Contributions of Transformer Attention Heads in Multi- and Cross-lingual Tasks

Weicheng Ma[1*], Kai Zhang[2*†], Renze Lou[3†], Lili Wang[1], and Soroush Vosoughi[4]

[1,4]Department of Computer Science, Dartmouth College
[2]Department of Computer Science and Technology, Tsinghua University
[3]Department of Computer Science, Zhejiang University City College
[1]{first.last}.gr@dartmouth.edu
[2]drogozhang@gmail.com
[3]marionojump0722@gmail.com
[4]soroush.vosoughi@dartmouth.edu

Abstract

This paper studies the relative importance of attention heads in Transformer-based models to aid their interpretability in cross-lingual and multi-lingual tasks. Prior research has found that only a few attention heads are important in each mono-lingual Natural Language Processing (NLP) task and pruning the remaining heads leads to comparable or improved performance of the model. However, the impact of pruning attention heads is not yet clear in cross-lingual and multi-lingual tasks. Through extensive experiments, we show that (1) pruning a number of attention heads in a multi-lingual Transformer-based model has, in general, positive effects on its performance in cross-lingual and multi-lingual tasks and (2) the attention heads to be pruned can be ranked using gradients and identified with a few trial experiments. Our experiments focus on sequence labeling tasks, with potential applicability on other cross-lingual and multi-lingual tasks. For comprehensiveness, we examine two pre-trained multi-lingual models, namely multi-lingual BERT (mBERT) and XLM-R, on three tasks across 9 languages each. We also discuss the validity of our findings and their extensibility to truly resource-scarce languages and other task settings.

1 Introduction

Prior research on mono-lingual Transformer-based (Vaswani et al., 2017) models reveals that a subset of their attention heads makes key contributions to each task, and the models perform comparably well (Voita et al., 2019; Michel et al., 2019) or even better (Kovaleva et al., 2019) with the remaining heads pruned [1]. While multi-lingual Transformer-based models, e.g. mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020), are widely applied in cross-lingual and multi-lingual NLP tasks [2] (Wang et al., 2019; Keung et al., 2019; Eskander et al., 2020), no attempt has been made to extend the findings on the aforementioned mono-lingual research to this context. In this paper, we explore the roles of attention heads in cross-lingual and multi-lingual tasks for two reasons. First, better understanding and interpretability of Transformer-based models leads to efficient model designs and parameter tuning. Second, head-pruning makes Transformer-based models more applicable to truly resource-scarce languages if it does not negatively affect model performance significantly.

The biggest challenge we face when studying the roles of attention heads in cross-lingual and multi-lingual tasks is locating the heads to prune. Existing research has shown that each attention head is specialized to extract a collection of linguistic features, e.g., the middle layers of BERT mainly extract syntactic features (Vig and Belinkov, 2019; Hewitt and Manning, 2019) and the fourth head on the fifth layer of BERT greatly contributes to the coreference resolution task (Clark et al., 2019). Thus, we hypothesize that important feature extractors for a task should be shared across languages and the remaining heads can be pruned. We evaluate two approaches used to rank attention heads, the first of which is layer-wise relevance propagation (LRP, Ding et al. (2017)). Voita et al. (2019) interpreted the adaptation of LRP in Transformer-based models on machine translation. Motivated by Feng et al. (2018) and Serrano and Smith (2019), we design a second ranking method based on gradients since the gradients on each attention head

*Equal contribution.

†Work done when interning at the Minds, Machines, and Society Lab at Dartmouth College.

[1]We regard single-source machine translation as a mono-lingual task since the inputs to the models are mono-lingual.

[2]We define a cross-lingual task as a task whose test set is in a different language from its training set. A multi-lingual task is a task whose training set is multi-lingual and the languages of its test set belong to the languages of the training set.

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 1956–1966
August 1–6, 2021.

reflect its contribution to the predictions.

We study the effects of pruning attention heads on three sequence labeling tasks, namely part-of-speech tagging (POS), named entity recognition (NER), and slot filling (SF). We focus on sequence labeling tasks since they are more difficult to annotate than document- or sentence-level classification datasets and require more treatment in cross-lingual and multi-lingual research. We choose POS and NER datasets in 9 languages, where English (EN), Chinese (ZH), and Arabic (AR) are candidate source languages. The MultiAtis++ corpus (Xu et al., 2020) is used in the SF evaluations with EN as the source language. We do not include syntactic chunking and semantic role labeling tasks due to lack of availability of manually written and annotated corpora. In these experiments, we rank attention heads based only on the source language(s) to ensure the extensibility of the learned knowledge to cross-lingual tasks and resource-poor languages. In our preliminary experiments comparing the gradient-based method and LRP, the average F1 score improvements on NER with mBERT are 0.69 (cross-lingual) and 0.24 (multi-lingual) for LRP and 0.81 (cross-lingual) and 0.31 (multi-lingual) for the gradient-based method, though both methods rank attention heads similarly. Thus we choose the gradient-based method to rank attention heads in all our experiments.

Our evaluations confirm that only a subset of attention heads in each Transformer-based model makes key contributions to each cross-lingual or multi-lingual task and that these heads are shared across languages. Performance of models generally drop when the highest-ranked or randomly selected heads are pruned, validating the head rankings generated by our gradient-based method. We also observe performance improvements on tasks with multiple source languages by pruning attention heads. Our findings potentially apply to truly resource-scarce languages since we show that the models perform better with attention heads pruned when fewer training instances are available in the target languages.

The contributions of this paper are three-fold:

- We explore the roles of attention heads in multi-lingual Transformer-based models and find that pruning certain heads leads to comparable or better performance in cross-lingual and multi-lingual sequence labeling tasks.
- We adapt a gradient-based method to locate attention heads that can be pruned without exhaustive experiments on all possible combinations.
- We show the correctness, robustness, and extensibility of the findings and our head ranking method under a wide range of settings through comprehensive experiments.

LC	Language Family	Training Size	
		POS	NER
EN	IE, Germanic	12,543	14,987
DE	IE, Germanic	13,814	12,705
NL	IE, Germanic	12,264	15,806
AR	Afro-Asiatic, Semitic	6,075	1,329
HE	Afro-Asiatic, Semitic	5,241	2,785
ZH	Sino-Tibetan	3,997	20,905
JA	Japanese	7,027	800
UR	IE, Indic	4,043	289,741
FA	IE, Iranian	4,798	18,463

Table 1: Details of POS and NER datasets in our experiments. LC refers to language code. Training size denotes the number of training instances.

2 Datasets

We use human-written and manually annotated datasets in experiments to avoid noise from machine translation and automatic label projection.

We choose POS and NER datasets in 9 languages, namely EN, ZH, AR, Hebrew (HE), Japanese (JA), Persian (FA), German (DE), Dutch (NL), and Urdu (UR). As Table 1 shows, these languages fall in diverse language families and the datasets are very different in size. EN, ZH, and AR are used as candidate source languages since they are resource-rich in many NLP tasks. Our POS datasets are all from Universal Dependencies (UD) v2.7 [3]. These datasets are labeled with a common label set containing 17 POS tags.

For NER, we use NL, EN, and DE datasets from CoNLL-2002 and 2003 challenges (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003). Additionally, we use the People's Daily dataset [4], iob2corpus [5], AQMAR (Mohit et al., 2012), ArmanPerosNERCorpus (Poostchi et al., 2016), MK-PUCIT (Kanwal et al., 2020), and a news-based NER dataset (Mordecai and Elhadad, 2012) for the languages CN, JA, AR, FA, UR, and

[3] http://universaldependencies.org/

[4] http://github.com/OYE93/Chinese-NLP-Corpus/tree/master/NER/People'sDaily

[5] http://github.com/Hironsan/IOB2Corpus

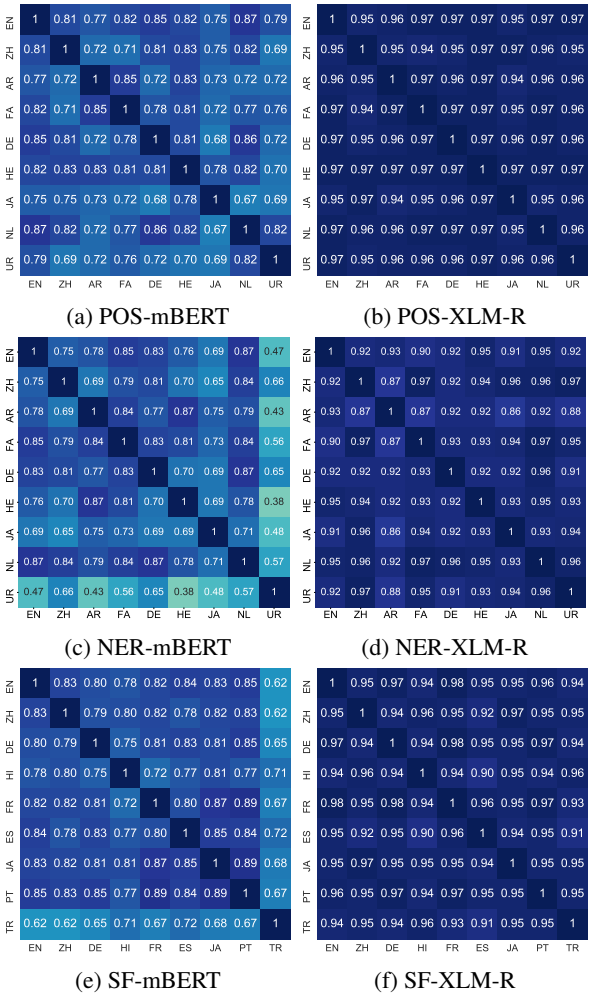
HE, respectively. Since the NER datasets are individually constructed in each language, their label sets do not fully agree. As there are four NE types (PER, ORG, LOC, MISC) in the three source-language datasets, we merge other NE types into the MISC class to allow cross-lingual evaluations.

We evaluate SF models on MultiAtis++ with EN as the source language and Spanish (ES), Portuguese (PT), DE, French (FR), ZH, JA, Hindi (HI), and Turkish (TR) as target languages. There are 71 slot types in the TR dataset, 75 in the HI dataset, and 84 in the other datasets. We do not use the intent labels in our evaluations since we study only sequence labeling tasks. Thus our results are not directly comparable with Xu et al. (2020).

3 Methodology

Here, we introduce the gradient-based method we use in the experiments to rank the attention heads. Feng et al. (2018) claim that gradients measure the importance of features to predictions. Since each head functions similarly as a standalone feature extractor in a Transformer-based model, we use gradients to approximate the importance of the feature set extracted by each head and rank the heads accordingly. Michel et al. (2019) determine importance of heads with accumulated gradients at each head in a training epoch. Different from their approach, we fine-tune the model on the training set and rank the heads using gradients on the development set to ensure that the head importance rankings are not significantly correlated with the training instances in one source language. Specifically, our method generates head rankings for each language in three steps:
(1) We fine-tune a Transformer-based model on a mono-lingual task for three epochs.
(2) We re-run the fine-tuned model on the development partition of the dataset with back-propagation but not parameter updates to obtain gradients.
(3) We sum up the absolute gradients on each head, layer-wise normalize the accumulated gradients, and scale them into the range [0, 1] globally.

We show Spearman's rank correlation coefficients (Spearman's ρ) between head rankings of each language pair generated by our method on POS, NER, and SF in Figure 1. The highest-ranked heads largely overlap in all three tasks, while the rankings of unimportant heads vary more in mBERT than XLM-R.



(a) POS-mBERT (b) POS-XLM-R

(c) NER-mBERT (d) NER-XLM-R

(e) SF-mBERT (f) SF-XLM-R

Figure 1: Spearman's ρ of head ranking matrices between languages in the POS, NER, and SF tasks. Darker colors indicate higher correlations.

After ranking the attention heads, we fine-tune the model, with the lowest-ranked head in the source language pruned. We keep increasing the number of heads to prune until it reaches a preset limit or when the performance starts to drop. We limit the number of trials to 12 since the models mostly show improved performance within 12 attempts [6].

4 Experiments and Analysis

This section displays and explains experimental results on cross-lingual and multi-lingual POS, NER, and SF tasks. Training sets in target languages are not used to train the model under the cross-lingual setting. Our experiments are based on the Huggingface (Wolf et al., 2020) implementations of mBERT

[6] On average 7.52 and 6.58 heads are pruned for POS, 7.54 and 7.28 heads for NER, and 6.19 and 6.31 heads for SF, respectively in mBERT and XLM-R models.

SL	TL	mBERT				XLM-R			
		Unpruned		Pruned		Unpruned		Pruned	
		CrLing	MulLing	CrLing	MulLing	CrLing	MulLing	CrLing	MulLing
EN	ZH	59.88	95.10	**59.99**	**95.31**	41.10	95.87	**46.18**	**95.99**
	AR	55.98	95.64	**56.71**	**95.68**	66.75	96.07	**67.02**	**96.13**
	FA	57.94	94.48	**58.34**	**94.81**	**66.60**	96.85	66.50	**97.09**
	DE	88.86	94.81	**89.13**	**94.94**	89.41	94.81	**89.78**	**95.19**
	HE	77.91	96.45	**78.01**	**96.58**	77.48	97.26	**80.37**	**97.30**
	JA	44.73	96.84	**45.95**	**96.97**	30.98	97.52	**33.64**	**97.62**
	NL	87.45	96.47	**87.48**	**96.69**	**88.06**	**97.04**	88.03	97.02
	UR	53.21	91.92	**54.78**	**92.17**	55.45	92.94	**56.04**	**93.07**
ZH	EN	55.63	96.52	**57.05**	**96.64**	42.35	97.19	**43.38**	**97.32**
	AR	38.41	95.62	**41.03**	**95.66**	36.71	95.99	**38.19**	**96.07**
	FA	43.68	94.55	**45.29**	**94.63**	33.43	97.07	**34.64**	**97.09**
	DE	63.50	94.62	**64.36**	**94.75**	46.58	95.06	**47.47**	**95.22**
	HE	57.14	96.51	**57.94**	**96.58**	**51.26**	97.06	50.42	**97.19**
	JA	43.63	96.73	**44.69**	**97.01**	49.12	97.32	**49.74**	**97.34**
	NL	59.95	96.78	**61.10**	**96.97**	40.78	97.30	**42.50**	**97.43**
	UR	43.82	92.21	**44.07**	**92.26**	**30.08**	92.90	29.26	**93.01**
AR	EN	54.77	96.50	**56.90**	**96.53**	61.73	97.21	**63.63**	**97.31**
	ZH	46.19	95.16	**47.14**	**95.31**	25.12	95.16	**34.71**	**96.04**
	FA	63.82	94.52	**64.02**	**94.64**	70.92	97.15	**71.55**	**97.20**
	DE	56.88	94.82	**57.85**	**94.98**	65.21	95.16	**68.28**	**95.29**
	HE	60.33	96.44	**61.88**	**96.70**	67.45	97.23	**67.72**	**97.34**
	JA	**44.32**	97.02	44.18	**97.15**	22.11	97.52	**29.21**	**97.65**
	NL	58.86	96.87	**60.31**	**97.03**	62.93	96.87	**64.80**	**97.50**
	UR	49.31	92.00	**49.76**	**92.16**	54.79	92.74	**56.06**	**92.88**

Table 2: F-1 scores of mBERT and XLM on POS. SL and TL refer to source and target languages and CrLing and MulLing stand for cross-lingual and multi-lingual settings, respectively. Unpruned results are produced by the full models and pruned results are the best scores each model produces with up to 12 lowest-ranked heads pruned. The higher performance in each pair of pruned and unpruned experiments is in bold.

and XLM-R. Specifically, we use the pre-trained bert-base-multilingual-cased and xlm-roberta-base models for their comparable model sizes. The models are fine-tuned for 3 epochs with a learning rate of 5e-5 in all the experiments. We use the official dataset splits and load training instances with sequential data samplers, so the reported evaluation scores are robust to randomness.

4.1 POS

Table 2 shows the evaluation scores on POS with three source language choices. In the majority (88 out of 96 pairs) of experiments, pruning up to 12 attention heads improves mBERT and XLM-R performance. Results are comparable in the other 8 experiments with and without head pruning. Average F-1 score improvements are 0.91 for mBERT and 1.78 for XLM-R in cross-lingual tasks, and 0.15 for mBERT and 0.17 for XLM-R in multi-lingual tasks. These results support that pruning heads generally has positive effects on model performance in cross-lingual and multi-lingual tasks, and that our method correctly ranks the heads.

Consistent with Conneau et al. (2020), XLM-R usually outperforms mBERT, with exceptions in cross-lingual experiments where ZH and JA datasets are involved. Word segmentation in ZH and JA is different from the other languages we choose, e.g. words are not separated by white spaces and unpaired adjacent word pieces often make up a new word. As XLM-R applies the SentencePiece tokenization method (Kudo and Richardson, 2018), it is more likely to detect wrong word boundaries and make improper predictions than mBERT in cross-lingual experiments involving ZH or JA datasets. We note that the performance improvements are solid regardless of the

SL	TL	mBERT				XLM-R			
		Unpruned		Pruned		Unpruned		Pruned	
		CrLing	MulLing	CrLing	MulLing	CrLing	MulLing	CrLing	MulLing
EN	ZH	47.64	93.24	**51.61**	**93.71**	29.97	90.99	**32.33**	**91.11**
	AR	38.81	70.55	**38.93**	**73.32**	41.21	71.77	**43.78**	**74.28**
	FA	**40.12**	96.70	39.81	**96.97**	54.90	96.62	**55.72**	**96.98**
	DE	56.43	79.11	**58.27**	**79.19**	63.71	82.31	**66.48**	**83.10**
	HE	**46.92**	**89.18**	46.55	88.49	**56.96**	88.02	56.87	**89.67**
	JA	42.45	**84.91**	**44.14**	84.34	33.87	81.48	**37.88**	**82.35**
	NL	64.51	84.90	**65.56**	**85.17**	77.15	90.21	**77.66**	**90.38**
	UR	37.34	**99.29**	**40.60**	99.22	58.25	**99.15**	**58.68**	99.07
ZH	EN	38.58	87.65	**41.40**	**87.99**	56.40	90.72	**58.55**	**91.05**
	AR	36.43	72.27	**36.99**	**72.86**	34.31	74.84	**36.11**	**75.68**
	FA	45.68	96.21	**46.57**	**96.23**	**51.60**	95.63	51.51	**95.66**
	DE	29.07	**79.04**	**33.81**	78.67	**56.22**	82.33	55.51	**82.54**
	HE	47.14	88.20	**47.68**	**89.35**	48.52	85.95	**48.94**	**87.79**
	JA	49.21	82.02	**51.69**	**83.20**	46.18	80.19	**47.06**	**82.63**
	NL	29.75	84.61	**31.46**	**85.28**	49.59	89.56	**52.27**	**90.56**
	UR	44.61	99.26	**46.33**	**99.28**	48.98	98.99	**55.95**	**99.10**
AR	EN	19.29	**87.86**	**20.07**	87.82	**51.33**	90.37	51.00	**91.01**
	ZH	**41.70**	93.46	40.43	**93.54**	25.78	90.51	**31.03**	**91.00**
	FA	46.57	96.82	**46.87**	**96.87**	**53.35**	96.55	52.60	**96.74**
	DE	24.47	75.78	**25.62**	**78.04**	**50.87**	82.63	50.00	**82.73**
	HE	**47.15**	86.77	46.72	**87.64**	49.52	87.37	**50.85**	**89.28**
	JA	41.49	79.90	**42.11**	**83.17**	36.98	**81.72**	**38.87**	80.92
	NL	26.00	84.83	**26.34**	**85.24**	**49.27**	90.73	48.87	**91.11**
	UR	**46.47**	99.26	45.66	**99.31**	48.48	99.10	**53.51**	**99.15**

Table 3: F-1 scores of mBERT and XLM on NER. SL and TL refer to source and target languages and CrLing and MulLing stand for cross-lingual and multi-lingual settings, respectively. Unpruned results are produced by the full models and pruned results are the best scores each model produces with up to 12 lowest-ranked heads pruned.

source language selection and severe differences of training data sizes in EN, ZH, and AR. This demonstrates the correctness of the head rankings our method generates and that the important attention heads for a task are almost language invariant.

We also examine to what extent the score improvements are affected by the relationships between source and target languages, e.g. language families, URIEL language distance scores (Littell et al., 2017), and the similarity of the head ranking matrices. There are three non-exclusive clusters of language families (containing more than one language) in our choice of languages, namely Indo-European (IE), Germanic, and Semitic languages. Average score improvements between models with and without head pruning are 0.40 (IE), 0.16 (Germanic), and 0.91 (Semitic) for mBERT and 0.19 (IE), 0.18 (Germanic), and 0.19 (Semitic) for XLM-R. In comparison, the overall average score improvements are 0.53 for mBERT and 0.97 for XLM-R. Despite the generally higher performance of models when the source and target languages are in the same family, the score improvements by pruning heads are not necessarily associated with language families. Additionally, we use Spearman's ρ to measure the correlations between improved F-1 scores and URIEL language distances. The correlation scores are 0.11 (cross-lingual) and 0.12 (multi-lingual) for mBERT, and -0.40 (cross-lingual) and 0.23 (multi-lingual) for XLM-R. Similarly, the Spearman's ρ between score improvements and similarities in head ranking matrices shown in Figure 1 are -0.34 (cross-lingual) and 0.25 (multi-lingual) for mBERT, and -0.52 (cross-lingual) and -0.10 (multi-lingual) for XLM-R. This indicate that except in the cross-lingual XLM-R model which faces word segmentation issues on ZH or JA experiments, pruning attention heads

SL	TL	mBERT				XLM-R			
		Unpruned		Pruned		Unpruned		Pruned	
		CrLing	MulLing	CrLing	MulLing	CrLing	MulLing	CrLing	MulLing
EN	ZH	69.83	94.11	**71.84**	**94.25**	62.58	93.97	**67.98**	**94.29**
	DE	60.69	94.60	**66.97**	**94.95**	82.85	94.81	**83.50**	**95.35**
	HI	44.28	85.93	**45.84**	**87.08**	58.32	86.72	**66.39**	**87.16**
	FR	60.44	93.96	**67.13**	**94.18**	76.53	93.51	**77.59**	**93.77**
	ES	72.27	87.71	**73.96**	**88.17**	81.70	**89.10**	**81.88**	88.83
	JA	68.28	93.73	**68.32**	**93.78**	32.39	93.65	**36.68**	**93.71**
	PT	59.37	**90.83**	**63.23**	90.82	77.42	90.76	**77.54**	**91.24**
	TR	28.11	83.41	**32.21**	**84.31**	45.91	83.20	**52.64**	**84.30**
	EN	**95.43**		95.27		94.59		**94.87**	

Table 4: Slot F-1 scores on the MultiAtis++ corpus. CrLing and MulLing refer to cross-lingual and multi-lingual settings, respectively. SL and TL refer to source and target languages, respectively. English mono-lingual results are reported for validity check purposes.

improves model performance regardless of the distances between source and target languages. Thus our findings are potentially applicable to all cross-lingual and multi-lingual POS tasks.

4.2 NER

As Table 3 shows, pruning attention heads generally has positive effects on our cross-lingual and multi-lingual NER models. Even in the multi-lingual AR-UR experiment where the full mBERT model achieves an F-1 score of 99.26, the score is raised to 99.31 by pruning heads. Scores are comparable with and without head pruning in the 19 cases where model performances are not improved. This also lends support to the specialized role of important attention heads and the consistency of head rankings across languages. In NER experiments, performance drops mostly happen when the source and target languages are from different families. This is likely caused by the difference between named entity (NE) representations across language families. We show in Section 5.2 that the gap is largely bridged when a language from the same family as the target language is added to the source languages.

Average score improvements are comparable on mBERT (0.81 under cross-lingual and 0.31 under multi-lingual settings) and XLM-R (1.08 under cross-lingual and 0.67 under multi-lingual settings) in NER experiments. The results indicate that the performance improvements introduced by head-pruning are not sensitive to the pre-training corpora of models. The correlations between F-1 score improvements and URIEL language distances are small, with Spearman's ρ of -0.05 (cross-lingual) and -0.27 (multi-lingual) for mBERT and 0.10 (cross-lingual) and 0.12 (multi-lingual) for XLM-R. Similarities between head ranking matrices do not greatly affect score improvements either, the Spearman's ρ of which are -0.08 (cross-lingual) and 0.06 (multi-lingual) for mBERT and 0.05 (cross-lingual) and 0.12 (multi-lingual) for XLM-R. The findings in POS and NER experiments are consistent, supporting our hypothesis that important heads for a task are shared by arbitrary source-target language selections.

4.3 Slot Filling

We report SF evaluation results in Table 4. In 31 out of 34 pairs of experiments, pruning up to 12 heads results in performance improvements, while the scores are comparable in the other three cases. These results agree with those in POS and NER experiments, showing that only a subset of heads in each model makes key contributions to cross-lingual or multi-lingual tasks.

We also evaluate the correlations between score changes and the closeness of source and target languages. In terms of URIEL language distances, the Spearman's ρ are 0.69 (cross-lingual) and 0.14 (multi-lingual) for mBERT and -0.59 (cross-lingual) and 0.14 (multi-lingual) for XLM-R. The coefficients are -0.25 (cross-lingual) and -0.73 (multi-lingual) for mBERT and -0.70 (cross-lingual) and -0.14 (multi-lingual) between score improvements and similarities in head ranking matrices. While these coefficients are generally higher than those in POS and NER evaluations, their p-values are also high (0.55 to 0.74), indicating the correlations between the score changes and source-

NER				
TL	Max-Pruning		Rand-Pruning	
	CrLing	MulLing	CrLing	MulLing
ZH	-1.74	+0.08	-2.44	+0.26
AR	-3.17	-2.42	-2.09	-0.43
DE	+0.88	-0.62	+0.57	-0.38
NL	-2.76	-0.23	+0.29	+0.36
FA	-0.86	-0.31	-2.52	-0.74
HE	-2.50	-2.15	-0.49	-4.21
JA	-1.48	-1.08	-2.65	-2.40
UR	-0.15	-0.10	-0.60	-0.12
POS				
TL	Max-Mask		Rand-Mask	
	CrLing	MulLing	CrLing	MulLing
ZH	+0.03	-0.39	-0.14	-0.20
AR	-0.65	-0.04	-0.66	-0.12
DE	-0.64	-0.04	-0.64	-0.14
NL	-0.13	-0.13	-0.11	-0.16
FA	-0.75	-0.03	-0.53	-0.25
HE	-1.27	-0.28	-1.06	+0.05
JA	-22.29	-0.05	-1.23	-0.05
UR	-1.78	-0.11	-0.77	-0.07

Table 5: F-1 score differences from the full mBERT model on NER (upper) and POS (lower) by pruning highest ranked (Max-Pruning) or random (Rand-Pruning) heads in the ranking matrices. The source language is EN. Blue and red cells indicate score drops and improvements, respectively.

target language closeness are not statistically significant. [7]

5 Discussions

In this section, we perform case studies to confirm the validity of our head ranking method. We also illustrate the extensibility of the knowledge we learn from the main experiments to a wider range of settings, e.g. when the training dataset is limited in size or constructed over multiple source languages.

5.1 Correctness of Head Rankings

We evaluate the correctness of our head ranking method through comparisons between results in Tables 2 and 3 and those produced by pruning (1) randomly sampled heads and (2) highest ranked heads. Specifically, we repeat the head-pruning experiments with mBERT on NER and POS using EN as the source language and display the score differences from the the full models in Table 5. Same as in the main experiments, we pick the best score from pruning 1 to 12 heads in each experiment. A random seed of 42 is used for sampling attention heads to prune under the random sampling setting.

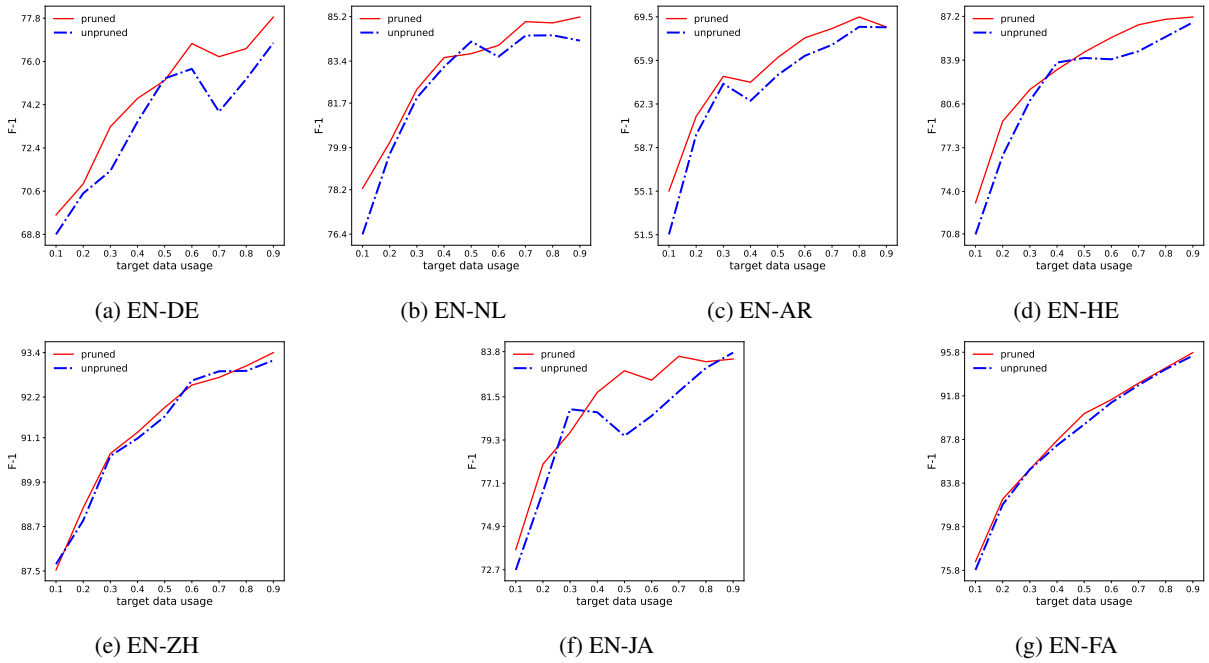
In 14 out of 16 NER experiments, pruning the heads ranked highest by our method results in noticeable performance drops compared to the full model. Consistently, pruning the highest-ranked attention heads harms the performance of mBERT in 15 out of 16 POS experiments. Though score changes are slightly positive for cross-lingual EN-DE and multi-lingual EN-ZH NER tasks and in the cross-lingual EN-ZH POS experiment, improvements introduced by pruning lowest-ranked heads are more significant, as Table 2 and Table 3 show. Pruning random attention heads also has mainly negative effects on the performance of mBERT. These results indicate that while pruning attention heads potentially boosts the performance of models, reasonably choosing the heads to prune is important. Our gradient-based method properly ranks the heads by their priority to prune.

5.2 Multiple Source Languages

Training cross-lingual models on multiple source languages is a practical way to improve their performance, due to enlarged training data size and supervision from source-target languages closer to each other (Wu et al., 2020; Moon et al., 2019; Chen et al., 2019; Rahimi et al., 2019; Täckström, 2012). We also explore the effects of pruning attention heads under the multi-source settings. In this section, we experiment with mBERT on EN, DE, AR, HE, and ZH datasets for both NER and POS tasks. These languages fall into three mutually exclusive language families, enabling our analysis on the influence of training cross-lingual models with source languages belonging to the same family as the target language. Similar to related research, the model is fine-tuned on the concatenation of training datasets in all the languages but the one on which the model is tested.

Since the head ranking matrices are not identical across languages, we design three heuristics to rank the heads in the multi-source experiments. The first method merges the head ranking matrices of all the source languages into one matrix and re-generates the rankings. The second method ranks the attention heads after summing up the head ranking

[7] The p-values for all the other Spearman's ρ we report are lower than 0.01, showing that those correlation scores are statistically significant.



(a) EN-DE (b) EN-NL (c) EN-AR (d) EN-HE

(e) EN-ZH (f) EN-JA (g) EN-FA

Figure 2: F-1 scores of mBERT on multi-lingual NER with 10% - 90% target language training data usage. Dashed blue lines indicate scores without head pruning and solid red lines show scores with head pruning.

NER					
	EN	DE	AR	HE	ZH
FL	60.77	59.16	35.90	51.19	44.18
MD	62.63	61.10	40.78	55.15	**47.59**
SD	63.38	61.66	**41.53**	54.20	47.08
EC	**64.63**	**61.71**	40.78	**56.26**	47.24
POS					
	EN	DE	AR	HE	ZH
FL	81.97	88.82	74.07	75.62	61.31
MD	82.99	89.19	74.65	77.00	61.74
SD	82.62	88.74	74.41	77.30	61.29
EC	**83.49**	**89.20**	**75.86**	**78.04**	**62.33**

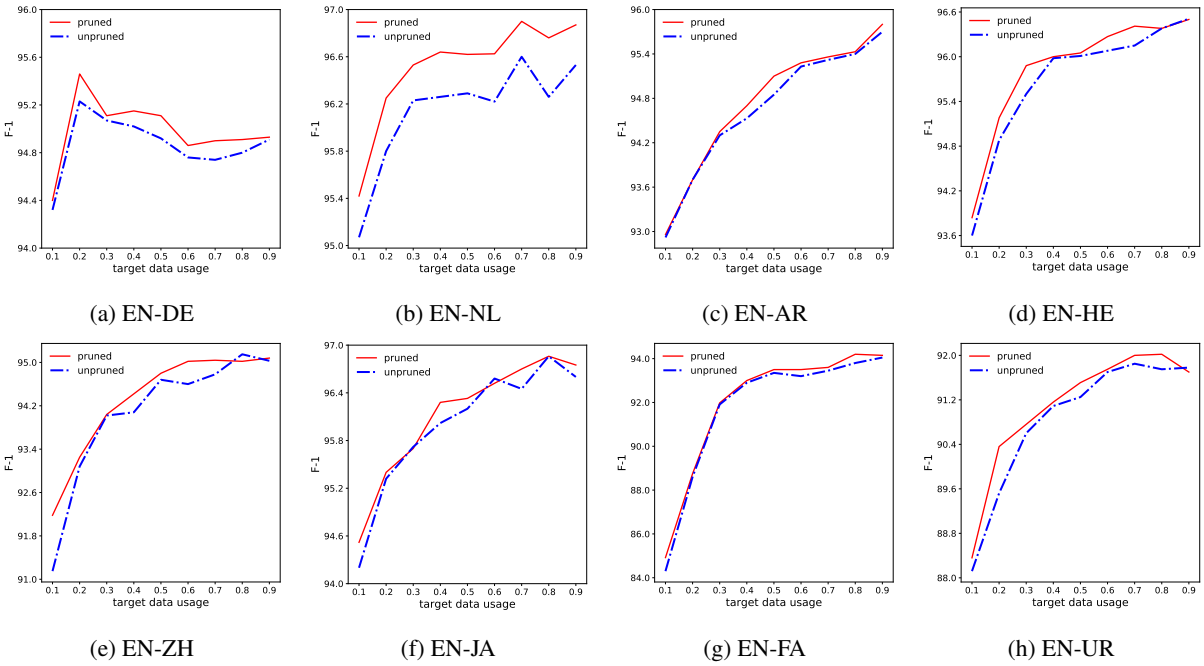
Table 6: Cross-lingual NER (upper) and POS (lower) evaluation results with multiple source languages. FL indicates unpruning. MD, SD, and EC are the three heuristics we examine.

matrices. We also examine the efficacy of pruning heads based on the head rankings from a single language. For this heuristic, we run experiments using the head ranking matrix from each language and report the highest score. We refer to the three heuristics as MD, SD, and EC, respectively.

Table 6 displays the results. We note that in the NER evaluations, the performance of mBERT on all the languages but ZH are higher than those in the single-source experiments. This supports our hypothesis that supervision from languages in the same family as the target language helps improve model performance. Different from NER, the evaluation results on POS are not much higher than the single-source evaluation scores, implying that syntactic features are more consistent across languages than appearances of named entities. However, it is consistent on both tasks that pruning attention heads brings performance boosts to all the multi-source experiments. While the EC heuristic provides the largest improvement margin in 3 out of 5 experiments, it requires a lot more trial experiments. MD and SD perform comparably well in most cases so they are also promising heuristics for ranking attention heads under the multi-source setting. The results support that pruning attention heads is beneficial to Transformer-based models in cross-lingual tasks even if the training dataset is already large and diverse in languages.

5.3 Extension to Resource-poor Languages

While the languages we use in the main experiments are not truly resource-poor, we examine our findings when training sets in the target languages are smaller. We design experiments under the multi-lingual setting with subsampled training datasets in target languages. Specifically, we randomly divide the training set of each target language into 10 disjoint subsets and compare model performance, with and without head pruning, using 1 to 9 sub-



(a) EN-DE (b) EN-NL (c) EN-AR (d) EN-HE

(e) EN-ZH (f) EN-JA (g) EN-FA (h) EN-UR

Figure 3: F-1 scores of mBERT on multi-lingual the POS task with 10% - 90% target language training data usage. Dashed blue lines indicate scores without head pruning and solid red lines show scores with head pruning.

sets. We do not use 0 or 10 subsets since they correspond to cross-lingual and fully multi-lingual settings, respectively. We run the evaluations on NER and POS tasks. These datasets vary greatly in size, allowing us to validate our findings on target-language datasets with as few as 80 training examples. The UR NER dataset is excluded from this case study since its training set is overly large. We note that the score differences with and without head pruning are, in the main experiments, consistent for all the choices of models and source languages. Thus, we only display the mBERT performance with EN as the source language on NER in Figure 2 and that on POS in Figure 3.

The evaluation results are consistent with those in our main experiments, where the model with up to 12 attention heads pruned generally outperforms the full mBERT model. This further supports our hypothesis that pruning lower-ranked attention heads has positive effects on the performance of Transformer-based models in truly resource-scarce languages. It is also worth noting that pruning attention heads often causes the mBERT model to reach peak evaluation scores with less training data in the target language. For example, in the EN-JA NER experiments, the full model achieves the highest F-1 score when all the 800 training instances in the JA dataset are used while the model with heads pruned achieves a comparable score with 20% less data. This suggests that pruning attention heads makes deep Transformer-based models easier to train with less training data and thus more applicable to truly resource-poor languages.

6 Conclusion and Future Work

This paper studied the contributions of attention heads in Transformer-based models. Past research has shown that in mono-lingual tasks, pruning a large number of attention heads can achieve comparable or higher performance than the full models. However, we were the first to extend these findings to cross-lingual and multi-lingual sequence labeling tasks. Using a gradient-based method, we identified the heads to prune and showed that pruning attention heads generally has positive effects on mBERT and XLM-R performances. Additional case studies empirically demonstrated the validity of our findings and showed further extensibility of them to a wider range of task settings. In addition to better understanding of Transformer-based models under cross- and multi-lingual settings, our findings can be applied to existing models to achieve better performance with reduced training data and resource consumption. Future work could include improving model interpretability in other cross-lingual and multi-lingual tasks, e.g. XNLI (Conneau et al., 2018) and other passage-level classification tasks.

References

Xilun Chen, Ahmed Hassan Awadallah, Hany Hassan, Wei Wang, and Claire Cardie. 2019. Multi-source cross-lingual model transfer: Learning what to share. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3098–3112, Florence, Italy. Association for Computational Linguistics.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. 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*. 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.

Yanzhuo Ding, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Visualizing and understanding neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1150–1159, Vancouver, Canada. Association for Computational Linguistics.

Ramy Eskander, Smaranda Muresan, and Michael Collins. 2020. Unsupervised cross-lingual part-of-speech tagging for truly low-resource scenarios. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4820–4831, Online. Association for Computational Linguistics.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of neural models make interpretations difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3719–3728, Brussels, Belgium. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. 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 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Safia Kanwal, Kamran Malik, Khurram Shahzad, Faisal Aslam, and Zubair Nawaz. 2020. Urdu named entity recognition: Corpus generation and deep learning applications. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 19(1):8:1–8:13.

Phillip Keung, Yichao Lu, and Vikas Bhardwaj. 2019. Adversarial learning with contextual embeddings for zero-resource cross-lingual classification and NER. 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 1355–1360, Hong Kong, China. 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.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14. 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*, volume 32, pages 14014–14024. Curran Associates, Inc.

Behrang Mohit, Nathan Schneider, Rishav Bhowmick, Kemal Oflazer, and Noah A. Smith. 2012. Recall-oriented learning of named entities in Arabic

Wikipedia. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 162–173, Avignon, France. Association for Computational Linguistics.

Taesun Moon, Parul Aswathy, Jian Ni, and Radu Florian. 2019. Towards lingua franca named entity recognition with bert. *arXiv preprint arXiv:1912.01389*.

Naama Mordecai and Michael Elhadad. 2012. Hebrew named entity recognition.

Hanieh Poostchi, Ehsan Zare Borzeshi, Mohammad Abdous, and Massimo Piccardi. 2016. PersoNER: Persian named-entity recognition. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3381–3389, Osaka, Japan. The COLING 2016 Organizing Committee.

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.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

Oscar Täckström. 2012. Nudging the envelope of direct transfer methods for multilingual named entity recognition. In *Proceedings of the NAACL-HLT Workshop on the Induction of Linguistic Structure*, pages 55–63, Montréal, Canada. Association for Computational Linguistics.

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

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

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.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

Yuxuan Wang, Wanxiang Che, Jiang Guo, Yijia Liu, and Ting Liu. 2019. Cross-lingual BERT transformation for zero-shot dependency parsing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5721–5727, Hong Kong, China. Association for Computational Linguistics.

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.

Qianhui Wu, Zijia Lin, Börje Karlsson, Jian-Guang Lou, and Biqing Huang. 2020. Single-/multi-source cross-lingual NER via teacher-student learning on unlabeled data in target language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6505–6514, Online. Association for Computational Linguistics.

Weijia Xu, Batool Haider, and Saab Mansour. 2020. End-to-end slot alignment and recognition for cross-lingual NLU. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5052–5063, Online. Association for Computational Linguistics.