.ku.dk

Abstract

Learning what to share between tasks has become a topic of great importance, 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: 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). A comprehensive error analysis indicates that the correlation of typological features between languages can partly explain when parameter sharing learned via meta-learning is beneficial.

1 Introduction

There are more than 7,000 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 (Joshi et al., 2020). Although there is 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 (Gu et al., 2018; Dou et al., 2019; 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[1], a cross-lingual meta-learning architecture, and study it for two natural language understanding tasks (Natural Language Inference and Question Answering);
- We test X-MAML on 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., 2019) and XLM-RoBERTa (Conneau et al., 2020);

[1]Our code is available at https://github.com/copenlu/X-MAML

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 4547–4562,
November 16–20, 2020.

- 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 idea of meta-learning for transferring zero-shot knowledge in a cross-lingual setting for natural language understanding, in particular for the tasks of NLI and QA. Specifically, we 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 $p(\mathcal{T})$ of tasks $\{\mathcal{T}_1, \mathcal{T}_2, ..., \mathcal{T}_k\}$. The parameters θ of model $\mathtt{M}$ for a particular task $\mathcal{T}_i$, sampled from the distribution $p(\mathcal{T})$, are updated to $\theta_i^{'}$. In particular, the parameters θ are updated using one or a few iterations of gradient descent steps on the training examples (i.e., D_i^{train}) of task $\mathcal{T}_i$. For example, for one gradient update,

$$\theta_i^{'} = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(\mathtt{M}_\theta) \tag{1}$$

where α is the step size, the $\mathtt{M}_\theta$ is the learned model from the neural network and $\mathcal{L}_{\mathcal{T}_i}$ is the loss on the specific task $\mathcal{T}_i$. The parameters of the model θ are trained to optimise the performance of $\mathtt{M}_{\theta_i'}$ on the unseen test examples (i.e., D_i^{test}) across tasks $p(\mathcal{T})$. The meta-learning objective is:

$$\min_\theta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(\mathtt{M}_{\theta_i'}) = \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(\mathtt{M}_{\theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(\mathtt{M}_\theta)}) \tag{2}$$

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 θ parameters of the model using stochastic gradient descent (SGD) as:

$$\theta \leftarrow \theta - \beta \nabla_\theta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(\mathtt{M}_{\theta_i'}) \tag{3}$$

where β 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 a high-resource language $\mathtt{h}$ (i.e., English): Given all training samples in a high-resource language $\mathtt{h}$, we first train the model $\mathtt{M}$ on $\mathtt{h}$ to initialise the model parameters θ.
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 $\mathcal{T}_i$. Then, we update the model parameters using K data points of $\mathcal{T}_i$ (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

Algorithm 1: X-MAML

Input: high-resource language h, set of low-resource languages L,
Model M, step size α and learning rate β

Pre-train M on h and provide initial model parameters θ
Select one or more languages from L as a set of auxiliary languages (A)
while *not done* **do**
  for $l \in$ A **do**
    Sample batch of tasks $\mathcal{T}_i$ using the development set of the auxiliary language l
    for *each* $\mathcal{T}_i$ **do**
      Sample K data-points to form $D_i^{train} = \{(X^k, Y^k)\}_{k=1}^{K} \in \mathcal{T}_i$
      Sample Q data-points to form $D_i^{test} = \{(X^q, Y^q)\}_{q=1}^{Q} \in \mathcal{T}_i$ for meta-update
      Compute $\nabla_\theta \mathcal{L}_{\mathcal{T}_i}(\mathtt{M}_\theta)$ on D_i^{train}
      Compute adapted parameters with gradient descent: $\theta' = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(\mathtt{M}_\theta)$
      Compute $\mathcal{L}_{\mathcal{T}_i}(\mathtt{M}_{\theta'})$ using D_i^{test}
  Update $\theta \leftarrow \theta - \beta \nabla_\theta \sum_i \mathcal{L}_{\mathcal{T}_i}(\mathtt{M}_{\theta'})$
Perform either (i) zero-shot or (ii) few-shot learning on $\{\mathtt{L} \setminus \mathtt{A}\}$ using meta-learned parameters θ

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 the meta-objective function and to perform a meta-update using Eq. (3). This step is performed in multiple iterations.

3. Zero-shot or few-shot learning on the target languages: In the last step of X-MAML, we first initialise the model parameters with those learned during meta-learning. We then continue by evaluating the model on the test set of the target languages (i.e., zero-shot learning) or fine-tuning the model parameters with standard supervised learning using the development set of the target languages and evaluate on the test set (i.e., few-shot learning).

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

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 $\mathcal{T}_i$ 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.

(2020) 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 implement X-MAML using the *higher* library[2]. 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 α and learning rate β to $1e-4$ and $1e-5$, 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. The results are reported for each experiment by averaging the performance over ten different runs. 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.

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 using Hugging Face's library (Devlin et al., 2019) 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 for more information) in X-MAML for XNLI, we fine-tune En-BERT and Multi-BERT on the MultiNLI dataset (English) to obtain the initial model parameters θ for each experiment.

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., 2020), 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 θ (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).

[2] https://github.com/facebookresearch/higher

x%	Baseline	MAML	
		$\mathcal{T}_{Train}$	$\mathcal{T}_{Dev}$
1	38.60	49.78	**50.92**
2	37.80	48.58	**50.66**
3	47.09	51.40	**52.85**
5	49.88	**52.22**	51.40
10	51.02	52.51	**53.95**
20	59.14	**61.38**	58.16
50	63.37	**63.85**	61.74
100	64.35	**64.99**	64.61

Table 1: Test accuracies with different settings of MAML on MultiNLI. x%: the percentage of training samples. **Baseline:** The test accuracy of trained ESIM using x% of training data. **MAML:** The test accuracy of ESIM after meta-learning, where $\mathcal{T}_{Train}$: 5 tasks are defined in MAML using the training set, and $\mathcal{T}_{Dev}$: 10 tasks are included in MAML using the development set. Bold font indicates best results for the various proportions of the used training data.

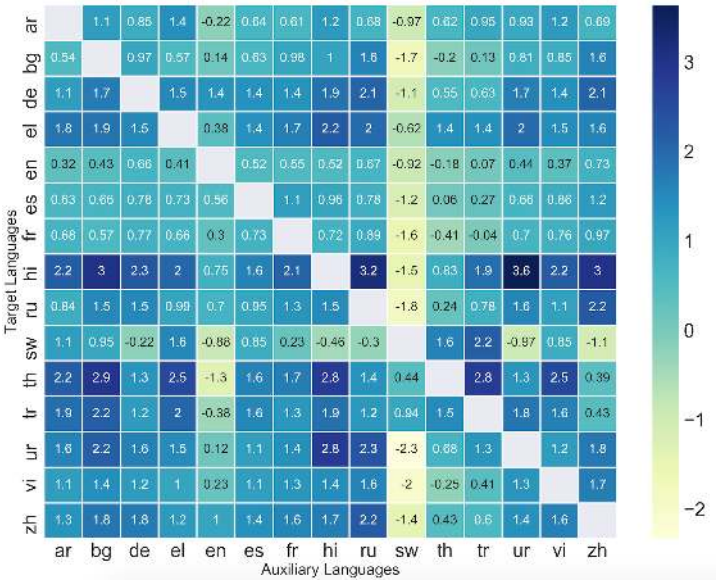

Figure 1: Differences in performance in terms of accuracy scores on the test set for zero-shot X-MAML on XNLI using the Multi-BERT model. Rows correspond to target and columns to auxiliary languages used in X-MAML. Numbers on the off-diagonal indicate performance differences between X-MAML and the baseline model in the same row. The coloring scheme indicates the differences in performance (e.g., blue for large improvement).

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) improve 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 after 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. (2019). 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. (2019)[3] and Wu and Dredze (2019). We report the average and maximum performance by using one auxiliary language for each target language. We also report

[3] https://github.com/google-research/bert/blob/master/multilingual.md

	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	avg
Zero-shot cross-lingual transfer																
Devlin et al. (2019)	81.4	-	74.3	70.5	-	-	-	-	62.1	-	-	63.8	-	-	58.35	-
Wu and Dredze (2019)	82.1	73.8	74.3	71.1	66.4	68.9	69.0	61.6	64.9	69.5	55.8	69.3	60.0	50.4	58.0	66.3
Multi-BERT (Our baseline)	81.36	73.45	73.85	69.74	65.73	67.82	67.94	59.04	64.63	70.12	52.46	68.90	58.56	47.58	58.70	65.33
X-MAML *(One aux. lang.)*																
AVG	81.69	73.86	74.43	71.00	67.16	68.39	68.90	60.41	65.33	70.95	54.08	70.09	60.51	47.97	59.94	-
MAX	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	67.33
$hi \rightarrow X$	81.88	74.17	74.81	71.59	67.95	68.86	69.44	60.93	65.86	71.57	55.26	70.59	-	47.12	61.51	-
X-MAML *(Two aux. lang.)*	*(hi,de)*	*(hi,ar)*	*(fr,de)*	*(bg,zh)*	*(ur,ru)*	*(hi,ru)*	*(de,bg)*	*(ur,sw)*	*(el,tr)*	*(de,bg)*	*(bg,tr)*	*(ru,el)*	*(ur,ru)*	*(el,tr)*	*(hi,de)*	
$(l_1, l_2) \rightarrow X$	**82.59**	**75.69**	**75.97**	**73.45**	**69.16**	**71.42**	**71.44**	**62.57**	**67.19**	**72.63**	**62.57**	**73.13**	**63.53**	**50.42**	**62.93**	**68.98**
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	69.65
X-MAML *(One aux. lang.)*																
AVG	82.22	75.24	76.06	73.34	69.97	71.80	71.28	64.76	67.82	73.41	61.57	74.02	64.83	58.02	63.66	-
MAX	82.39	75.32	76.18	73.46	70.03	71.94	71.45	64.92	67.95	73.52	61.74	74.21	64.97	58.23	63.81	70.01
$sw \rightarrow X$	82.24	75.31	75.94	73.34	69.98	71.77	71.31	64.89	67.87	73.38	61.5	73.99	64.94	-	63.63	-
X-MAML *(Two aux. lang.)*	*(ar,ru)*	*(ru,th)*	*(ru,th)*	*(el,hi)*	*(sw,vi)*	*(ar,zh)*	*(de,tr)*	*(es,sw)*	*(bg,hi)*	*(bg,ru)*	*(el,vi)*	*(ar,th)*	*(sw,vi)*	*(ar,tr)*	*(en,ru)*	
$(l_1, l_2) \rightarrow X$	**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**	70.69
Machine translate at test (TRANSLATE-TEST)																
Devlin et al. (2019)	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	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. (2019) 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. (l_1, l_2) 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., 2019), 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.

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 results from Devlin et al. (2019) 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 F_1 score compared to XLM-15, XLM-R_{base} and XLM-R_{large}, 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 F_1 score on zero-shot QA. More details can be found in the Appendix (Section A.1).

5 Related Work

The main motivation for this work is the low availability of labelled training datasets for most of the world's languages. To alleviate this issue, a number of methods, including the so-called few-shot learning approaches have been proposed. Few-shot learning methods have initially been introduced within the area of image classification (Vinyals et al., 2016; Ravi and Larochelle, 2017; Finn et al., 2017), but have recently also been applied to NLP tasks such as relation extraction (Han et al., 2018), text classification (Yu et al., 2018; Rethmeier and Augenstein, 2020) and machine translation (Gu et al., 2018). Specifically, in NLP, these few-shot learning approaches include: (i) the transformation of the problem into a different task (e.g., relation extraction is transformed to question answering (Levy et al., 2017; Abdou et al., 2019)); or (ii) meta-learning (Andrychowicz et al., 2016; Finn et al., 2017).

Meta-Learning: Meta-learning or learning-to-learn has recently received a lot of attention from the NLP community. First-order MAML has been applied to the task of machine translation (Gu et al., 2018), where they propose to use meta-learning for improving the machine translation performance for low-resource languages by learning to adapt to target languages based on multilingual high-resource languages. However, in the proposed framework, they include 18 high-resource languages as auxiliary languages and five diverse low-resource languages as target languages. In our work, we assume access to only English as a high-resource language. For the task of dialogue generation, Qian and Yu (2019) address domain adaptation using meta-learning. Dou et al. (2019) explore MAML variants thereof for low-resource NLU tasks in the GLUE dataset (Wang et al., 2018). They consider different high-resource NLU tasks such as MultiNLI (Williams et al., 2018) and QNLI (Rajpurkar et al., 2016) as auxiliary tasks to learn meta-parameters using MAML. They then fine-tune the low-resource tasks using the adapted parameters from the meta-learning phase. All the above-mentioned works on meta-learning in NLP assume that there are multiple high-resource tasks or languages, which are then adapted to new target tasks or languages with a handful of training samples. However, in a cross-lingual NLI and QA setting, the available high-resource language is usually only English. Our work thus fills an important gap in the literature, as we only require a single source language.

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, morphological inflection and parsing. In contrast to these tasks, which have seen much cross-lingual attention (Plank et al., 2016; Bjerva, 2017; Kementchedjhieva et al., 2018; 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 (Agic and Schluter, 2018; Conneau et al., 2018; Zhao et al., 2020), and to a lesser degree on RE (Faruqui and Kumar, 2015; Verga et al., 2016) and QA (Abdou et al., 2019; Lewis et al., 2020). 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.

		Model	en	ar	de	es	hi	vi	zh	avg
XLM		Our baseline	**69.80**	48.95	52.64	**58.15**	46.67	48.46	42.64	52.47
	X-MAML	*(One aux. lang.)* $l \rightarrow X$	69.39 *ar*	48.45 *hi*	53.04 *es*	57.68 *en*	46.90 *zh*	49.79 *zh*	44.36 *hi*	52.80
	X-MAML	*(Two aux. lang.)* $(l_1, l_2) \rightarrow X$	68.88 *(es,ar)*	**49.76** *(vi,zh)*	**53.18** *(vi,zh)*	58.00 *(en,zh)*	**48.43** *(vi,zh)*	**50.86** *(hi,zh)*	**45.44** *(es,hi)*	**53.51**
XLM-R$_{base}$		Liang et al. (2020)	80.1	56.4	62.1	67.9	60.5	67.1	61.4	65.1
		Our baseline	**80.38**	57.23	63.08	67.91	61.46	67.14	62.73	65.70
	X-MAML	*(One aux. lang.)* $l \rightarrow X$	80.19 *vi*	57.97 *hi*	63.57 *ar*	67.46 *vi*	61.70 *vi*	67.97 *hi*	64.01 *hi*	66.12
	X-MAML	*(Two aux. lang.)* $(l_1, l_2) \rightarrow X$	80.31 *(ar,vi)*	**58.14** *(hi,vi)*	**64.07** *(ar,hi)*	**68.08** *(ar,hi)*	**62.67** *(es,ar)*	**68.82** *(ar,hi)*	**64.06** *(ar,hi)*	**66.59**
XLM-R$_{large}$		Hu et al. (2020)	83.5	66.6	70.1	74.1	70.6	74	62.1	71.6
		Our baseline	83.95	66.09	70.62	74.59	70.64	74.13	69.80	72.83
	X-MAML	*(One aux. lang.)* $l \rightarrow X$	84.31 *ar*	66.61 *hi*	70.84 *ar*	74.32 *hi*	**70.94** *vi*	**74.84** *ar*	**70.74** *hi*	73.23
	X-MAML	*(Two aux. lang.)* $(l_1, l_2) \rightarrow X$	**84.60** *(hi,vi)*	**66.95** *(hi,vi)*	**71.00** *(ar,vi)*	**74.62** *(en,vi)*	70.93 *(ar,vi)*	74.73 *(es,hi)*	70.29 *(en,vi)*	**74.30**

Table 3: F_1 scores (average over 10 runs) on the MLQA test set using zero-shot X-MAML. Columns indicate the target languages. The avg column indicates row-wise average F_1 score. We also report the most beneficial auxiliary language(s) for X-MAML in improving the test F_1 of each target language.

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; Bjerva et al., 2019a,b,c; Zhao et al., 2020) 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.

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

Mostafa Abdou, Cezar Sas, Rahul Aralikatte, Isabelle Augenstein, and Anders Søgaard. 2019. X-WikiRE: A Large, Multilingual Resource for Relation Extraction as Machine Comprehension. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 265–274, Hong Kong, China. Association for Computational Linguistics.

Zeljko Agic and Natalie Schluter. 2018. Baselines and Test Data for Cross-Lingual Inference. In *LREC*. European Language Resources Association (ELRA).

Marcin Andrychowicz, Misha Denil, Sergio Gomez Colmenarejo, Matthew W. Hoffman, David Pfau, Tom Schaul, and Nando de Freitas. 2016. Learning to learn by gradient descent by gradient descent. In *Advances in Neural Information Processing Systems*, pages 3981–3989.

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

Johannes Bjerva. 2017. *One Model to Rule them all – Multitask and Multilingual Modelling for Lexical Analysis*. Ph.D. thesis, University of Groningen.

Johannes Bjerva and Isabelle Augenstein. 2018a. From Phonology to Syntax: Unsupervised Linguistic Typology at Different Levels with Language Embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 907–916, New Orleans, Louisiana. Association for Computational Linguistics.

Johannes Bjerva and Isabelle Augenstein. 2018b. Tracking Typological Traits of Uralic Languages in Distributed Language Representations. In *Proceedings of the Fourth International Workshop on Computational Linguistics of Uralic Languages*, pages 76–86, Helsinki, Finland. Association for Computational Linguistics.

Johannes Bjerva, Yova Kementchedjhieva, Ryan Cotterell, and Isabelle Augenstein. 2019a. A Probabilistic Generative Model of Linguistic Typology. 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 1529–1540, Minneapolis, Minnesota. Association for Computational Linguistics.

Johannes Bjerva, Yova Kementchedjhieva, Ryan Cotterell, and Isabelle Augenstein. 2019b. Uncovering Probabilistic Implications in Typological Knowledge Bases. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3924–3930, Florence, Italy. Association for Computational Linguistics.

Johannes Bjerva, Robert Östling, Maria Han Veiga, Jörg Tiedemann, and Isabelle Augenstein. 2019c. What Do Language Representations Really Represent? *Computational Linguistics*, 45(2):381–389.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, and Hui Jiang. 2016. Enhancing and Combining Sequential and Tree LSTM for Natural Language Inference. *CoRR*, abs/1609.06038.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual Language Model Pretraining. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 7057–7067.

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. 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.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2019. Ethnologue: Languages of the World. `https://www.ethnologue.com/statistics/size`. Accessed: 2019-05-25.

Manaal Faruqui and Shankar Kumar. 2015. Multilingual Open Relation Extraction Using Cross-lingual Projection. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1351–1356. 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-Volume 70*, pages 1126–1135. JMLR. org.

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.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. FewRel: A Large-Scale Supervised Few-Shot Relation Classification Dataset with State-of-the-Art Evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4803–4809.

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 Generalization. *CoRR*, abs/2003.11080.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Yova Kementchedjhieva, Johannes Bjerva, and Isabelle Augenstein. 2018. Copenhagen at CoNLL–SIGMORPHON 2018: Multilingual Inflection in Context with Explicit Morphosyntactic Decoding. In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 93–98, Brussels. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*.

Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. 2015. Siamese Neural Networks for One-shot Image 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 Empirical Methods in Natural Language Processing*, pages 4243–4252, Brussels, Belgium. Association for Computational 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 Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating Cross-lingual Extractive Question Answering. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Miryam de Lhoneux, Johannes Bjerva, Isabelle Augenstein, and Anders Søgaard. 2018. Parameter sharing between dependency parsers for related languages. 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, Shuguang Liu, Fan Yang, and Ming Zhou. 2020. XGLUE: A New Benchmark Dataset for Cross-lingual Pre-training, Understanding and Generation. *ArXiv*, abs/2004.01401.

Alex Nichol, Joshua Achiam, and John Schulman. 2018. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*.

Barbara Plank, Anders Søgaard, and Yoav Goldberg. 2016. Multilingual Part-of-Speech Tagging with Bidirectional Long Short-Term Memory Models and Auxiliary Loss. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 412–418.

Kun Qian and Zhou Yu. 2019. Domain Adaptive Dialog Generation via Meta Learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2639–2649, Florence, Italy. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100, 000+ Questions 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 (ICLR)*. OpenReview.net.

Nils Rethmeier and Isabelle Augenstein. 2020. Long-Tail Zero and Few-Shot Learning via Contrastive Pretraining on and for Small Data. *arXiv preprint arXiv:2010.01061*.

Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy P. Lillicrap. 2016. Meta-Learning with Memory-Augmented Neural Networks. In *ICML*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 1842–1850. JMLR.org.

Patrick Verga, David Belanger, Emma Strubell, Benjamin Roth, and Andrew McCallum. 2016. Multilingual Relation Extraction using Compositional Universal Schema. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 886–896. Association for Computational Linguistics.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, and Daan Wierstra. 2016. Matching Networks for One Shot Learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3630–3638. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

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 Classification with Multiple Metrics. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1206–1215.

Wei Zhao, Steffen Eger, Johannes Bjerva, and Isabelle Augenstein. 2020. Inducing Language-Agnostic Multilingual Representations. *arXiv preprint arXiv:2008.09112*.

A Appendices

A.1 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), a new dataset (X-WikiRE) is introduced for multilingual QA-based relation extraction in five languages (i.e., English, French, Spanish, Italian and German) by Abdou et al. (2019). 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 this 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))[4] 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` is 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/nusnlp/namanda

		Auxiliary language				Baseline	Abdou et al. (2019)			
		es	fr	it	de		BERT		fastText	
zero-shot	es	-	49.01	**50.11**	**50.59**	49.85	5.49		16.17	
	fr	**52.20**	-	**52.13**	**51.96**	51.72	17.42		15.28	
	it	50.53	**50.65**	-	50.58	50.58	10.70		4.44	
	de	**49.92**	48.78	48.63	-	48.98	2.87		14.09	
		1k					1k	10k	1k	10k
few-shot	es	-	78.09	**78.33**	77.89	78.26	42.97	71.66	65.78	77.99
	fr	**80.68**	-	**80.81**	**80.74**	80.67	42.69	72.43	65.67	74.15
	it	**82.04**	81.76	-	81.77	81.78	56.25	80.06	64.02	**83.45**
	de	78.29	78.48	**78.66**	-	78.63	56.01	70.43	62.47	72.17

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.

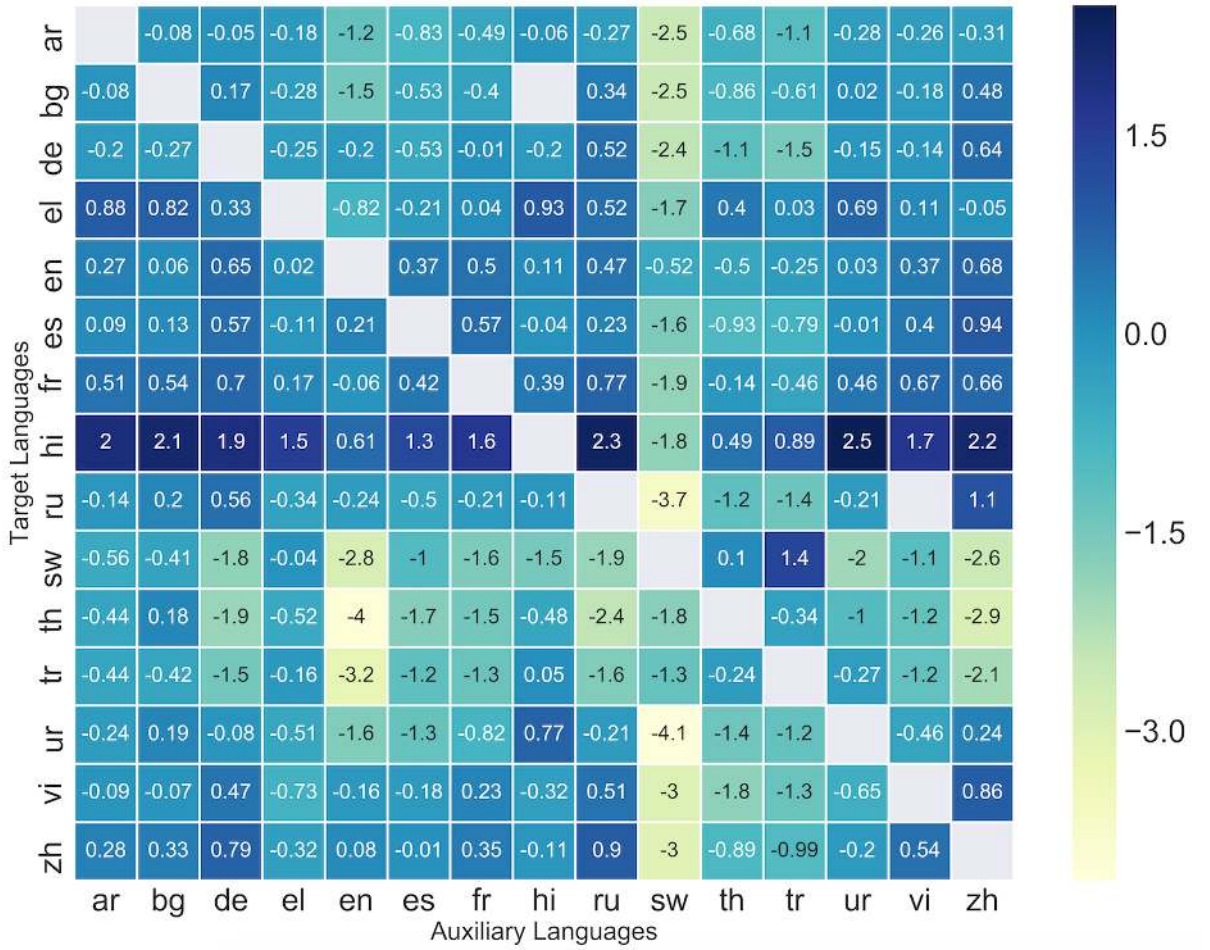

Figure 2: Differences in performance in terms of accuracy scores on the test set for the zero-shot case using training (without meta-learning) on XNLI with the Multi-BERT model. Rows correspond to target and columns to auxiliary languages. Numbers on the off-diagonal indicate performance differences 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).

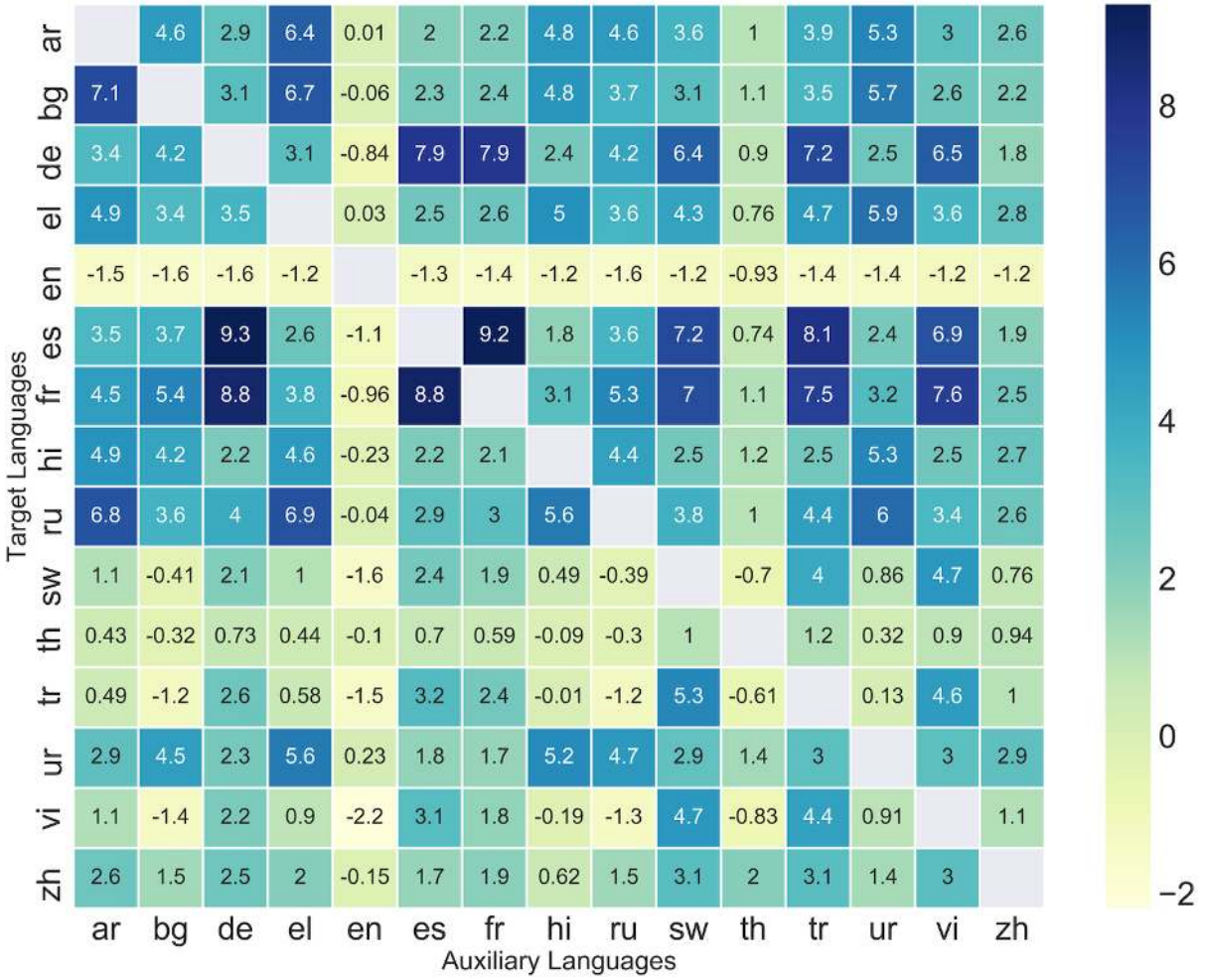

Figure 3: Differences in performance in terms of accuracy scores on the test set for zero-shot X-MAML on XNLI using the En-BERT (English) model. Rows correspond to target and columns to auxiliary languages used in X-MAML. Numbers on the off-diagonal indicate performance differences between X-MAML and the baseline model in the same row. The coloring scheme indicates the differences in performance (e.g., blue for large improvement).

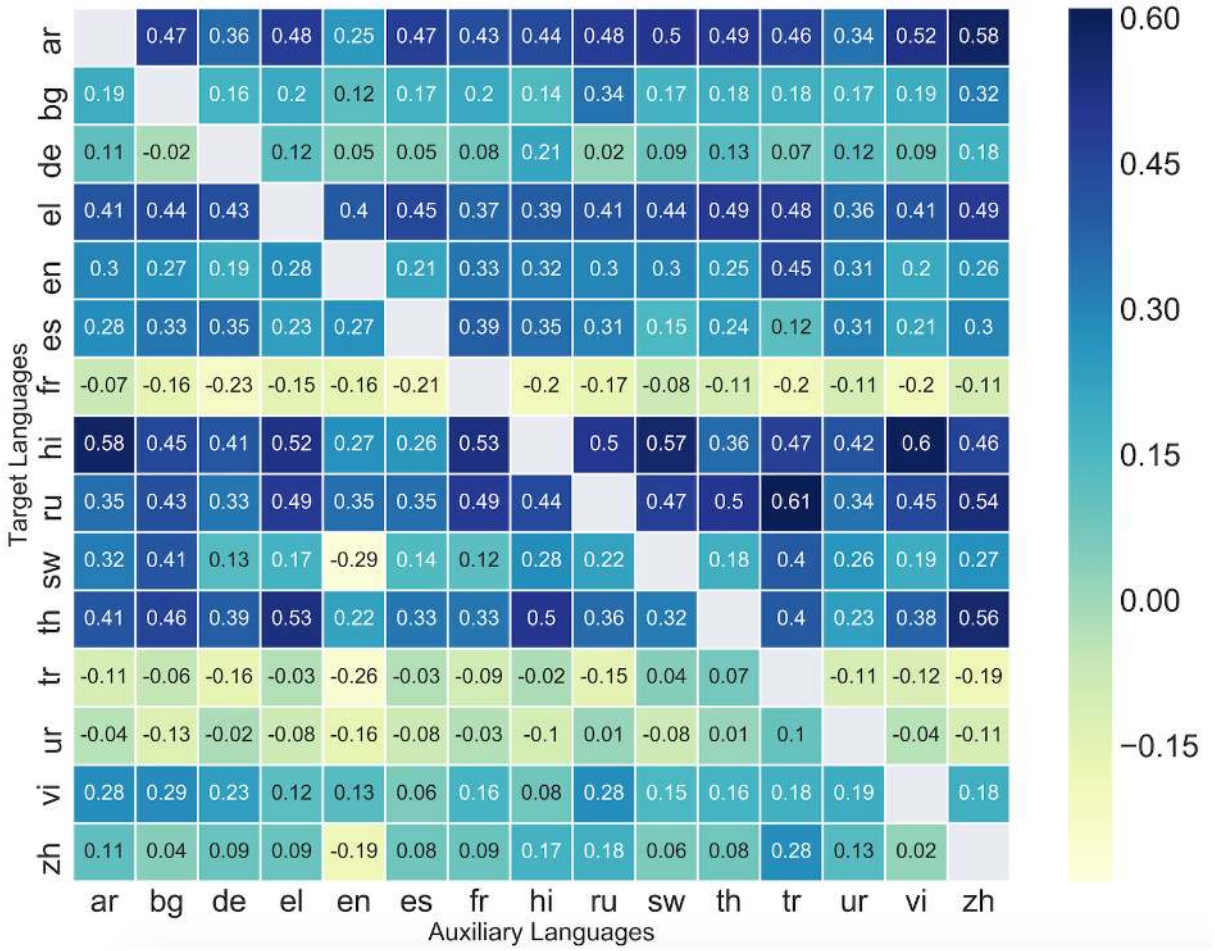

Figure 4: Differences in performance in terms of accuracy scores on the test set for few-shot X-MAML on XNLI using the Multi-BERT model. Rows correspond to target and columns to auxiliary languages used in X-MAML. Numbers on the off-diagonal indicate performance differences between X-MAML and the baseline model in the same row. The coloring scheme indicates the differences in performance (e.g., blue for large improvement).

	Auxiliary language															baseline
	ar	**bg**	**de**	**el**	**en**	**es**	**fr**	**hi**	**ru**	**sw**	**th**	**tr**	**ur**	**vi**	**zh**	
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	67.95	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	67.34	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	71.14	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	70.59	71.11	67.47	69.33	69.50	70.29	70.54	-	68.90

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.

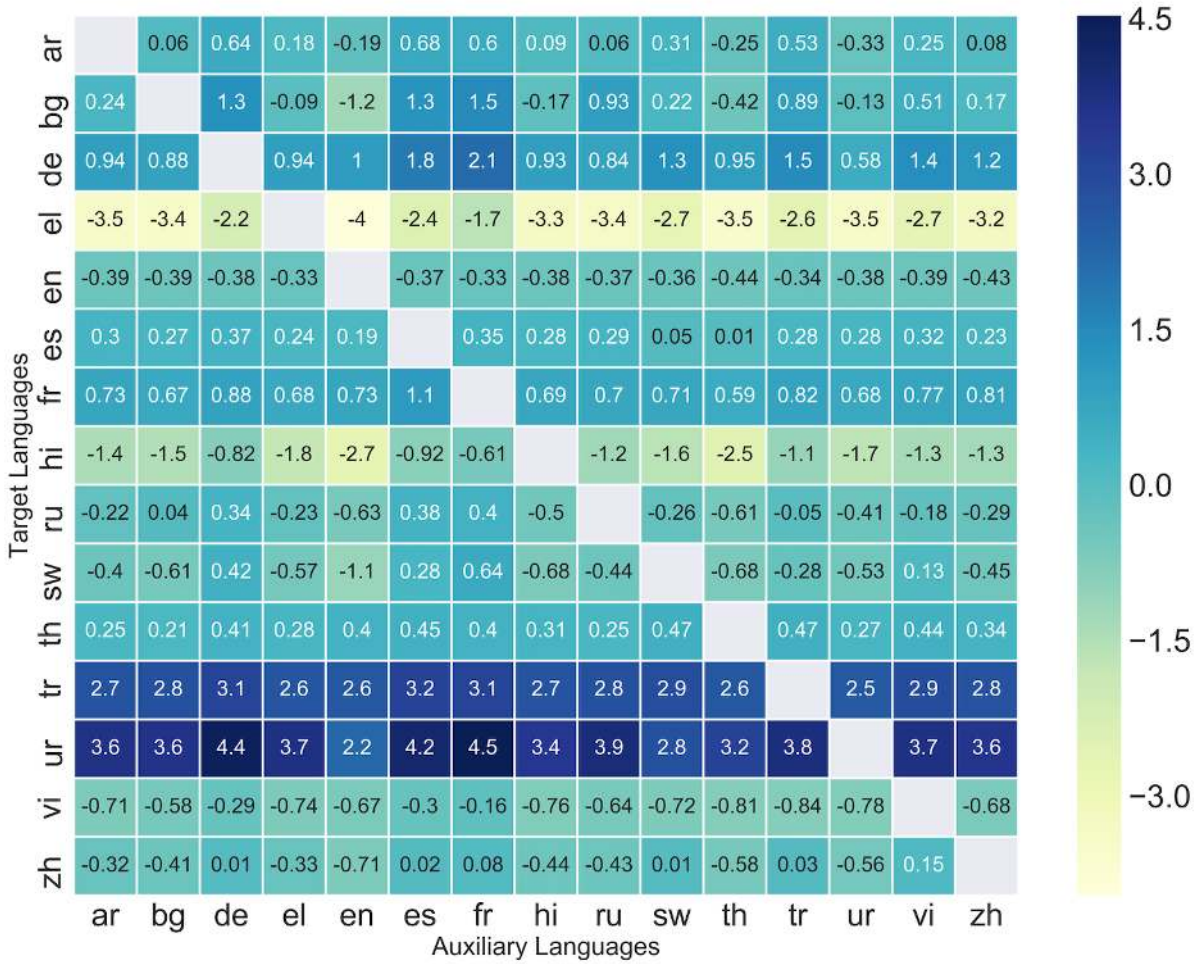

Figure 5: Differences in performance in terms of accuracy scores on the test set for few-shot X-MAML on XNLI using the En-BERT (English) model. Rows correspond to target and columns to auxiliary languages used in X-MAML. Numbers on the off-diagonal indicate performance differences between X-MAML and the baseline model in the same row. The coloring scheme indicates the differences in performance (e.g., blue for large improvement).

	Auxiliary language															baseline
	ar	**bg**	**de**	**el**	**en**	**es**	**fr**	**hi**	**ru**	**sw**	**th**	**tr**	**ur**	**vi**	**zh**	
ar	-	67.84	67.73	67.85	67.62	67.84	67.80	67.81	67.85	67.87	67.86	67.83	67.71	67.89	67.95	67.37
bg	71.79	-	71.76	71.80	71.72	71.77	71.80	71.74	71.94	71.77	71.78	71.78	71.77	71.79	71.92	71.60
de	73.36	73.23	-	73.37	73.30	73.30	73.33	73.46	73.27	73.34	73.38	73.32	73.37	73.34	73.43	73.25
el	69.95	69.98	69.97	-	69.94	69.99	69.91	69.93	69.95	69.98	70.03	70.02	69.90	69.95	70.03	69.54
en	82.24	82.21	82.13	82.22	-	82.15	82.27	82.26	82.24	82.24	82.19	82.39	82.25	82.14	82.20	81.94
es	76.07	76.12	76.14	76.02	76.06	-	76.18	76.14	76.10	75.94	76.03	75.91	76.10	76.00	76.09	75.79
fr	75.32	75.23	75.16	75.24	75.23	75.18	-	75.19	75.22	75.31	75.28	75.19	75.28	75.19	75.28	75.39
hi	64.95	64.82	64.78	64.89	64.64	64.63	64.90	-	64.87	64.94	64.73	64.84	64.79	64.97	64.83	64.37
ru	71.19	71.27	71.17	71.33	71.19	71.19	71.33	71.28	-	71.31	71.34	71.45	71.18	71.29	71.38	70.84
sw	58.14	58.23	57.95	57.99	57.53	57.97	57.94	58.10	58.04	-	58.00	58.22	58.08	58.01	58.09	57.82
th	61.59	61.64	61.57	61.71	61.40	61.51	61.51	61.68	61.54	61.50	-	61.58	61.41	61.56	61.74	61.18
tr	64.74	64.79	64.69	64.82	64.59	64.82	64.76	64.83	64.70	64.89	64.92	-	64.74	64.73	64.66	64.85
ur	63.67	63.58	63.69	63.63	63.55	63.63	63.68	63.61	63.72	63.63	63.72	63.81	-	63.67	63.60	63.71
vi	73.51	73.52	73.46	73.35	73.36	73.29	73.39	73.31	73.51	73.38	73.39	73.41	73.42	-	73.41	73.23
zh	74.04	73.97	74.02	74.02	73.74	74.01	74.02	74.10	74.11	73.99	74.01	74.21	74.06	73.95	-	73.93

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.