Enhancing Cross-Lingual Training with Knowledge Learned from Multi-Lingual Training

Anonymous ACL submission

Abstract

On multi-lingual natural language processing (NLP) tasks, it is generally agreed that multi-lingual models perform better than cross-lingual models even with limited training data in the target languages. Though this is expected, its cause has not been well-studied. In this paper, we examine the differences between cross- and multi-lingual models fine-tuned on syntactic, semantic, or sentiment analysis (SA) tasks, from the perspectives of parameter updates, feature extraction, and domain changes to investigate the advantage of multi-lingual training. Additionally, we incorporate the knowledge we learn from our analyses into the training process of cross-lingual models to improve their performance. Results show that jointly applying feature augmentation and domain adaptation approaches effectively improves the performance of the vanilla cross-lingual models, with average F1-macro score improvements from 0.38% to 20.75% on four NLP tasks. Our studies indicate cross-lingual training effectiveness could be enhanced without requiring additional labeled data in the target languages. This provides an alternative choice to data augmentation for future research on resource-scarce languages.

1 Introduction

Two common settings for training a machine learning model on a multi-lingual NLP task are: 1) training the model on a (source) language and evaluating it on another (target) language, and 2) training the model on both the source and target languages and evaluating it on the target language. We refer to the former as cross-lingual training and the latter as multi-lingual training. While cross-lingual training features better extensibility to truly resource-scarce target languages, multi-lingual models outperform cross-lingual models in most cases (Liang et al., 2020; Hu et al., 2020). However, it remains unknown what factors lead to the superior performance of multi-lingual models. To uncover the secrets of multi-lingual training and improve cross-lingual models, we examine the differences between cross- and multi-lingual models trained on four syntactic, semantic, and SA tasks. We use multi-lingual BERT (mBERT, Devlin et al. (2019)), one of the top-performing multi-lingual NLP models, in these experiments and analyses. Our analyses show two main differences between cross- and multi-lingual models.

First, through model probing and attention-head analyses on the four tasks, we find that different linguistic features are emphasized by cross- and multi-lingual models fine-tuned on the same task. This potentially results from the fact that the importance of attention heads are not uniformly distributed for cross- and multi-lingual models, as each attention head extracts a relatively stable set of linguistic features (Michel et al., 2019). For example, attention heads in the middle to upper layers of mBERT are updated more intensely by a multi-lingual paraphrase identification (PI) model than its cross-lingual counterpart in our experiments, potentially suggesting that critical syntactic and semantic knowledge in target languages is learned via multi-lingual training (Tenney et al., 2019; Vig and Belinkov, 2019). Our model probing and attention-head probing experiments on the four NLP tasks provide additional evidence for the different importance ranking of linguistic features between cross- and multi-lingual models. For example, our experimental results suggest that key linguistic features for a PI task are emphasized more heavily by a multi-lingual SA model than a cross-lingual SA model fine-tuned on the same dataset.

Second, knowledge about the target language domains potentially contributes to the higher performance of multi-lingual models. Through language modeling (LM) evaluations, we find that multi-lingual models produce pseudo perplexity scores (Salazar et al., 2020) that are 11.45% to 90.43% lower than cross-lingual models fine-tuned

using the same datasets on the training corpora of all the languages. This shows the substantially lower encoding ability of the cross-lingual model on documents not in the source language, which is a potential cause of the performance gap between the cross- and multi-lingual models in our evaluations.

Our findings suggest that cross-lingual models function differently than multi-lingual models from both feature and domain perspectives. Taking advantage of these findings, we design three approaches to improve cross-lingual models without introducing additional task-oriented labeled training instances. The approaches include: 1) fusing linguistic features that are important for the multi-lingual models into cross-lingual training via joint modeling, 2) adapting the cross-lingual models to the target language domain via two-stage training, and 3) combining the two approaches into a training pipeline. Our evaluations on four NLP tasks show that while the first and second approaches generally bring positive effects to the performance of cross-lingual models, the combination of both approaches results in the best average performance on all the tasks (0.38% to 20.75% higher performance than the vanilla cross-lingual models).

The contributions of this paper are two-fold:

- We develop a systematic understanding of the differences between cross- and multi-lingual models via model interpretation.
- We augment cross-lingual training with knowledge learned from these differences to improve the performance of cross-lingual models without additional labeled data in the target languages.

2 Tasks and Datasets

To examine the differences between cross- and multi-lingual models, we evaluate and analyze mBERT models on four tasks, including two syntactic tasks (part-of-speech tagging (POS) and named entity recognition (NER)), one semantic task (PI), and one SA task. We use datasets in four languages (English (EN), German (DE), Spanish (ES), and French (FR)) in our experiments, where EN is the source language for training the cross-lingual models. Table 1 displays the number of training and test instances in these datasets. The tasks and datasets are:

Universal Dependencies (UD) [1] provides annotated datasets for grammar-related tasks (e.g., POS and dependency parsing) in over 100 languages. We use POS in our experiments since the objective and evaluation metric of dependency parsing are very different from those of the other tasks we use, which can lead to potential problems when comparing models fine-tuned on different tasks.

WikiANN (Rahimi et al., 2019) is an NER dataset constructed over Wikipedia documents. Different from POS, language-specific lexical features are important for NER models since the named entity (NE) expressions are language-specific. The WikiANN dataset is annotated with three NE labels, i.e., LOC (location), PER (person), and ORG (organization), on word level in the IOB-2 format.

PAWS-X (Yang et al., 2019) is used in our experiments to study the feature extraction patterns of models fine-tuned on semantic tasks. Each instance in PAWS-X is a pair of sentences in the same language which is labeled for whether the two sentences express the same semantic meaning (1) or not (0). The instances in PAWS-X are sampled from the PAWS-WIKI dataset (Zhang et al., 2019) in English and translated into other languages.

Multilingual Amazon Reviews Corpus (MARC) (Keung et al., 2020) is an SA dataset containing documents from Amazon product reviews. It labels each review with a 5-scale sentiment label based on the stars (rating) it receives from its author.

Metric: We evaluate the F1-macro score for all the datasets to avoid bias caused by imbalanced

	EN	DE	ES	FR
UD Training	12,543	13,814	14,035	14,449
UD Test	2,077	977	1,721	416
WikiANN Training	20,000	20,000	20,000	20,000
WikiANN Test	10,000	10,000	10,000	10,000
PAWS-X Training	49,401	49,401	49,401	49,401
PAWS-X Test	2,000	2,000	2,000	2,000
MARC Training	200,000	200,000	200,000	200,000
MARC Test	5,000	5,000	5,000	5,000

Table 1: Number of training and test instances in the datasets we use in our evaluations and analyses.

[1] https://universaldependencies.org/

	EN	DE	ES	FR
POS-cross-ling	**93.85**	69.73	60.07	62.91
POS-multi-ling	93.07	**74.92**	**86.29**	**75.70**
NER-cross-ling	92.10	83.28	75.28	83.35
NER-multi-ling	**92.60**	**94.12**	**94.76**	**94.51**
PI-cross-ling	94.22	85.77	87.34	86.03
PI-multi-ling	**94.52**	**89.27**	**91.30**	**91.45**
SA-cross-ling	57.63	43.42	45.22	44.70
SA-multi-ling	**57.75**	**60.28**	**57.75**	**57.05**

Table 2: F1-macro score of mBERT models fine-tuned on the POS (UD), NER (WikiANN), PI (PAWS-X), and SA (MARC) tasks. The higher performance on each data set is in bold.

	POS	NER	PI	SA
POS-cross-ling	-	3.61	3.39	4.80
POS-multi-ling	-	0.05	2.79	1.96
NER-cross-ling	-1.14	-	1.00	0.76
NER-multi-ling	-0.14	-	2.22	-2.36
PI-cross-ling	0.81	3.09	-	2.97
PI-multi-ling	1.61	3.69	-	-4.37
SA-cross-ling	-0.40	1.83	0.73	-
SA-multi-ling	-0.29	5.02	4.96	-

Table 3: Probing results of cross- and multi-lingual mBERT models fine-tuned on the POS (UD), NER (WikiANN), PI (PAWS-X), and SA (MARC) tasks. The results are in F1-macro score.

label distributions.

3 Experiments and Analyses

This section examines the differences between cross- and multi-lingual mBERT models fine-tuned on each of the POS, NER, PI, and SA tasks from three perspectives. Section 3.1 presents the evaluation performance of both models to confirm the advantage of multi-lingual training. Section 3.2 examines the different linguistic features emphasized by the cross- and multi-lingual models via probing, and Section 3.3 conducts more in-depth studies about the feature extraction patterns of both models through attention-head analyses. Section 3.4 evaluates the domain compatibility of both models with the domains of the test sets for each task to examine the importance of domain knowledge for multi-lingual tasks. For clarity, we refer to the cross- and multi-lingual models on each task as [TASK]-cross-ling and [TASK]-multi-ling, respectively, where [TASK] is the task on which the models are fine-tuned. Specifically, we apply the Huggingface (Wolf et al., 2020) implementation of the pre-trained bert-base-multilingual-cased model with 12 layers and 12 attention heads per layer. All the models are fine-tuned for five epochs with a learning rate of 1e-5 and a batch size of 32.

3.1 Evaluations

The evaluation results of cross- and multi-lingual mBERT models on the POS, NER, PI, and SA tasks are displayed in Table 2. We note that the multi-lingual model outperforms the cross-lingual model in almost all the experiments except for the EN experiment on POS, demonstrating the advantage of multi-lingual training.

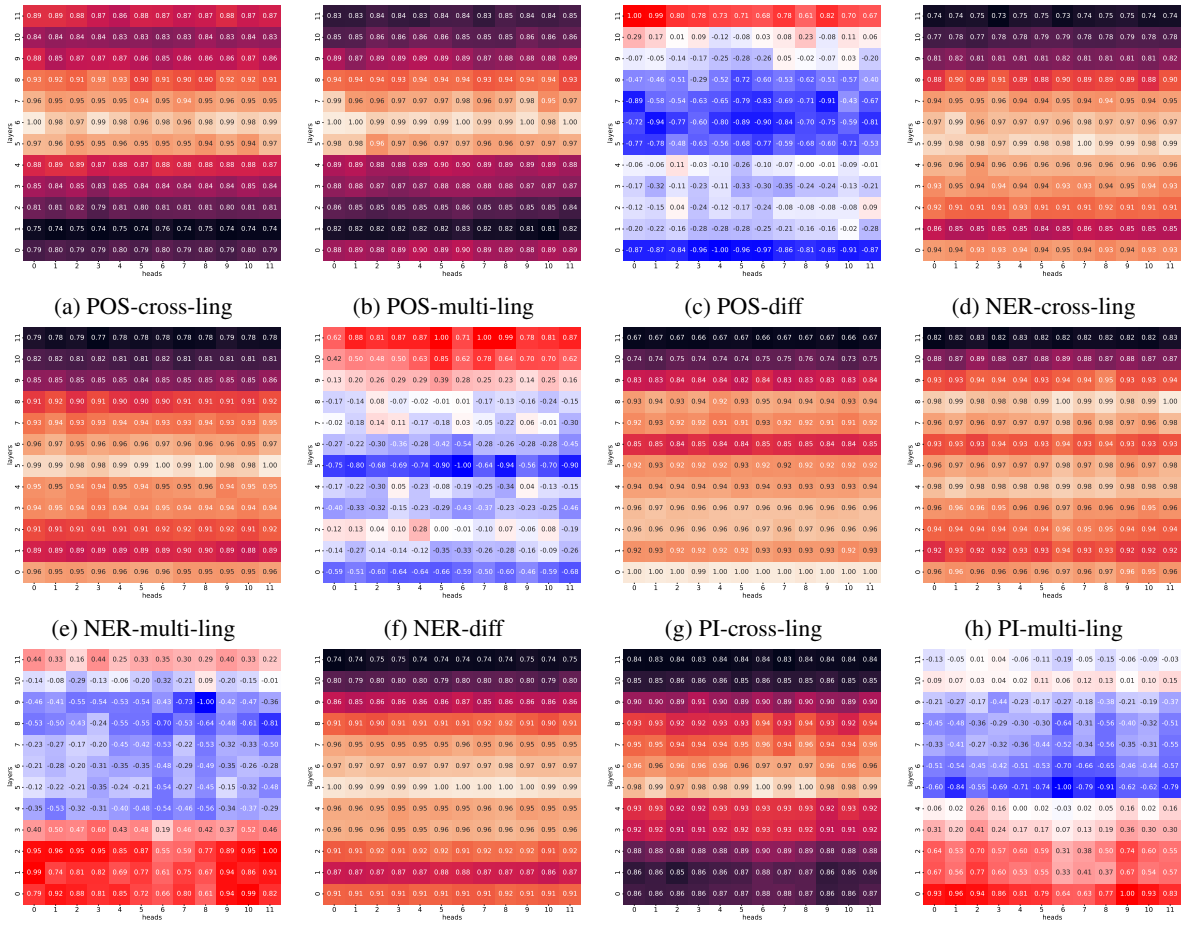
Additionally, we find that though multi-lingual training does not introduce additional training instances in the source language (i.e., EN), the performance of multi-lingual models is higher than the cross-lingual models in the EN evaluations on NER, PI, and SA tasks. On the other hand, the performance gains brought by multi-lingual training are imbalanced on the four languages (e.g., the performance gain is 3.50 for DE but 5.42 for FR on the PI task) even on the PAWS-X dataset which is constructed on a parallel corpus. These findings suggest that the amount of training data in each language is not the only factor affecting the performance of cross- or multi-lingual models. Instead, we hypothesize that linguistic features and domain knowledge could also be important for mBERT models to perform well on multi-lingual tasks. To verify our hypothesis and explain the advantage of multi-lingual training, we conduct probing and domain compatibility experiments in the rest of this section.

3.2 Model Probing

To examine important linguistic features underlying the predictions of cross- and multi-lingual models, we probe the 8 [TASK]-cross-ling and [TASK]-multi-ling models on the POS, NER, PI, and SA tasks. Similar to Wu et al. (2021), we probe each model in four steps:

1) We construct a prediction head on top of a pre-trained mBERT model and train the model on a probing task with the mBERT weights frozen to get the “probe" model.

2) We construct a prediction head on top of a fine-tuned mBERT model (e.g., POS-cross-ling) and train the model on the same probing task with the



(a) POS-cross-ling (b) POS-multi-ling (c) POS-diff (d) NER-cross-ling

(e) NER-multi-ling (f) NER-diff (g) PI-cross-ling (h) PI-multi-ling

(i) PI-diff (j) SA-cross-ling (k) SA-multi-ling (l) SA-diff

Figure 1: Normalized attention-head updates of [TASK]-cross-ling and [TASK]-multi-ling and the difference between the two matrices for each [TASK] ([TASK]-diff) among POS, NER, PI, and SA. For attention-head update figures (captioned [TASK]-cross-ling and [TASK]-multi-ling), brighter colors indicate more intense updates. In [TASK]-diff figures, cells in blue and red colors represent attention heads updated more heavily by the [TASK]-multi-ling and [TASK]-cross-ling, respectively.

mBERT weights frozen to get the “ceiling" model. 3) We evaluate both models on the probing task and subtract the evaluation score of the “probe" model from the “ceiling" model to get the probing result for the fine-tuned model on the probing task.

All these models are trained on the EN training sets and evaluated on the combination of test sets in four languages (i.e., the cross-lingual setting) to avoid discrepancies between mBERT weights and prediction heads in cross-lingual models.

We display the probing results in Table 3. From the results, we find that NER-multi-ling gets noticeably higher probing results than NER-cross-ling on POS and PI. We speculate the cause of this phenomenon to be that NER-multi-ling learns to put more emphasis on extracting important linguistic features for these two tasks. Similarly, the extraction of important features for POS and NER may be contributing to the higher performance of NER-multi-ling than NER-cross-ling, and SA-multi-ling may have benefited from emphasizing critical features for NER and PI. While POS-multi-ling outperforms POS-cross-ling in the evaluations, its probing results are lower than those of POS-cross-ling on NER, PI, and SA. Two possible causes of this may be that POS-multi-ling learns to emphasize linguistic features that are not critical for the NER, PI, and SA tasks, or that other types of information (e.g., domain knowledge) mainly account for the superior performance of POS-multi-ling over POS-cross-ling. These results show that there are key linguistic features that are emphasized by multi-lingual models but not by cross-lingual models fine-tuned on the same datasets. Our probing experiments help specify the differences in feature extraction behaviors between cross- and multi-lingual

models, which is shown by our later experiments to be useful for improving the performance of cross-lingual models.

3.3 Attention Head Analyses

As attention heads are the feature extraction units in a Transformer-based model (Vaswani et al., 2017), we examine the updates and roles of attention heads to illustrate differences between cross- and multi-lingual models in the probing experiments.

3.3.1 Attention Head Updates

The absolute weight updates in each attention head reflect its importance in the training process of the model. We plot the normalized absolute attention-head updates and the differences between the updates of each pair of cross- and multi-lingual models in Figure 1.

We find that the attention-head updates correlate strongly between the [TASK]-cross-ling and [TASK]-multi-ling models, with Spearman's rank correlation coefficient (Spearman's ρ) above 0.70 for all the tasks. [2] Meanwhile, the most heavily updated attention heads lie in different areas of the model for different tasks. For example, the middle layers of mBERT models fine-tuned on POS and NER are updated the most intensely, which may result from the heavy dependence of POS and NER models on syntactic features. Furthermore, the NER models update lower-layer attention heads more heavily than the highest-layer attention heads, and so do the PI and SA models, suggesting that semantic and lexical features are important for these tasks (see Kovaleva et al. (2019); Tenney et al. (2019); Vig and Belinkov (2019)).

In addition, we find that though the attention-head updates are ranked similarly between cross- and multi-lingual models, the extent to which each attention head is updated differ substantially. Specifically, the multi-lingual POS and NER models distribute more emphasis on the middle syntactic layers while the cross-lingual models update the upper-most task-specific layers more intensely. We infer that additional syntactic features are learned by POS-multi-ling and NER-multi-ling on the multi-lingual training data, which help them perform better in the non-EN evaluations. Moreover, both PI-multi-ling and SA-multi-ling place more emphasis on upper-layer semantic heads while PI-cross-ling and SA-cross-ling update the lexical attention heads on lower layers more heavily. Though lexical features are important for PI and SA, these features are language-dependent and may have caused the performance differences between cross- and multi-lingual models shown in Table 2. These findings show that attention heads are weighed differently in the cross- and multi-lingual training processes, which lends additional support to our assumption that cross- and multi-lingual models emphasize different sets of linguistic features.

3.3.2 Attention Head Probing

To gain a better understanding of the feature-extraction roles of each attention head, we probe each attention head in the 8 [TASK]-cross-ling and [TASK]-multi-ling models on the POS, NER, PI, and SA tasks using the same probing method as introduced in Section 3.2. On each task, we compare the attention-head probing results of all the models with those of the cross-lingual model fine-tuned specifically for that task, e.g., POS-cross-ling for POS. For clarity, we refer to the model with which all the other models are compared on each task as [TASK]-baseline. Specifically, we examine the number of overlaps among the topK contributive attention heads and Spearman's ρ of the overlapped heads between each [TASK]-cross-ling or [TASK]-multi-ling model and each [TASK]-baseline model. We assume that the most contributive attention heads in a model play important roles in the extraction of critical linguistic features for the task where the model is fine-tuned. As such, these analyses could help us identify how differently cross- and multi-lingual models fine-tuned on the same task extract linguistic features. Note that this assumption is not strict, and the method cannot be used as a quantitative metric for evaluating feature dependence of these models, though it helps us qualitatively explain the performance of the models from the perspective of linguistic feature extraction.

As Table 4 shows, each multi-lingual model shares a more consistent attention-head ranking with the cross-lingual model fine-tuned on the same task than with other cross-lingual models. We speculate the reason to be that both cross- and multi-lingual models fine-tuned on the same task learn to put higher emphasis on a core linguistic feature set for that task. The different attention-head rankings between each pair of cross- and multi-lingual

[2] All these Spearman's ρ are statistically significant with p-values lower than 0.01.

Models	TopK	POS		NER		PI		SA	
		Ovlp	Corr	Ovlp	Corr	Ovlp	Corr	Ovlp	Corr
POS-cross-ling	40	-	-	18	0.42	21	0.27	12	0.51
	60	-	-	32	0.50*	35	0.55*	24	0.33*
	80	-	-	50	0.66*	52	0.74*	44	0.51*
POS-multi-ling	40	20	0.22	15	0.19	14	0.54*	12	0.23
	60	36	0.61*	37	0.17	29	0.62*	24	0.47*
	80	55	0.42*	56	0.51*	46	0.78*	42	0.57*
NER-cross-ling	40	14	0.00	-	-	18	-0.30	17	0.59*
	60	34	0.04	-	-	29	0.09	27	0.65*
	80	52	0.22	-	-	49	0.70*	50	0.59*
NER-multi-ling	40	16	0.16	27	0.46*	21	-0.32	18	0.23
	60	33	0.31	41	0.65*	29	0.29	31	0.33
	80	53	0.30*	52	0.47*	49	0.75*	50	0.65*
PI-cross-ling	40	15	-0.17	11	0.05	-	-	14	0.44
	60	33	0.17	27	-0.10	-	-	24	0.44*
	80	49	0.24	48	0.32*	-	-	49	0.54*
PI-multi-ling	40	14	0.56*	17	-0.18	23	0.44*	12	0.39
	60	32	0.13	28	0.53*	38	0.71*	23	0.45*
	80	50	0.24	52	0.47*	57	0.82*	45	0.62*
SA-cross-ling	40	14	-0.04	12	-0.17	21	0.09	-	-
	60	28	0.30	29	-0.12	32	0.20	-	-
	80	49	0.25	50	0.41*	49	0.59*	-	-
SA-multi-ling	40	18	-0.42	13	0.45	21	0.23	26	0.73*
	60	31	0.07	27	0.30	29	0.13	43	0.80*
	80	45	0.13	49	0.15	50	0.30*	64	0.69*

Table 4: Number of overlapped attention heads (Ovlp) and Spearman's ρ of the rankings of these overlapped heads (Corr) among the TopK contributive heads (TopK) between [TASK]-cross-ling or [TASK]-multi-ling models and [TASK]-baseline models on the POS, NER, PI, and SA tasks. Statistically significant ρ are marked with *.

models further reflect the different weightings of linguistic features in cross- and multi-lingual fine-tuning processes on each task.

Additionally, we compare the attention-head probing results of multi-lingual models and their cross-lingual counterparts to help explain the model probing results in Table 3. We provide the full attention-head probing results in Appendix A. For POS, we find that POS-cross-ling has more overlapped attention heads with NER-baseline and PI-baseline than POS-multi-ling, especially among the top-40 and top-60 contributive heads. The rank correlations of the overlapped heads are also substantially higher for POS-cross-ling than POS-multi-ling in most of the cases. On the other hand, though POS-cross-ling and POS-multi-ling have the same amount of overlapped attention heads with SA-baseline among the top-40 contributive heads, the rank correlation is higher for POS-cross-ling. These results are consistent with the higher probing performance of POS-cross-ling than POS-multi-ling on all the three tasks. For NER, we note that the attention-head rankings of NER-multi-ling are more consistent with POS-baseline and PI-baseline than NER-cross-ling, which explains the higher model probing results of NER-multi-ling on POS and PI. Similarly, attention-head rankings of PI-multi-ling are more consistent with POS-baseline and NER-baseline than PI-cross-ling, and those of SA-multi-ling are more consistent with NER-baseline and PI-baseline. These all correspond to cases where the probing results of the multi-lingual models are noticeably higher than those of their cross-lingual counterparts. In other cases where the model probing results of the cross-lingual models are higher, the higher attention-head overlap or rank correlations are also reflected in the attention-head probing results. These findings support our hypothesis that multi-lingual training provides additional knowledge about feature weight-

	EN	DE	ES	FR
	POS			
POS-cross-ling	694.92	6237.60	3123.90	2636.49
POS-multi-ling	615.34	596.71	191.01	202.78
mBERT	23.34	10.64	6.30	5.85
	NER			
NER-cross-ling	616.47	4613.01	4943.07	4252.96
NER-multi-ling	517.44	515.25	326.17	323.52
mBERT	13.60	15.50	14.02	14.91
	PI			
PI-cross-ling	3193.42	2753.90	3318.81	3824.42
PI-multi-ling	239.94	104.55	166.35	114.31
mBERT	10.43	9.66	7.06	12.35
	SA			
SA-cross-ling	5559.11	6458.82	4157.43	4764.42
SA-multi-ling	613.33	731.77	436.07	305.81
mBERT	33.76	25.39	27.92	25.77

Table 5: Pseudo perplexities achieved by cross- and multi-lingual mBERT models on the EN, DE, ES, and FR test sets of the POS (UD), NER (WikiANN), PI (PAWS-X), and SA (MARC) datasets. mBERT represents the vanilla mBERT model.

Models	Languages			
	EN	DE	ES	FR
POS-cross-ling	**93.85**	**69.73**	60.07	62.91
+FA	-	-	-	-
+DA	91.60	69.54	**60.45**	**63.32**
+combined	-	-	-	-
NER-cross-ling	92.10	83.28	75.28	83.35
+FA	93.53	91.81	75.31	83.77
+DA	92.18	89.23	89.76	92.29
+combined	**93.83**	**93.65**	**92.50**	**94.83**
PI-cross-ling	94.22	**85.77**	87.34	86.03
+FA	93.75	85.20	87.05	86.80
+DA	86.09	77.16	76.83	78.28
+combined	**94.24**	85.46	**87.91**	**87.06**
SA-cross-ling	57.63	43.42	45.22	44.70
+FA	58.04	43.94	44.94	45.14
+DA	58.35	58.07	54.58	53.70
+combined	**58.86**	**59.71**	**55.37**	**54.04**

Table 6: Evaluation performance (in F1-macro score) of cross-lingual models fine-tuned on POS (UD), NER (WikiANN), PI (PAWS-X), and SA (MARC) tasks ([TASK]-cross-ling) and cross-lingual models augmented with feature augmentation (+FA), domain adaptation (+DA), and both (+combined) approaches. Highest performance on each dataset is in bold.

ing, which as we show later, can be leveraged to improve the performance of cross-lingual models.

3.4 Domain Compatibility

Since the cross- and multi-lingual models are fine-tuned on documents in different languages, their text domain compatibility with the test corpora (in four languages for each task) may also differ. This could be a potential cause of the performance differences between cross- and multi-lingual models. To investigate, we evaluate the pseudo perplexity of each model on the four test corpora of its own training dataset and display the results in Table 5.

We note that all the fine-tuned models produce higher pseudo perplexities than the vanilla mBERT model on these test corpora, since the models are fine-tuned with the classification or sequence labeling objectives which are not directly related to the masked language modeling (MLM) objective. However, the pseudo perplexities produced by the cross-lingual models are much (12.93% to 945.33%) higher than those produced by the corresponding multi-lingual models on all the non-EN corpora. Meanwhile, the pseudo perplexity differences are noticeably higher for POS models than for NER or PI models, which are consistent with the greater performance gaps between POS-cross-ling and POS-multi-ling on non-EN test data, as Table 2 shows. These findings demonstrate that textual domain knowledge specific to each language, e.g., vocabulary and expressions, is an unignorable difference between cross- and multi-lingual models which potentially affects their performance.

4 Enhancing Cross-Lingual Models

From our probing and attention-head analyses, we find that cross- and multi-lingual models differ substantially in their feature extraction behaviors and the domain knowledge they learn. This motivates us to examine practical ways of improving the performance of cross-lingual models by emphasizing linguistic features or domain knowledge, without requiring additional labeled training data in target languages.

4.1 Feature Augmentation

As our independent experiments in Sections 3.2 and 3.3 reach the same conclusion that cross- and multi-lingual models put different weights on linguistic features, we hypothesize that augmenting cross-lingual models with proper external feature sets could improve their performance. To verify this hypothesis, we conduct multi-task learning (MTL) experiments on these tasks, with the auxiliary tasks chosen based on the probing results. Specifically, we use POS and PI as auxiliary tasks for NER, POS and NER as auxiliary tasks for PI, and NER and

PI as auxiliary tasks for SA. As POS-cross-ling always achieves higher probing results than POS-multi-ling in our experiments, we do not apply the feature augmentation method on POS-cross-ling. Only the EN training data of both the primary and auxiliary tasks is used for the feature augmentation approach to avoid information leakage.

As Table 6 shows, feature augmentation helps improve the performance of cross-lingual models in 8 out of 12 experiments. These results suggest that the weightings of linguistic features have an effect on the performance of mBERT models, and that feature augmentation is potentially a practical approach for improving the performance of cross-lingual models. For the PI-cross-ling model, however, feature augmentation mainly results in negative effects on its performance. This possibly results from the different task objectives of PI (classification) and its auxiliary tasks (sequence labeling). Choosing auxiliary tasks from a broader set of NLP tasks may help relieve this problem.

4.2 Domain Adaptation

Since we find domain compatibility to be a potential cause of the performance differences between cross-lingual and multi-lingual models fine-tuned on POS, NER, PI, and SA, we examine a domain adaptation approach to help improve the performance of the cross-lingual models. Specifically, we first fine-tune a pre-trained mBERT model on the multi-lingual training corpus of a task using the MLM objective and then re-fine-tune the model on the EN training dataset using the classification or sequence labeling objective.

According to Table 6, the domain adaptation approach provides noticeable performance gains to the NER and SA tasks. This potentially results from the more important role of lexical features in target languages for the two tasks. However, domain adaptation harms the performance of POS-cross-ling on EN and DE, and that of PI-cross-ling on all the languages. The negative effects on the performance of POS-cross-ling may have resulted from the lower importance of lexical features to POS, as discussed in Section 3.3.1. We speculate the cause of the low performance of PI-cross-ling to be the mismatch of input format between the MLM fine-tuning stage and the cross-lingual fine-tuning stage for PI. Since we break the premise and hypothesis sentences in each PI instance down to two parts, the LM capability of the PI-cross-ling model potentially decreases on the PAWS-X corpora, which could lead to the worse performance of the model.

4.3 Joint Knowledge Enhancement

As feature augmentation and domain adaptation have been shown to be effective in improving the performance of cross-lingual models on the POS, NER, and SA tasks, we also examine whether the two approaches can be applied in combination. Specifically, we first fine-tune an mBERT model on the multi-lingual training data of a task using the MLM objective and then fine-tune the trained model jointly with the auxiliary tasks we choose. We do not apply joint knowledge enhancement on POS-cross-ling since we cannot choose a proper auxiliary task set for POS based on our probing experiments. As Table 6 shows, the combined approach leads to the highest performance improvements to NER-cross-ling and SA-cross-ling. For PI-cross-ling whose performance is harmed by separately applying the feature augmentation and domain adaptation methods, the combined approach is able to compensate for the negative effects and slightly improve the performance of PI-cross-ling on the EN, ES, and FR test sets. One possible explanation of the higher effectiveness of the joint knowledge enhancement approach is that domain adaptation helps mBERT better generalize the knowledge it learns from cross-lingual training and feature augmentation to other languages, which boosts the effectiveness of the feature augmentation approach.

5 Conclusion and Future Work

Our analyses of the differences between cross- and multi-lingual mBERT models fine-tuned on various NLP tasks demonstrate that two key factors impact the higher performance of multi-lingual models: the weightings of features and domain compatibility with target languages. Based on these findings, we design two approaches to improve the performance of cross-lingual models, i.e., feature augmentation and domain adaptation. Evaluations on four NLP tasks show that these two approaches, either used individually or in combination, generally have positive effects on the performance of cross-lingual models without additional task-specific annotations in target languages.

Future work can extend the scope of this paper to include NLP tasks of other types, e.g., natural language generation.

6 Ethics Statement & Broader Impact

All the datasets used in this paper are publicly available to the entire NLP community, and we adopt the official data annotations and splits in all our evaluations and analyses. The datasets do not contain sensitive or identifiable information about the authors or annotators. In addition, the mBERT model we use in the experiments is implemented and made publicly available by Huggingface. We do not foresee any ethical issue in this paper. However, we should note that large-scale pre-trained language models such as mBERT have been shown to be biased. This should be taken into consideration when utilizing such models for real-world applications.

The work presented in this paper has a broader impact of improving the performance of cross-lingual NLP models on truly resource-scarce languages without the need for acquiring additional annotated data, which can be expensive. This can potentially broaden the application of pre-trained NLP models to a wider range of languages.

References

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.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

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

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4365–4374, Hong Kong, China. Association for Computational Linguistics.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 14014–14024.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. Association for Computational Linguistics.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

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

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76, Florence, Italy. Association for Computational Linguistics.

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

Demonstrations, pages 38–45, Online. Association for Computational Linguistics.

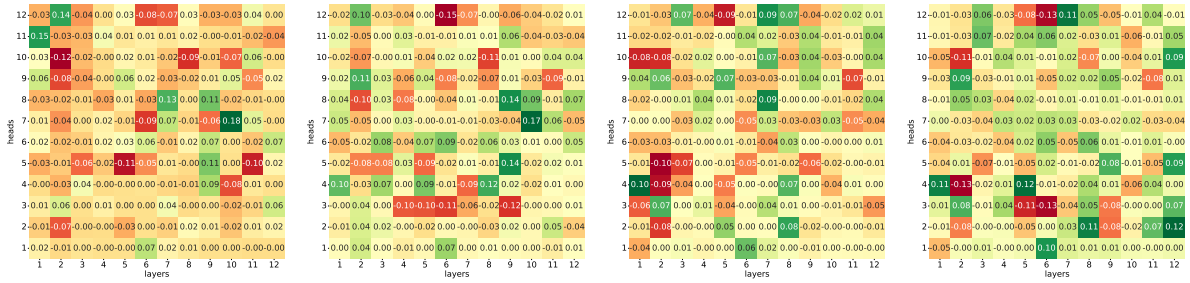
Zhaofeng Wu, Hao Peng, and Noah A. Smith. 2021. Infusing Finetuning with Semantic Dependencies. *Transactions of the Association for Computational Linguistics*, 9:226–242.

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

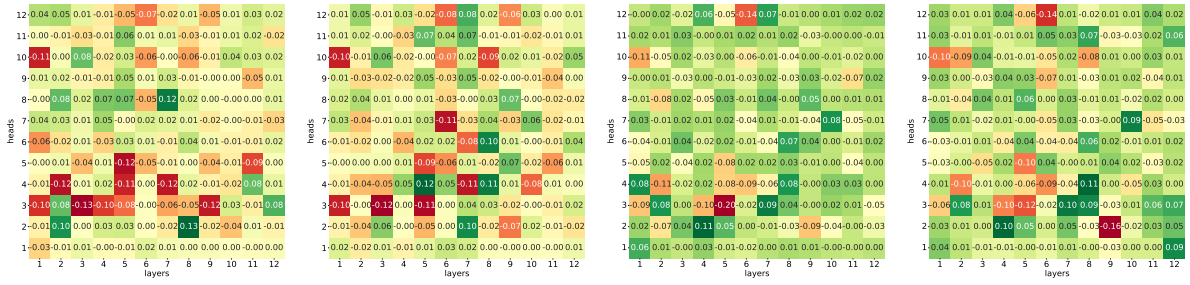
Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

A Attention Head Probing Results

We display the attention-head probing results on POS in Figure A1, the results on NER in Figure A2, the results on PI in Figure A3, and those on SA in Figure A4. For each attention-head probing experiment, we train the probing models on the EN training data and evaluate on the test data of all the four languages (i.e., the cross-lingual setting). In the figures, the most contributive attention heads in each [TASK]-cross-ling or [TASK]-multi-ling model for the target task are marked in green, and the least important attention heads are marked in red. From the figures, we find that the most contributive attention heads for each task heavily overlap across models, e.g., the 8-th attention head on the 7-th layer of mBERT for the POS task. This implies that each task or dataset relies on a set of foundational linguistic features.

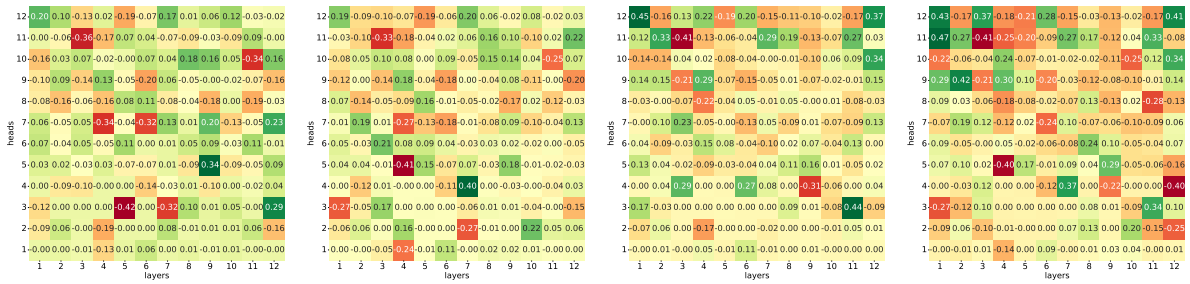


(a) POS-cross-ling (b) POS-multi-ling (c) NER-cross-ling (d) NER-multi-ling

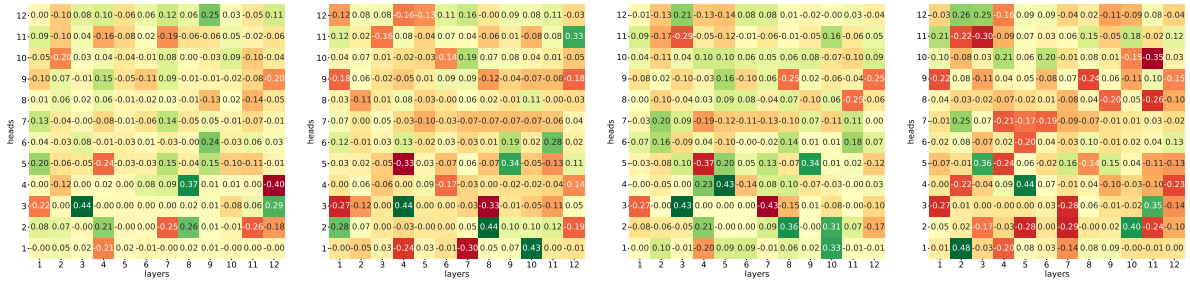


(e) PI-cross-ling (f) PI-multi-ling (g) SA-cross-ling (h) SA-multi-ling

Figure A1: The probing results of the POS-cross-ling, POS-multi-ling, NER-cross-ling, NER-multi-ling, PI-cross-ling, PI-multi-ling, SA-cross-ling, and SA-multi-ling on the POS task (the UD dataset). The most contributive attention heads are marked in green, and the least contributive heads are marked in red.

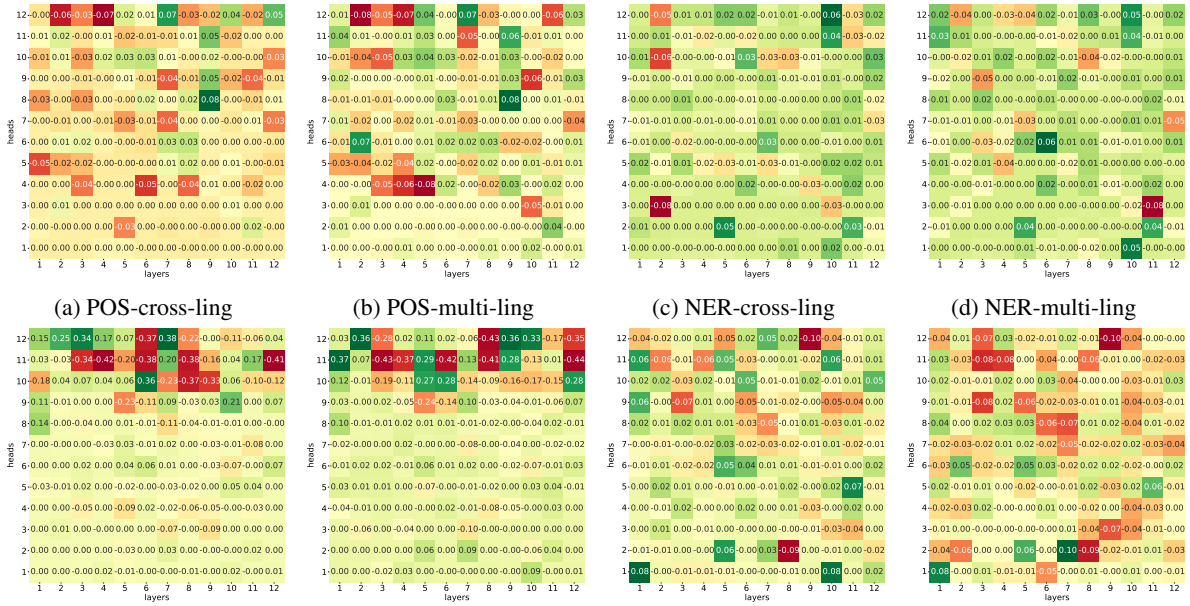


(a) POS-cross-ling (b) POS-multi-ling (c) NER-cross-ling (d) NER-multi-ling



(e) PI-cross-ling (f) PI-multi-ling (g) SA-cross-ling (h) SA-multi-ling

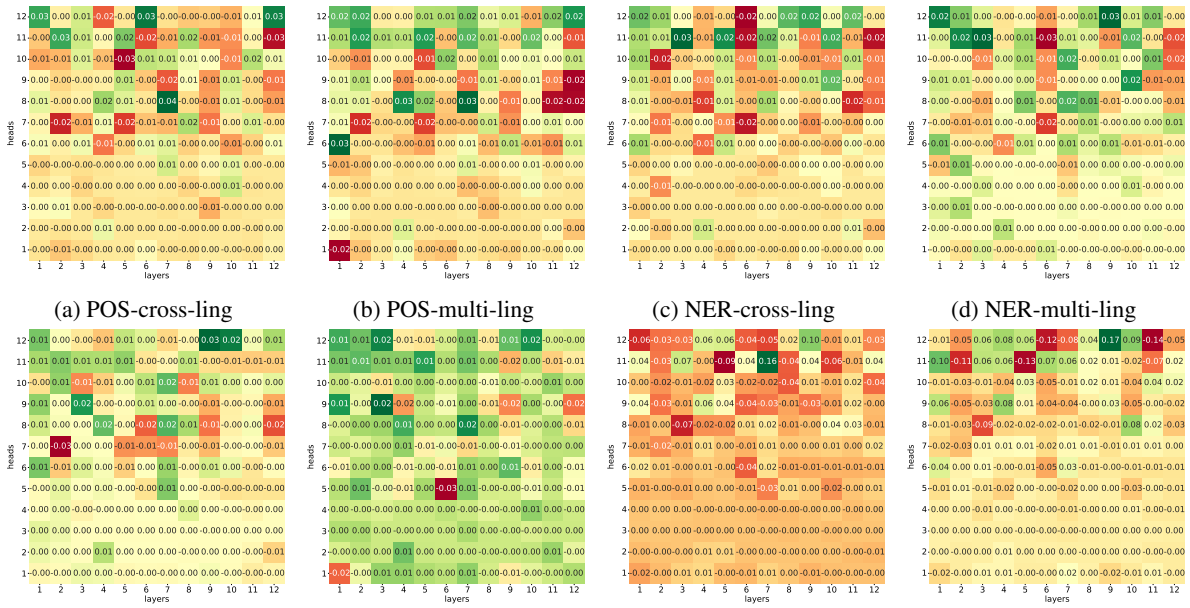
Figure A2: The probing results of the POS-cross-ling, POS-multi-ling, NER-cross-ling, NER-multi-ling, PI-cross-ling, PI-multi-ling, SA-cross-ling, and SA-multi-ling on the NER task (the WikiANN dataset). The most contributive attention heads are marked in green, and the least contributive heads are marked in red.



(a) POS-cross-ling (b) POS-multi-ling (c) NER-cross-ling (d) NER-multi-ling

(e) PI-cross-ling (f) PI-multi-ling (g) SA-cross-ling (h) SA-multi-ling

Figure A3: The probing results of the POS-cross-ling, POS-multi-ling, NER-cross-ling, NER-multi-ling, PI-cross-ling, PI-multi-ling, SA-cross-ling, and SA-multi-ling on the PI task (the PAWS-X dataset). The most contributive attention heads are marked in green, and the least contributive heads are marked in red.



(a) POS-cross-ling (b) POS-multi-ling (c) NER-cross-ling (d) NER-multi-ling

(e) PI-cross-ling (f) PI-multi-ling (g) SA-cross-ling (h) SA-multi-ling

Figure A4: The probing results of the POS-cross-ling, POS-multi-ling, NER-cross-ling, NER-multi-ling, PI-cross-ling, PI-multi-ling, SA-cross-ling, and SA-multi-ling on the SA task (the MARC dataset). The most contributive attention heads are marked in green, and the least contributive heads are marked in red.