From Zero to Hero: On the Limitations of Zero-Shot Language Transfer with Multilingual Transformers

Anne Lauscher[1]*, **Vinit Ravishankar**[2]*, **Ivan Vulić**[3], **and Goran Glavaš**[1]

[1]Data and Web Science Group, University of Mannheim, Germany
[2]Language Technology Group, University of Oslo, Norway
[3]Language Technology Lab, University of Cambridge, UK

[1]{anne,goran}@informatik.uni-mannheim.de,
[2]vinitr@ifi.uio.no, [3]iv250@cam.ac.uk

Abstract

Massively multilingual transformers (MMTs) pretrained via language modeling (e.g., mBERT, XLM-R) have become a default paradigm for zero-shot language transfer in NLP, offering unmatched transfer performance. Current evaluations, however, verify their efficacy in transfers (a) to languages with sufficiently large pretraining corpora, and (b) between close languages. In this work, we analyze the limitations of downstream language transfer with MMTs, showing that, much like cross-lingual word embeddings, they are substantially less effective in resource-lean scenarios and for distant languages. Our experiments, encompassing three lower-level tasks (POS tagging, dependency parsing, NER) and two high-level tasks (NLI, QA), empirically correlate transfer performance with linguistic proximity between source and target languages, but also with the size of target language corpora used in MMT pretraining. Most importantly, we demonstrate that the inexpensive few-shot transfer (i.e., additional fine-tuning on a few target-language instances) is surprisingly effective across the board, warranting more research efforts reaching beyond the limiting zero-shot conditions.

1 Introduction and Motivation

Labeled datasets of sufficient size support supervised learning in NLP. The notorious tediousness, subjectivity, and cost of linguistic annotation (Dandapat et al., 2009; Sabou et al., 2012; Fort, 2016), coupled with plethora of structurally different NLP tasks, lead to existence of such datasets only for a handful of resource-rich languages (Bender, 2011; Ponti et al., 2019; Joshi et al., 2020). This data scarcity renders the need for effective *cross-lingual transfer* strategies: how can we exploit abundant labeled data from resource-rich languages to make predictions in resource-lean languages? In the most extreme scenario, termed *zero-shot cross-lingual transfer*, not a single labeled instance exists for a target language. Recent work has placed much emphasis on this scenario exactly; in theory, it offers the widest portability across the world's 7,000+ languages (Pires et al., 2019; Artetxe et al., 2020b; Lin et al., 2019; Cao et al., 2020; Hu et al., 2020).

The current mainstay of cross-lingual transfer in NLP are approaches based on continuous cross-lingual representation spaces such as cross-lingual word embeddings (CLWEs) (Ruder et al., 2019) and, most recently, massively multilingual transformer networks (MMTs), pretrained on multilingual corpora with language modeling (LM) objectives (Devlin et al., 2019; Conneau and Lample, 2019; Conneau et al., 2020). The latter have *de facto* become the default language transfer paradigm, with multiple studies reporting their unparalleled transfer performance (Pires et al., 2019; Wu and Dredze, 2019; Rönnqvist et al., 2019; Karthikeyan et al., 2020; Wu et al., 2020).

Key Questions and Contributions. In this work, we dissect the current state-of-the-art MMT-based approach to (zero-shot) cross-lingual transfer, and analyze a variety of conditions and factors that critically impact or limit effective cross-lingual transfer. Our aim is to provide answers to the following crucial questions.

(Q1) *What is the role of language (dis)similarity and language-specific corpora size in pretraining?*

Current cross-lingual transfer via MMTs is still primarily focused on either (1) languages that are typologically or etymologically close to English (e.g., German, Scandinavian languages, French, Spanish), or (2) languages with large monolingual corpora, well-represented in the multilingual pretraining corpora (e.g., Arabic, Hindi, Chinese). Wu et al. (2020) suggest that LM-pretrained transform-

*Equal contribution.

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 4483–4499,
November 16–20, 2020.

ers, much like static word embeddings models, produce topologically similar representation spaces that can easily be aligned between languages, offering this as explanation of language transfer efficacy of MMTs. However, transfer with static CLWEs has been shown ineffective between dissimilar languages (Søgaard et al., 2018; Vulić et al., 2019) or languages with small corpora (Vulić et al., 2020).

We thus scrutinize MMTs in diverse zero-shot transfer settings and find, in line with prior work on CLWEs, that MMTs' transfer performance critically depends on (1) linguistic (dis)similarity between the source and target language and (2) size of the pretraining corpus of the target language.

(Q2) *What is the role of a particular task in consideration for transfer performance?*

We conduct all analyses across five different tasks, which we roughly divide into two groups: (1) "low-level" tasks (POS-tagging, dependency parsing, and NER); and (2) "high-level" language understanding (LU) tasks (NLI and QA). We show that transfer performance in both zero-shot and few-shot scenarios largely depends on the "task level".

(Q3) *Can we (even) predict transfer performance?*

Running a simple regression on available transfer results, we show that we can (roughly) predict the transfer performance from (1) language proximity (Littell et al., 2017) for low-level tasks; (2) combination of language proximity and size of target-language pretraining corpora for high-level tasks.

(Q4) *Should we focus more on few-shot transfer scenarios and quick annotation cycles?*

Complementing the efforts on improving zero-shot transfer (Cao et al., 2020), we point to few-shot transfer as a very effective mechanism for improving target-language performance. Similar to the seminal "pre-neural" work of Garrette and Baldridge (2013), our results suggest that only several hours (or even minutes) of annotation work can "buy" substantial performance gains for low-resource targets. For all five tasks in our study, we obtain substantial (and in some cases surprisingly large) improvements with minimal annotation effort. For instance, we improve dependency parsing for some target languages up to 40 UAS points with as few as 10 target language sentences. Crucially, the few-shot gains are most pronounced exactly where zero-shot transfer fails: for distant target languages with small monolingual corpora.

2 Background and Related Work

For completeness, we provide a brief overview of **1)** cross-lingual transfer approaches, with a focus on **2)** massively multilingual transformer (MMT) models, and then **3)** position our work w.r.t. other studies that examine different properties of MMTs.

2.1 Cross-Lingual Transfer Paradigms

Language transfer entails representing texts from both the source and target language in a shared cross-lingual space. Transfer paradigms based on discrete text representations include *machine translation* (MT) of target language text to the source language (or vice-versa) (Mayhew et al., 2017; Eger et al., 2018), and grounding texts from both languages in *multilingual knowledge bases* (KBs) (Navigli and Ponzetto, 2012; Lehmann et al., 2015). While reliable MT hinges on availability of large parallel corpora, transfer via multilingual KBs (Camacho-Collados et al., 2016; Mrkšić et al., 2017) is impaired by the limited KB coverage and inaccurate entity linking (Moro et al., 2014; Raiman and Raiman, 2018).

Therefore, recent years have seen a surge of language transfer methods based on continuous representation spaces. The previous state-of-the-art, cross-lingual word embeddings (CLWEs) (Mikolov et al., 2013; Ammar et al., 2016; Artetxe et al., 2017; Smith et al., 2017; Glavaš et al., 2019; Vulić et al., 2019) and sentence embeddings (Artetxe and Schwenk, 2019), have most recently been replaced by massively multilingual transformers (MMTs) pretrained with LM objectives (Devlin et al., 2019; Conneau and Lample, 2019; Conneau et al., 2020).

2.2 Massively Multilingual Transformers

Multilingual BERT (mBERT). At BERT's (Devlin et al., 2019) core is a multi-layer transformer network (Vaswani et al., 2017), parameters of which are pretrained using masked language modeling (MLM) and next sentence prediction (NSP). In MLM, some tokens are masked out and they need to be recovered from the context; NSP predicts adjacency of sentences in text, informing the transformer of longer dependencies, beyond sentence boundaries. Liu et al. (2019) introduce RoBERTa, a more robust instance of BERT trained on larger corpora using only the MLM objective. Multilingual BERT (mBERT) is an instance of BERT trained on concatenation of 104 largest Wikipedias. The effects of underfitting for lan-

guages with small Wikipedias and overfitting to languages with large Wikipedias, are respectively attenuated with exponentially smoothed up-sampling and down-sampling.

XLM on RoBERTa (XLM-R). XLM-R (Conneau et al., 2020) is an instance of RoBERTa, robustly trained on a large multilingual CommonCrawl-100 (CC-100) corpus (Wenzek et al., 2019) covering 100 languages. mBERT's corpus and CC-100 share 88 languages, with corresponding CC-100's portions being much larger than mBERT's Wikipedias.

The "Curse of Multilinguality". For XLM-R, Conneau et al. (2020) observe that for a fixed model capacity, downstream cross-lingual transfer improves with more pretraining languages up to a point after which adding more pretraining languages hurts downstream transfer. This effect, termed the "curse of multilinguality", can be mitigated by increasing model's capacity (Artetxe et al., 2020b) or additional training for particular language pairs (Pfeiffer et al., 2020). This points to MMTs' capacity (i.e., computational budgets), as a critical factor for effective zero-shot transfer.

In contrast, we identify few-shot transfer as a much more cost-effective strategy for improving downstream target language performance (§4). We show for a number of target languages and downstream tasks, that one can obtain large performance gains at very small annotation cost, without having to pretrain from scratch an MMT of larger capacity.

2.3 Cross-Lingual Transfer with MMTs

A body of recent work probed the knowledge encoded in MMTs, primarily mBERT. Libovický et al. (2020) analyze language-specific versus language-universal knowledge encoded in mBERT. Pires et al. (2019) demonstrate mBERT to be effective for POS-tagging and NER zero-shot transfer between related languages. Wu and Dredze (2019) extend this analysis to more tasks and languages, and show that mBERT-based transfer is on a par with the best task-specific zero-shot transfer approaches. Similarly, Karthikeyan et al. (2020) prove mBERT to be effective for NER and NLI transfer to Hindi, Spanish, and Russian.[1] Importantly, they show that transfer effectiveness does not depend on the vocabulary overlap between the languages.

In most recent work, concurrent to this, Hu et al. (2020) introduce XTREME, a benchmark for evaluating multilingual encoders encompassing 9 tasks and 40 languages.[2] While the primary focus is a large-scale zero-shot transfer evaluation, they also experiment with target-language fine-tuning (1,000 instances for POS and NER). While Hu et al. (2020) focus on the evaluation aspects and protocols, in this work, we provide a more detailed analysis of the factors that hinder effective zero-shot transfer across several tasks.[3] We also put more emphasis on few-shot transfer, and approach it differently: by sequentially fine-tuning MMTs, first on (larger) source language training data and then on few target-language instances.

Artetxe et al. (2020b) and Wu et al. (2020) analyze different monolingual BERTs to explain transfer efficacy of mBERT. They find topological similarities between monolingual spaces, suggesting these are responsible for effective language transfer with MMTs. In essence, their work recasts the well-known assumption of approximate isomorphism of monolingual representation spaces (Søgaard et al., 2018). For CLWEs, this assumption does not hold for distant languages (Søgaard et al., 2018; Vulić et al., 2019), and in face of monolingual corpora of small size (Vulić et al., 2020). We demonstrate that the same is the case for zero-shot language transfer with MMTs: target-language performance drastically decreases as we move to more distant target languages with smaller pretraining corpora.

3 Zero-Shot Transfer: Analyses

We first address Q1 and Q2 (see §1): we conduct zero-shot language transfer experiments for five different tasks and analyze the factors behind the varying performance drops across target languages.

3.1 Experimental Setup

Tasks and Languages. We experiment with – **a)** low-level structured prediction tasks: POS-tagging, dependency parsing, and NER and **b)** high-level language understanding (LU) tasks: NLI and QA. We investigate if the factors that drive transfer performance differ between the two task groups.

Dependency Parsing (DEP). We use Universal Dependency treebanks (UD, Nivre et al., 2017) for English and following target languages (from 8 language families): Arabic (AR), Basque (EU), (Man-

[1] Note that all three are high-resource Indo-European languages with large Wikipedias.

[2] Note that none of the individual tasks in XTREME covers all 40 languages, but much smaller language subsets.

[3] We leave an even more general analysis that combines transfer both across tasks (Pruksachatkun et al., 2020; Glavaš and Vulić, 2020) *and* across languages for future work.

darin) Chinese (ZH), Finnish (FI), Hebrew (HE), Hindi (HI), Italian (IT), Japanese (JA), Korean (KO), Russian (RU), Swedish (SV), and Turkish (TR).

Part-of-speech Tagging (POS). Again, we use UD and obtain the Universal POS-tag (UPOS) annotations from the same treebanks as with DEP.

Named Entity Recognition (NER). We resort to the NER WikiANN dataset from Rahimi et al. (2019). We experiment with the same set of 12 target languages as in DEP and POS.

Cross-lingual Natural Language Inference (XNLI). We evaluate on the XNLI corpus (Conneau et al., 2018) created by translating dev and test portions of the English Multi-NLI dataset (Williams et al., 2018) into 14 languages by professional translators (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)).

Cross-lingual Question Answering (XQuAD). We rely on the XQuAD dataset (Artetxe et al., 2020b), created by translating the 240 dev paragraphs (from 48 documents) and corresponding 1,190 QA pairs of SQuAD v1.1 (Rajpurkar et al., 2016) to 11 languages (ES, DE, EL, RU, TR, AR, VI, TH, ZH, and HI). In order to allow for a comparison between zero-shot and few-shot transfer (see §4), we reserve 10 documents as the development set for our experiments and evaluate on the remaining 38 articles.[4]

Fine-tuning. For higher-level tasks, we perform standard downstream fine-tuning of LM-pretrained mBERT and XLM-R. For lower-level tasks, we instead freeze the transformer and train only task-specific classifiers.[5,6]

We add the following task-specific architectures on top of MMTs: for DEP we add the biaffine parsing head (Dozat and Manning, 2017; Kondratyuk and Straka, 2019); for POS, we attach a simple feed-forward token-level classifier; for NER, we feed MMT's token-level outputs to a CRF classifier, similar to Peters et al. (2017). For XNLI, we apply a simple softmax classifier on the vector of the sequence start token (`[CLS]` for mBERT; `<s>` for XLM-R); for XQuAD, we pool MMT's representations of all subwords and input it to a span classification head – a linear layer computing the start and the end of the answer.

Training and Evaluation Details. We experiment with mBERT *Base cased* and XLM-R *Base*, both with $L = 12$ transformer layers, hidden state size of $H = 768$, and $A = 12$ self-attention heads.

For XNLI, we limit the inputs to $T = 128$ subword tokens and train in batches of 32 instances. For XQuAD, we limit paragraphs to $T = 384$ tokens and questions to $Q = 64$ tokens. We slide over paragraphs with a window of 128 tokens and train in batches of size 12. For XNLI and XQuAD, we search the following hyperparameter grid: learning rate $\lambda \in \{5 \cdot 10^{-5}, 3 \cdot 10^{-5}\}$; training epochs $n \in \{2, 3\}$. For DEP, POS and NER, we fix the number of training epochs to 20. We train in batches of 32 sentences, with maximal length of $T = 512$ subword tokens. We optimize all models with Adam (Kingma and Ba, 2015).

We report DEP performance in terms of Unlabeled Attachment Scores (UAS).[7] For POS, NER, and XNLI we report accuracy, and for XQuAD, we report the Exact Match (EM) score.

3.2 Results and Preliminary Discussion

A summary of the zero-shot cross-lingual transfer results, per target language, is provided in Table 1. As expected, we observe drops in performance for all tasks and all target languages w.r.t. reference EN performance. However, the drops vary greatly across languages. For example, NER (mBERT) drops mere 2.6% for IT, but enormous 32% for AR; XNLI transfer (XLM-R) yields a moderate 6.1% drop for FR, but a large 20% drop for SW, etc.

At first glance, it appears – as suggested in prior work – that the transfer drops primarily correlate with language proximity: they are more pronounced for languages that are more distant from EN (e.g., JA, ZH, AR, TH, SW). While we see no notable exception to this in the three lower-level tasks, language proximity alone does not explain many

[4] As a general note, while the effects of "translationese" might have some impact on the absolute numbers (Artetxe et al., 2020a), they are not prominent enough to have any impact on the relative trends in the reported results (e.g., zero-shot vs. few-shot performance). For both XNLI and XQuAD, the translations were done completely manually and not via post-editing of MT (which would pose a higher "translationese" risk). Moreover, having an independently created test set in each language would impede comparability across languages.

[5] This gave slightly better performance than fine-tuning.

[6] We tokenize the input for each model with the corresponding pretrained fixed-vocabulary tokenizer: WordPiece tokenizer (Wu et al., 2016) with the vocabulary of 110K tokens for mBERT, and the SentencePiece BPE tokenizer (Sennrich et al., 2016) with the vocabulary of 250K tokens for XLM-R.

[7] Using Labeled Attachment Score (LAS) would make differences in annotation schemes between languages a confounding factor and impede our analysis of effects of language proximity and size of the target language corpora.

Task	Model	EN	ZH Δ	TR Δ	RU Δ	AR Δ	HI Δ	EU Δ	FI Δ	HE Δ	IT Δ	JA Δ	KO Δ	SV Δ	VI Δ	TH Δ	ES Δ	EL Δ	DE Δ	FR Δ	BG Δ	SW Δ	UR Δ
DEP	B	91.2	-43.9	-46.0	-28.1	-56.4	-36.1	-50.2	-30.7	-36.1	-17.1	**-60.1**	-56.1	-14.3	-	-	-	-	-	-	-	-	-
	X	92.0	**-85.4**	-44.2	-29.7	-54.6	-39	-49.5	-26.7	-39	-23.5	-80.5	-56.0	-16.3	-	-	-	-	-	-	-	-	-
POS	B	95.8	-38.0	-35.9	-16.0	-40.1	-33.4	-34.6	-21.9	-33.4	-19.8	**-46.1**	-42.0	-9.6	-	-	-	-	-	-	-	-	-
	X	96.3	-69.2	-27.7	-14.3	-37.1	-27.3	-31.9	-17.9	-27.3	-19.0	**-77.0**	-37.3	-10.7	-	-	-	-	-	-	-	-	-
NER	B	92.4	-23.3	-11.6	-10.7	**-31.7**	-11.1	-12.8	-3.8	-11.1	-2.6	-25.7	-13.8	-6.7	-	-	-	-	-	-	-	-	-
	X	91.6	**-34.8**	-6.2	-13.7	-24.6	-16.5	-8.0	-0.9	-16.5	-2.4	-30.1	-15.6	-2.2	-	-	-	-	-	-	-	-	-
XNLI	B	82.8	-13.6	-20.6	-13.5	-17.3	-21.3	-	-	-	-	-	-	-	-11.9	-28.1	-8.1	-14.1	-10.5	-7.8	-13.3	**-33.0**	-23.4
	X	84.3	-11.0	-11.3	-9.0	-13.0	-14.2	-	-	-	-	-	-	-	-9.7	-12.3	-5.8	-8.9	-7.8	-6.1	-6.6	**-20.2**	-17.3
XQuAD	B	71.1	-22.9	-34.2	-19.2	-24.7	-28.6	-	-	-	-	-	-	-	-22.1	**-43.2**	-16.6	-28.2	-14.8	-	-	-	-
	X	72.5	**-26.2**	-18.7	-15.4	-24.1	-22.8	-	-	-	-	-	-	-	-19.7	-14.8	-14.5	-15.7	-16.2	-	-	-	-

Table 1: Zero-shot cross-lingual transfer performance on five tasks (DEP, POS, NER, XNLI, and XQuAD) with mBERT (B) and XLM-R (X). We show the monolingual EN performance and report drops in performance relative to EN for all target languages. Numbers in bold indicate the largest zero-shot performance drops for each task.

of the XNLI and XQuAD results. For instance, RU XNLI (for both mBERT and XLM-R) is comparable to that of ZH, and lower than that for HI and UR: this is despite the fact that, as Indo-European languages, RU, HI, and UR are linguistically closer to EN than ZH. Similarly, we observe comparable performance on XQuAD for TH, RU, and ES.

3.3 Analysis

For each task, we now analyze the correlations between transfer performance and **a)** several measures of linguistic proximity (i.e., similarity) between languages and **b)** the size of MMT pretraining corpora of each target language.

Language Vectors and Corpora Sizes. For estimates of linguistic similarity, we rely on language vectors from LANG2VEC, which encode various linguistic features from the URIEL database (Littell et al., 2017). We consider the following LANG2VEC vectors: `syntax` (SYN) vectors encode syntactic properties, e.g., if a subject appears before or after a verb; `phonology` (PHON) vectors encode phonological properties such as the consonant-vowel ratio; `inventory` (INV) vectors denote presence or absence of natural classes of sounds (e.g., voiced uvulars); FAM vectors encode memberships in `language families`; and GEO vectors express orthodromic distances for languages w.r.t. fixed points on the Earth's surface. Language proximity is computed as cosine similarity between the languages' corresponding LANG2VEC vectors: each vector type (e.g., SYN) produces one similarity score (i.e., feature). We couple LANG2VEC features with the z-normalized size of the target language corpus used in MMT pretraining (SIZE).[8]

Correlation Analysis. We first correlate individual features with the zero-shot transfer scores for each task and show the results in Table 2. Quite intuitively, the zero-shot performance for low-level syntactic tasks – POS and DEP – highly correlates with syntactic language similarity (SYN). SYN also correlates well with transfer results for high-level tasks (except with XLM-R results on XQuAD). Somewhat surprisingly, the phonological language similarity (PHON) correlates best with transfer performance with XLM-R, for all tasks except XNLI, and also for mBERT on POS. For both high-level tasks and both MMTs, we observe very high correlations between transfer performance and size of pretraining corpora of the target language (SIZE). In contrast, SIZE exhibits lower correlations for lower-level tasks (DEP, POS, NER). We believe that this reflect the fact that high-level LU tasks rely on rich representations of semantic phenomena of a language, whereas low-level tasks require simpler structural representation of a language – it simply takes more distributional data to acquire the former than the latter.

Meta-Regression. Across the tasks, we observe high correlations between zero-shot transfer results and several features (e.g., SYN, PHON and SIZE). We next test if we can predict the transfer performance for a new language, by (linearly) combining individual features. For each task, we fit a linear regression using transfer results for target languages as labels. With only between 11 and 14 target languages (i.e., instances for fitting the regressor) per task, we resort to leave-one-out cross-validation (LOOCV) to obtain correlations for feature com-

[8] For XLM-R, we take reported sizes of language-specific CC-100 portions (Conneau et al., 2020); for mBERT, we work with sizes of language-specific Wikipedias.

Task	Model	SYN		PHON		INV		FAM		GEO		SIZE	
		P	S	P	S	P	S	P	S	P	S	P	S
DEP	XLM-R	0.77	0.78	**0.83**	**0.77**	0.46	-0.04	0.68	0.61	0.80	0.81	0.62	0.47
	mBERT	**0.92**	**0.91**	0.79	0.74	0.55	-0.01	0.76	0.62	0.64	0.69	0.79	0.59
POS	XLM-R	0.68	0.79	**0.81**	**0.81**	0.38	0.02	0.58	0.74	0.80	0.73	0.54	0.46
	mBERT	**0.90**	**0.87**	0.86	0.81	0.57	0.02	0.82	0.80	0.66	0.72	0.47	0.39
NER	XLM-R	0.49	0.49	**0.80**	**0.83**	0.27	0.14	0.47	0.55	0.77	0.81	0.37	0.35
	mBERT	0.60	0.74	**0.81**	**0.84**	0.34	-0.04	0.53	0.58	0.59	0.73	0.42	0.38
XNLI	XLM-R	**0.88**	**0.90**	0.29	0.27	0.31	-0.11	0.63	0.54	0.54	0.74	0.70	0.76
	mBERT	**0.87**	0.86	0.21	0.08	0.29	0.04	0.61	0.47	0.55	0.67	0.77	**0.91**
XQuAD	XLM-R	0.69	0.53	**0.85**	**0.81**	0.62	-0.01	**0.81**	0.54	0.43	0.50	**0.81**	0.55
	mBERT	0.84	0.89	0.56	0.48	0.55	0.22	0.79	0.64	0.51	0.55	**0.89**	**0.96**

Table 2: Correlations between zero-shot transfer performance with mBERT and XLM-R for different downstream tasks with linguistic proximity features (SYN, PHON, INV, FAM and GEO) and pretraining size of target-language corpora (SIZE). Results reported in terms of Pearson (P) and Spearman (S) correlation coefficients.

Task	Model	Selected features	P	S	MAE
POS	X	PHON (.75); GEO (.25)	0.77	0.75	10.99
	B	SYN (.99)	0.94	0.90	4.60
DEP	X	PHON (.25); SYN (.18) GEO (.57)	0.81	0.89	10.14
	B	SYN(.99)	0.93	0.92	5.77
NER	X	PHON(.99)	0.80	0.88	4.64
	B	PHON(.99)	0.69	0.82	9.45
XNLI	X	SYN (.51); SIZE (.49)	0.84	0.85	2.01
	B	SYN (.35); SIZE (.34), FAM (.31)	0.89	0.90	2.78
XQuAD	X	PHON (.99)	0.95	0.83	2.89
	B	SIZE (.99)	0.89	0.93	4.76

Table 3: Results of the meta-regression analysis, i.e., predicting zero-shot transfer performance for mBERT (B) and XLM-R (X). For each task-model pair we list only features with weights ≥ 0.01. P=Pearson; S=Spearman; MAE=Mean Absolute Error.

binations. We perform greedy forward feature selection: in each iteration we add the feature which boosts correlation (obtained via LOOCV) the most; we stop when none of the remaining features further improves the Pearson correlation.

We summarize the results of this meta-regression analysis in Table 3. For each task-model pair, we list features selected with the greedy feature selection and show (normalized) weights assigned to each feature. Except for NER, combinations of features manage to yield higher correlations with zero-shot transfer results than any of the features on their own. These results empirically confirm our previous intuition that linguistic proximity between the source and target language only partially explains zero-short transfer performance. On XNLI, transfer performance is best explained with the combination of structural similarity between languages (SYN) and the size of the target-language pretraining corpora (SIZE); on XQuAD with mBERT, SIZE alone best explains zero-short transfer scores. Note that the features are mutually quite correlated as well (e.g., languages closer to EN also tend to have larger pretraining corpora): thus if the regressor selects only one feature, this does not mean that other features do not correlate with transfer performance (as shown by Table 2).

The coefficients in Table 3 again indicate the importance of SIZE for the language understanding tasks and highlight our core finding: pretraining corpora sizes are stronger features for predicting zero-shot performance in higher-level tasks, whereas the results in lower-level tasks are more affected by typological language proximity.

4 From Zero to Hero: Few-Shot

Motivated by the low zero-shot transfer performance for many tasks and languages obtained in §3, we now investigate Q4 from §1: we aim to mitigate transfer losses with inexpensive few-shot cross-lingual transfer.

Experimental Setup. We rely on the same models, tasks, and evaluation protocols as described in §3.1. However, instead of fine-tuning the MMTs on task-specific data in EN only, we continue the fine-tuning process by feeding k additional training examples randomly chosen from reserved target language data portions, disjoint with the test sets.[9]

[9]Note that for XQuAD, we performed the split on the article level to avoid topical overlap. Consequently, for XQuAD

Task	Model	k	$k=10$		$k=50$		$k=100$		$k=500$		$k=1000$	
		$k=0$	score	Δ	score	Δ	score	Δ	score	Δ	score	Δ
DEP	MBERT	52.96	66.69	13.73	72.67	19.70	74.8	21.84	80.47	27.5	82.74	29.77
	XLM-R	48.60	65.57	16.97	72.19	23.59	74.08	25.48	81.16	32.56	83.33	34.73
POS	MBERT	67.2	80.17	12.96	85.34	18.14	87.09	19.88	91.16	23.96	92.64	25.44
	XLM-R	65.5	80.68	15.18	85.7	20.2	87.59	22.09	91.35	25.85	92.80	27.3
NER	MBERT	79.34	83.18	3.84	84.54	5.20	85.25	5.91	87.9	8.56	89.31	9.97
	XLM-R	85.43	88.06	2.63	91.07	5.64	91.49	6.06	93.69	8.26	93.82	8.39
XNLI	MBERT	65.92	65.89	-0.03	65.08	-0.84	64.92	-1.00	67.41	1.49	68.16	2.24
	XLM-R	73.32	73.73	0.41	73.76	0.45	75.03	1.71	75.34	2.02	75.84	2.52
			$k=2$		$k=4$		$k=6$		$k=8$		$k=10$	
XQUAD	MBERT	45.62	48.12	2.50	48.66	3.04	49.34	3.72	49.91	4.29	50.19	4.57
	XLM-R	53.68	53.73	0.05	53.84	0.17	54.76	1.08	55.56	1.88	55.78	2.10

Table 4: Results of the few-shot experiments with varying numbers of target-language examples k. For each k, we report performance averaged across languages and the difference (Δ) with respect to the zero-shot setting.

For our low-level tasks, we compare three sampling methods: (i) random sampling (RAND) of k target language sentences, (ii) selection of the k shortest (SHORTEST) and (iii) the k longest (LONGEST) sentences.[10] For XNLI and XQuAD, we run the experiments five times and report the average scores.

4.1 Results and Discussion

The results on each task, conditioned on the number of examples k and averaged across all target languages, are presented in Table 4. We note substantial improvements in few-shot learning setups for all tasks. However, the results also reveal notable differences between different types of tasks. For higher-level language understanding tasks the improvements are less pronounced; the maximum gains for XNLI and XQuAD after seeing $k = 1,000$ target-language instances and 10 articles, respectively, are between 2.52 (XLM-R) and 4.57 points (mBERT). On the other hand, the average gains for the lower-level tasks are massive: between 10 (NER) and 30 (DEP) points for mBERT and 8 (NER) and 35 (DEP) points for XLM-R. Moreover, the gains in all lower-level tasks are substantial even when we add only 10 annotated sentences in the target language (on average, up to 17 points on DEP, and 15 points on POS). What is more, our additional experiments (omitted for brevity) show substantial gains for DEP and POS even with fewer than 5 annotated target language sentences. A comparison of different sampling strategies for the lower-level tasks is shown in Figure 1 for mBERT.[11] For DEP and POS, the pattern is very clear and quite expected – adding longer sentences results in better scores. For NER, however, random sampling (RAND) appears to perform best: we hypothesize that this is because: (i) very long sentences are relatively sparse with named entities, resulting in our model seeing mostly negative examples; (ii) shorter sentences contribute less than for DEP and POS because they typically consist of (confirmed by manual inspection) a single named entity mention, without any non-NE tokens.

Figure 2 illustrates few-shot performance for individual languages on two lower-level (DEP, NER) and two higher-level tasks (XNLI, XQuAD), for different values of k.[12] Across languages, we see a clear trend – more distant target languages benefit much more from the few-shot data. Observe, e.g., SV for DEP or DE for XQuAD. Both are closely related to EN, exhibit high zero-shot transfer performance, and benefit only marginally from few in-language instances. We hypothesize that for such closely related languages, with enough pretraining data, MMT is able to extrapolate the missing language-specific knowledge from few in-language examples; its priors for languages close to EN are already quite sensible and *a priori* offer less room for improvements. In stark contrast, KO (DEP, a) and TH (XQuAD, b), for example, both exhibit poor zero-shot performance and understandably so, given their linguistic distance to EN. Given in-language data, however, both see rapid leaps in performance, displaying gains of almost 40% UAS on

k refers to the number of articles.

[10] In all three cases, we only choose between sentences with ≥ 3 and ≤ 50 tokens.

[11] A similar analysis for XLM-R is in the supplementary.

[12] We show per-language scores for POS with mBERT, and all tasks with XLM-R in the Appendix.

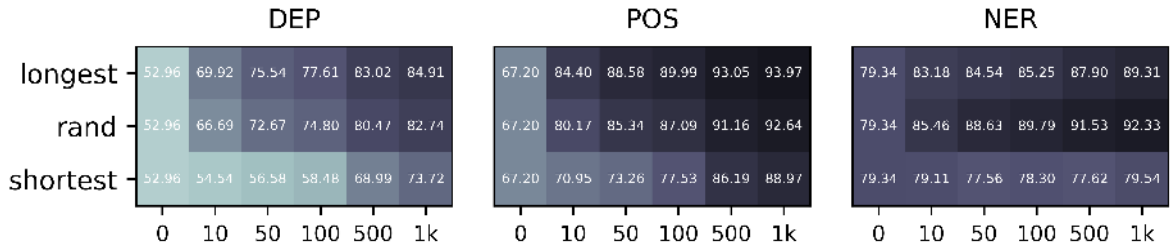
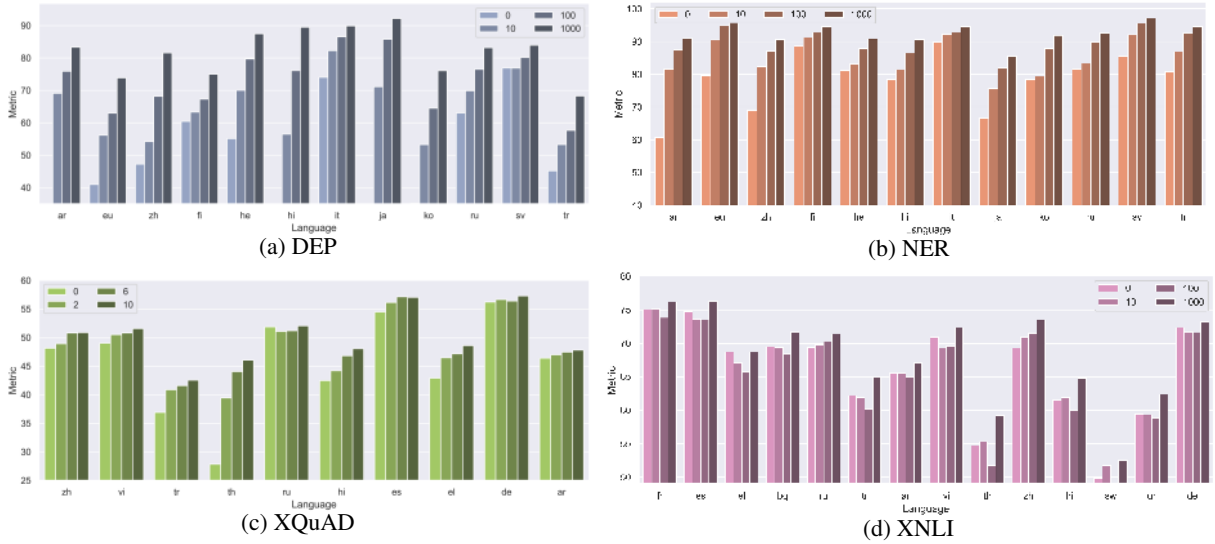


Figure 1: Heatmap of performance gains for low-level tasks from few-shot transfer with mBERT for different sampling strategies. X-axis: number of target-language instances k; Y-axis: sampling strategy.



(a) DEP (b) NER

(c) XQuAD (d) XNLI

Figure 2: Few-shot transfer results with mBERT for each language with varying k for two low-level tasks: a) DEP, b) NER, and two higher-level tasks: c) XQuAD, d) XNLI. For DEP, NER, and XNLI k denotes the number of sampled sentences, for XQuAD, the number of sampled articles.

DEP (KO), and almost 5% on XQuAD (TH). This can be seen as MMTs' ability to rapidly learn to utilize the multilingual space to adjust its task-specific knowledge for the target language. Other interesting patterns emerge. Particularly interesting are DEP results for JA and AR, where we observe massive UAS improvements with only 10 annotated sentences. For XQuAD, we observe a substantial improvement from only 2 in-language documents for TH. In sum, we see the largest gains from few-shot transfer exactly for languages for which the zero-shot transfer setup yields largest performance drops: languages distant from EN and represented with small corpora in MMT pretraining.

Direct Target Language Few-Shot Fine-Tuning. We have additionally run a set of control experiments in which we bypass the task-specific fine-tuning on the Enhlish data and directly fine-tune the MMTs on the few target language instances. Expectedly, for high-level LU tasks, fine-tuning the MMTs with only a handful of target language examples (i.e., *without* prior fine-tuning in English) yields subpar performance w.r.t. the corresponding model variant that had been previously fine-tuned on English data. For instance, direct few-shot target language fine-tuning of mBERT yields the average XNLI performance of 33.95 for $k = 100$ and 40.19 for $k = 1,000$, respectively (compared to 64.92 and 68.16, respectively, when prior fine-tuning on English data is performed). These findings suggest that fine-tuning with abundant (English) in-task data plus fine-tuning with scarce in-language in-task data yields a truly synergistic effect for higher-level language understanding tasks: the small number of examples in the target language is not sufficient to adapt the MMT directly, but they can provide a substantial edge over fine-tuning only on the English data (i.e., zero-shot transfer).

Somewhat surprisingly, however, for the simpler lower-level tasks, omitting task-specific fine-

Task	#inst.	Cost est.	Δ mBERT	Δ XLM-R
POS	1K sents	$73	+25.4	+27.3
DEP	1K sents	$280	+29.8	+34.7
NER	1K sents	$60	+10	+8.4
NLI	1K sent. pairs	$10	+2.24	+2.54
QA	10 docs	$30	+4.5	+2.1

Table 5: Conversion rates between target language annotation costs and corresponding average performance gains from MMT-based few-shot language transfer.

tuning on the English data and fine-tuning only on few target language instances does not lead to the major deterioration of performance (in fact, in some cases, omitting to fine-tune the MMTs on English data even slightly improves the results): for NER (mBERT) we obtain the average performance of 82.89 and 89.76 for $k = 100$ and $k = 1,000$ respectively, compared to 85.25 and 89.31 obtained respectively with prior English fine-tuning; for POS, the direct few-shot target language fine-tuning yields 87.08 ($k = 100$) and 92.64 ($k = 1,000$). We observe the same trends for the remaining tasks and with XLM-R. This suggests that MMTs can be fine-tuned for lower-level (i.e., simpler) tasks with only a handful of instances.

4.2 Cost of Language Transfer Gains

As shown in §4.1, moving to few-shot transfer can massively improve performance and reduce the gaps observed with zero-shot transfer, especially for low-resource languages. While additional fine-tuning on few target-language examples is computationally cheap, data annotation may be expensive, especially for minor languages. What are the annotation costs, and how do they translate into performance gains? Table 5 provides ballpark estimates for our five evaluation tasks; the estimates are based on annotation costs from the literature (Hovy et al., 2014; Tratz, 2019; Bontcheva et al., 2017; Marelli et al., 2014; Rajpurkar et al., 2016). We explain these cost-to-gain conversion estimates in more detail in Appendix C).

A provocative high-level question that calls for further discussion in future work can be framed as: are GPU hours effectively more costly[13] than data annotations are in the long run? While MMTs are extremely useful as general-purpose models of language, their potential for some (target) languages can be quickly unlocked by pairing them with a small number of annotated target-language examples. Effectively, this suggests leveraging the best of both worlds, i.e., coupling knowledge encoded in large MMTs with a small annotation effort.

[13]Financially, but also ecologically (Strubell et al., 2019).

5 Conclusion

Research on zero-shot language transfer in NLP is motivated by inherent data scarcity: the fact that most languages have no annotated data for most NLP tasks. Massively multilingual transformers (MMTs) have recently been praised for their zero-shot transfer capabilities that mitigate the data scarcity issue. In this work, we have demonstrated that, similar to earlier language transfer paradigms, MMTs perform poorly in zero-shot transfer to distant target languages, and to languages with smaller monolingual corpora available for exploitation in MMT pretraining. We have presented a detailed empirical analysis of factors affecting zero-shot transfer performance of MMTs across diverse tasks and languages. Our results have revealed that structural language similarity determines the transfer success for lower-level tasks like POS-tagging and dependency parsing; on the other hand, the pretraining corpora size of the target language is crucial for explaining transfer results for higher-level language understanding tasks, such as question answering and natural language inference.

Finally and most importantly, we have shown that the MMT potential on distant and low-resource target languages can be quickly unlocked if they are provided a handful of annotated instances in the target language. This finding provides a strong incentive for intensifying future research efforts that focus on cheap or naturally occurring supervision (Vulić et al., 2019; Artetxe et al., 2020c; Marchisio et al., 2020), quick and simple annotation procedure, and the more effective few-shot transfer learning setups.

Acknowledgements

The work of Anne Lauscher and Goran Glavaš has been supported by the Eliteprogramm of the Baden-Württemberg Stiftung (Grant "AGREE: Algebraic Reasoning over Events from Text and External Knowledge"). The work of Ivan Vulić is supported by the ERC Consolidator Grant LEXICAL (no 648909). We thank the anonymous reviewers for their suggestions and insightful comments.

References

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah A. Smith. 2016. Many languages, one parser. *Transactions of the Association for computational Linguistics*, 4:431–444.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of ACL*, pages 451–462.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020a. Translation artifacts in cross-lingual transfer learning. *CoRR*, abs/2004.04721.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020b. On the cross-lingual transferability of monolingual representations. In *Proceedings of ACL*, pages 4623–4637.

Mikel Artetxe, Sebastian Ruder, Dani Yogatama, Gorka Labaka, and Eneko Agirre. 2020c. A call for more rigor in unsupervised cross-lingual learning. In *Proceedings of ACL*, pages 7375–7388.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Emily M. Bender. 2011. On achieving and evaluating language-independence in NLP. *Linguistic Issues in Language Technology*, 6(3):1–26.

Kalina Bontcheva, Leon Derczynski, and Ian Roberts. 2017. Crowdsourcing named entity recognition and entity linking corpora. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*, pages 875–892.

José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence*, 240:36–64.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In *Proceedings of ICLR*.

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 ACL*, pages 8440–8451.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Proceedings of NeurIPS*, pages 7057–7067.

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 EMNLP*, pages 2475–2485.

Sandipan Dandapat, Priyanka Biswas, Monojit Choudhury, and Kalika Bali. 2009. Complex linguistic annotation—no easy way out!: A case from Bangla and Hindi POS labeling tasks. In *Proceedings of the 3rd Linguistic Annotation Workshop*, pages 10–18.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. In *Proceedings of ICLR*.

Steffen Eger, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2018. Cross-lingual argumentation mining: Machine translation (and a bit of projection) is all you need! In *Proceedings of COLING*, pages 831–844.

Karën Fort. 2016. *Collaborative Annotation for Reliable Natural Language Processing: Technical and Sociological Aspects*. John Wiley & Sons.

Dan Garrette and Jason Baldridge. 2013. Learning a part-of-speech tagger from two hours of annotation. In *Proceedings of NAACL-HLT*, pages 138–147.

Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulić. 2019. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In *Proceedings of ACL*, pages 710–721.

Goran Glavaš and Ivan Vulić. 2020. Is supervised syntactic parsing beneficial for language understanding? An empirical investigation. *CoRR*, abs/2008.06788.

Dirk Hovy, Barbara Plank, and Anders Søgaard. 2014. Experiments with crowdsourced re-annotation of a POS tagging data set. In *Proceedings of ACL*, pages 377–382.

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. In *Proceedings of ICML*.

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 ACL*, pages 6282–6293.

K Karthikeyan, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual bert: An empirical study. In *Proceedings of ICLR*.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of ICLR*.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing universal dependencies universally. In *Proceedings of EMNLP-IJCNLP*, pages 2779–2795.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, 6(2):167–195.

Jindřich Libovickỳ, Rudolf Rosa, and Alexander Fraser. 2020. On the language neutrality of pretrained multilingual representations. *arXiv preprint arXiv:2004.05160*.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of ACL*, pages 3125–3135.

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 EACL*, pages 8–14.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692*.

Kelly Marchisio, Kevin Duh, and Philipp Koehn. 2020. When does unsupervised machine translation work? *CoRR*, abs/2004.05516.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of LREC*, pages 216–223.

Stephen Mayhew, Chen-Tse Tsai, and Dan Roth. 2017. Cheap translation for cross-lingual named entity recognition. In *Proceedings of EMNLP*, pages 2536–2545.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics*, 2:231–244.

Nikola Mrkšić, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. 2017. Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the Association for Computational Linguistics*, 5:309–324.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Joakim Nivre, Željko Agić, Lars Ahrenberg, Lene Antonsen, Maria Jesus Aranzabe, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, and Liesbeth Augustinus. 2017. Universal Dependencies 2.1.

Matthew Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of ACL*, pages 1756–1765.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An adapter-based framework for multi-task cross-lingual transfer. In *Proceedings of EMNLP*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of ACL*, pages 4996–5001.

Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3):559–601.

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. Intermediate-task transfer learning with pretrained language models: When and why does it work? In *Proceedings of ACL*, pages 5231–5247.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of ACL*, pages 151–164.

Jonathan Raphael Raiman and Olivier Michel Raiman. 2018. DeepType: Multilingual entity linking by neural type system evolution. In *Proceedings of AAAI*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of EMNLP*, pages 2383–2392.

Samuel Rönnqvist, Jenna Kanerva, Tapio Salakoski, and Filip Ginter. 2019. Is multilingual BERT fluent in language generation? In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, pages 29–36.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Marta Sabou, Kalina Bontcheva, and Arno Scharl. 2012. Crowdsourcing research opportunities: Lessons from Natural Language Processing. In *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies*, pages 1–8.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of ACL*, pages 1715–1725.

Samuel L. Smith, David H.P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *Proceedings of ICLR*.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of ACL*, pages 778–788.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of ACL*, pages 3645–3650.

Stephen Tratz. 2019. Dependency Tree Annotation with Mechanical Turk. In *Proceedings of the First Workshop on Aggregating and Analysing Crowdsourced Annotations for NLP*, pages 1–5.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of NeurIPS*, pages 5998–6008.

Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. 2019. Do we really need fully unsupervised cross-lingual embeddings? In *Proceedings of EMNLP-IJCNLP*, pages 4398–4409.

Ivan Vulić, Sebastian Ruder, and Anders Søgaard. 2020. Are all good word vector spaces isomorphic? In *Proceedings of EMNLP*.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2019. CCNet: Extracting high quality monolingual datasets from Web crawl data. *CoRR*, abs/1911.00359.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of NAACL-HLT*, pages 1112–1122.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv*, abs/1910.03771.

Shijie Wu, Alexis Conneau, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Emerging cross-lingual structure in pretrained language models. In *Proceedings of ACL*.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of EMNLP-IJCNLP*, pages 833–844.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

A Reproducibility

We first provide details on where to obtain datasets and code used in this work.

Code and Dependencies. Our code can be obtained from https://www.dropbox.com/s/o5cxyy92re48xmu/zerohero_code.zip?dl=0. The code is separated in two parts: for experiments related to low-level tasks (DEP, POS, NER) the code is based on the AllenNLP framework; for the experiments on high-level tasks (XNLI, XQuAD), our code directly builds on top of the HuggingFace Transformers framework (Wolf et al., 2019). We provide links to code dependencies and pretrained models in Table 6.

Datasets. Table 7 provide links to all datasets that we used in our study, for each of the five tasks (low-level tasks: DEP, POS, NER; high-level tasks: XNLI, XQuAD).

B Full Per-Language Few-Shot Results

We show full per-language few-shot transfer results for all five tasks (DEP, POS, NER, XNLI, XSQuAD) for mBERT and XLM-R in Tables 8 and 9, respectively. We visually illustrate the gains from few-shot transfer for individual languages, for mBERT (for the POS task not covered in the main paper) in Figure 3 and for XLM-R (for all five tasks) in Figure 4. Finally, we show how the few-shot transfer results with XLM-R for lower-level tasks (DEP, POS, NER) depend on the instance sampling strategy (RAND, SHORTEST, LONGEST) in Figure 5.

C Few-Shot Transfer: Annotation Costs versus Performance Gains

We now present the more detailed explanations for the conversion between the annotation costs and few-shot transfer performance gains, summarized in Table 5 in the main paper.

Natural Language Inference. Marelli et al. (2014) reportedly paid $\$2,030$ for 200k judgements, which would amount to $\$0.01015$ per NLI instance and, in turn, to $\$10.15$ for $1,000$ annotations. In our few-shot experiments this would yield an average improvement of 2.24 and 2.52 accuracy points for mBERT and XLM-R, respectively. It is also possible to translate the English data directly via professional translation services as done with the XNLI dataset and XQuAD: the platforms for hiring professionals such as Upwork show that it is possible to find qualified translators even for lower-resource languages: e.g., the translation cost estimate for Zulu is \$12.5-\$16/h, or \$19/h for the Basque language.

Question Answering. Rajpurkar et al. (2016) report a payment cost of \$9 per hour and a time effort of 4 minutes per paragraph. With an average of 5 paragraphs per article, our few-shot scenario (10 articles) roughly requires 50 paragraphs-level annotations, i.e., 200 minutes of annotation effort and would in total cost around \$30 (for respective performance improvements of 4.6 and 2.1 points for mBERT and XLM-R).

On the one hand, compared to language understanding tasks, our lower-level (DEP, POS) tasks are presumably more expensive to annotate, as they require some linguistic knowledge and annotation training. On the other hand, as shown in our few-shot experiments, we typically need much fewer annotated instances (i.e., we observe high gains with already 10 target language sentences) for substantial gains in these tasks.

Dependency Parsing. Tratz (2019) provide an overview of crowd-sourcing annotations for dependency parsing; they report obtaining a fully correct dependency tree from at least one annotator for 72% of sentences. At the reported cost of \$0.28 per sentence this amounts to spending \$280 for annotating $1,000$ sentences. Somewhat shockingly, annotating 10 sentences with dependency trees – which for particular target languages like AR and JA corresponds to performance gains of 30-40 UAS points (see Figure 2) – amounts to spending merely \$3-5.

Part-of-Speech Tagging. Hovy et al. (2014) measure agreement of crowdsourced POS annotations with expert annotations; they crowdsource annotations for 1,000 tweets, at a cost of \$0.05 for every 10 tokens. With a total of $14,619$ tokens in the corpus, this amounts to approximately \$73 for $1,000$ tweets, which is $\geq 1,000$ sentences.[14] Based on Table 4, 2 hours of POS annotation work translates to gains of up to 20-22 points on average over zero-shot transfer methods.

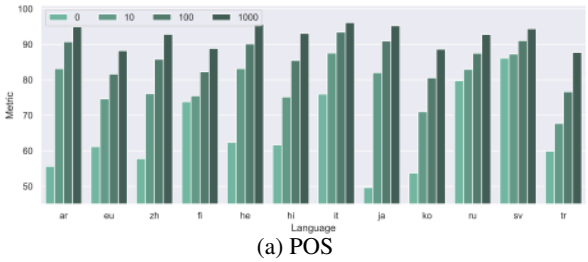
[14] Note, however, that lower-level tasks do come with an additional risk of poorer quality annotation, due to crowdsourced annotators not being experts. Garrette and Baldridge (2013) report that even for truly low-resource languages (e.g., Kinyarwanda, Malagasy), it is possible to obtain $\approx$ 100 POS-annotated sentences in 2 hours.

Codebase	MMT	**Vocab**	**Params**	**URL**
Allen NLP	–	–	–	https://github.com/allenai/allennlp
HF Trans.	–	–	–	https://github.com/huggingface/transformers
	mBERT	119K	125M	https://huggingface.co/bert-base-multilingual-cased
	XLM-R	250K	125M	https://huggingface.co/xlm-roberta-base

Table 6: Links to codebases and pretrained models used in this work. For low-level tasks (DEP, POS, NER), we carried out our experiments using the AllenNLP library. For high-level tasks (XNLI, XQuAD), we built our models directly on top of the HuggingFace (HF) Transformers library.

Task	**Dataset**	**URL**
Dependency Parsing	UD	https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3105
POS Tagging	UPOS	https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3105
Named Entity Recognition	WikiAnn	https://elisa-ie.github.io/wikiann/
Natural Language Inference	XNLI	https://github.com/facebookresearch/XNLI
Question Answering	XQuAD	https://github.com/deepmind/xquad

Table 7: Links to the datasets used in our work.



(a) POS

Figure 3: Graphical illustration of few-shot transfer gains for each language with mBERT, for the remaining task not covered in the main paper: POS.

Named Entity Recognition. Bontcheva et al. (2017) provide estimates for crowdsourcing annotation for named entity recognition; they pay \$0.06 per sentence, resulting in \$60 cost for $1,000$ annotated sentences. At a median pay of \$11.37/hr, this amounts to around 190 sentences annotated in an hour. In other words, in less than 3 hours, we can collect more than 500 annotated examples. According to Table 4, this can result in gains of 8+ points on average, and even more for some languages (e.g., 27 points for AR).

POS	ar	eu	zh	fi	he	hi	it	ja	ko	ru	sv	tr
0	55.65	61.19	57.8	73.85	62.38	61.7	76.02	49.65	53.75	79.79	86.15	59.9
10	83.16	74.65	76.1	75.5	83.18	75.19	87.56	82.04	71.02	82.95	87.28	67.73
50	89.18	79.84	83.84	81.4	88.91	83.12	92.04	88.27	77.17	86.07	89.5	74.2
100	90.73	81.63	85.82	82.28	90.12	85.46	93.47	90.95	80.57	87.5	91.06	76.66
500	94.08	86.84	90.78	86.8	94.75	89.69	95.73	94.25	86.48	91.21	93.43	85.29
1000	94.97	88.23	92.83	88.86	95.7	93.09	96.15	95.24	88.64	92.77	94.39	87.72

NER	ar	eu	zh	fi	he	hi	it	ja	ko	ru	sv	tr
0	60.69	79.53	69.01	88.59	81.26	78.46	89.77	66.64	78.51	81.64	85.62	80.78
10	81.69	90.51	82.27	91.28	83.12	81.44	92.14	75.64	79.36	83.39	92.09	86.91
50	86.3	93.36	85.6	92.38	87.02	85.04	92.34	78.88	86.94	88.07	95.51	91.93
100	87.37	94.84	87.19	92.88	87.8	86.52	92.79	81.98	88	89.98	95.53	92.5
500	89.74	95.28	89.5	94.01	89.86	89.27	93.8	84.6	90.93	92.18	96.84	94.34
1000	90.92	96.01	90.71	94.57	90.8	90.67	94.5	85.62	91.96	92.71	97.17	94.65

DEP	ar	eu	zh	fi	he	hi	it	ja	ko	ru	sv	tr
0	34.72	40.96	47.25	60.44	55.1	33.59	74.05	31.03	35.11	63.03	76.9	45.17
10	69.08	56.16	54.18	63.3	70.02	56.49	82.26	71.12	53.25	69.89	76.88	53.26
50	73.65	61.11	64.39	65.88	78.78	71.48	84.46	82.58	61.11	73.95	79.37	56.78
100	75.91	62.98	68.17	67.31	79.71	76.1	86.53	85.77	64.51	76.51	80.13	57.66
500	81.48	70.33	78.64	71.4	84.81	85.34	89.39	90.38	73.65	81.19	82.87	65.16
1000	83.31	73.85	81.59	74.97	87.47	89.49	89.9	92.18	76.08	83.18	83.95	68.26

XNLI	fr	es	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	de
0	75.05	74.71	68.68	69.50	69.34	62.18	65.53	70.88	54.69	69.26	61.50	49.84	59.38	72.34
10	75.09	73.62	67.04	69.35	69.80	61.86	65.56	69.26	55.30	70.89	61.92	51.79	59.28	71.63
50	74.60	73.91	66.44	68.37	69.05	60.99	64.63	70.29	51.17	71.32	60.08	49.95	58.83	71.43
100	73.85	73.50	65.67	68.47	70.24	60.13	64.93	69.59	51.68	71.46	60.01	48.96	58.78	71.60
500	75.36	74.97	68.04	71.03	70.59	63.21	66.71	72.38	58.12	72.81	64.06	52.26	61.15	73.09
1000	76.20	76.24	68.73	71.73	71.41	65.01	67.04	72.35	59.19	73.47	64.75	52.47	62.38	73.21

XQuAD	zh	vi	tr	th	ru	hi	es	el	de	ar
0	48.14	49.02	36.90	27.84	51.86	42.47	54.48	42.90	56.22	46.40
2	48.93	50.50	40.87	39.43	51.07	44.19	56.14	46.46	56.66	46.99
4	49.72	51.38	40.22	41.24	51.33	45.90	56.62	47.25	56.38	46.57
6	50.81	50.81	41.59	44.04	51.20	46.81	57.14	47.16	56.40	47.45
8	51.53	51.29	41.99	45.28	51.29	47.10	57.45	47.95	57.07	48.21
10	50.87	51.57	42.55	46.05	52.05	48.06	57.03	48.60	57.29	47.82

Table 8: Detailed per-language few-shot language results with mBERT for different number of target-language data instances k. For low-level tasks, we report results with RAND sampling.

POS	ar	eu	zh	fi	he	hi	it	ja	ko	ru	sv	tr
0	59.23	64.41	27.06	78.34	68.94	65.63	77.25	19.28	58.98	81.96	85.54	68.61
10	82.72	76.54	68.3	81.04	84.81	77.08	88.44	78.92	70.5	83.95	87.87	72.33
50	89.14	80.19	77.49	84.94	89.13	84.07	92.51	86.94	76.09	87.29	90.8	79.19
100	90.67	83.38	80.83	86.44	90.3	87.23	93.52	88.78	78.91	88.84	91.79	81.65
500	94.36	88.4	86.61	90.23	94.23	91.4	95.7	92.11	84.37	91.87	94.35	87.64
1000	95.29	89.66	88.86	91.87	95.31	94.26	96.18	93.49	86.88	93.19	95.41	89.71

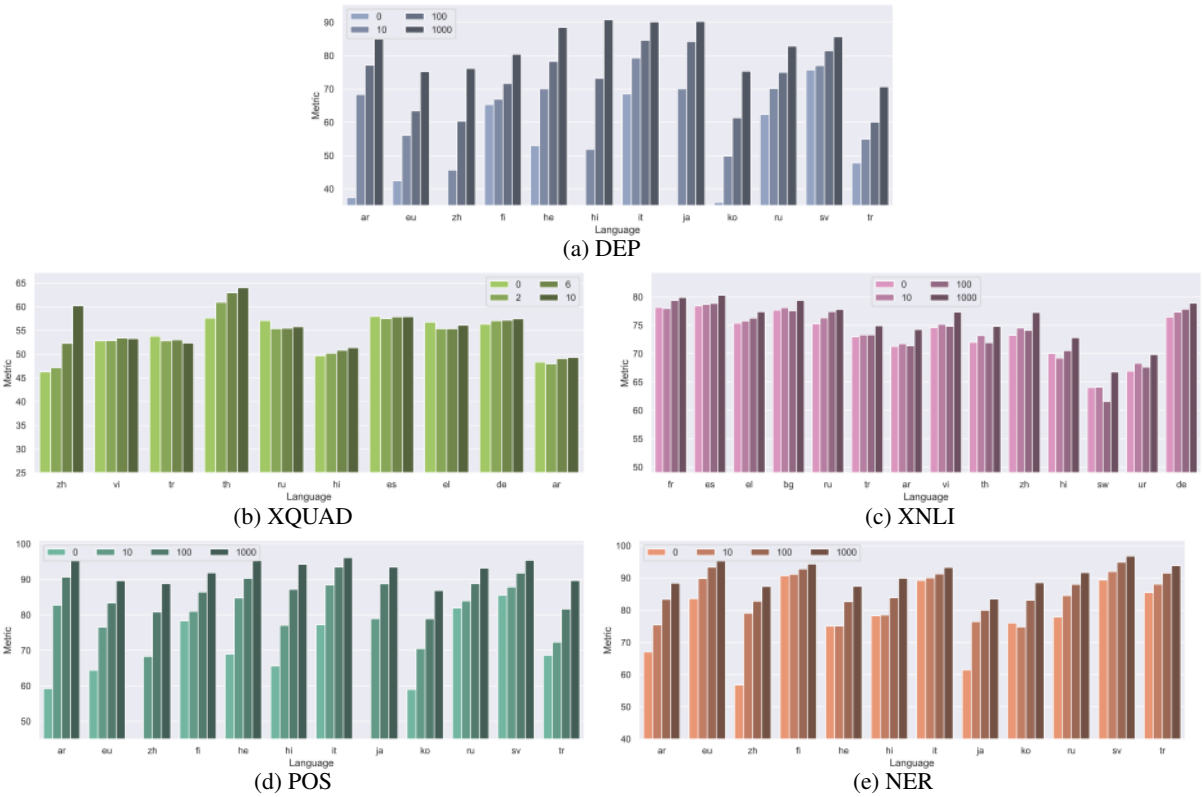
NER	ar	eu	zh	fi	he	hi	it	ja	ko	ru	sv	tr
0	67.03	83.58	56.77	90.69	75.05	78.28	89.25	61.46	76	77.87	89.36	85.43
10	75.45	89.81	79.02	91.14	75.1	78.5	90.02	76.45	74.8	84.5	92.01	88.06
50	82.56	91.63	80.81	92.01	80.34	81.23	91.01	78.13	81.8	87.21	94.72	91.07
100	83.37	93.33	82.77	92.77	82.63	83.88	91.23	79.97	83.06	88.01	94.89	91.49
500	86.95	94.82	85.77	93.78	86.09	87.79	92.44	82.38	87.17	91.02	96.33	93.69
1000	88.36	95.24	87.34	94.3	87.4	89.87	93.25	83.45	88.52	91.66	96.78	93.82

DEP	ar	eu	zh	fi	he	hi	it	ja	ko	ru	sv	tr
0	37.46	42.48	6.61	65.33	53.06	32.94	68.54	11.48	36	62.37	75.72	47.83
10	68.37	56.09	45.67	66.97	70.06	51.93	79.32	70.05	49.88	70.14	77.03	54.93
50	74.9	60.92	57.39	71.35	77.95	67.09	83.97	81.64	59.22	73.55	78.72	59.77
100	77.15	63.46	60.33	71.65	78.27	73.2	84.63	84.3	61.37	75.03	81.52	60.06
500	83.29	72.37	71.52	77.22	86.21	87.06	88.82	88.83	73.1	80.41	85.38	68.88
1000	84.99	75.25	76.2	80.46	88.48	90.81	90.14	90.28	75.35	82.88	85.68	70.68

XNLI	fr	es	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	de	
0	84.25	78.16	78.44	75.39	77.68	75.25	72.99	71.28	74.59	72	73.21	70.02	64.03	66.93	76.45
10	84.26	77.96	78.67	75.77	78.11	76.32	73.31	71.75	75.17	73.18	74.53	69.23	64.09	68.32	77.32
50	84.39	78.69	79.81	76.13	77.57	76.16	73.96	71.2	75.01	71.74	74.47	69.84	61.98	68.06	77.6
100	83.64	79.37	78.87	76.28	77.58	77.42	73.31	71.4	74.83	71.94	74.1	70.54	61.55	67.63	77.84
200	81.57	79.29	79.84	77.01	78.94	77.54	74.81	73.22	76.52	73.91	76.37	71.54	64	68.98	78.42
500	82.69	79.65	79.95	77.34	79.09	77.78	74.08	73.6	77.22	74.32	77.03	71.75	65.37	68.85	78.71
1000	83.74	79.91	80.29	77.39	79.39	77.8	74.92	74.26	77.34	74.8	77.26	72.83	66.77	69.84	78.91

XQuAD	zh	vi	tr	th	ru	hi	es	el	de	ar
0	46.29	52.84	53.82	57.64	57.10	49.67	57.97	56.77	56.33	48.36
2	47.16	52.86	52.84	60.96	55.39	50.20	57.51	55.37	57.05	47.97
4	48.06	53.43	51.88	61.57	54.21	50.28	57.62	55.68	56.72	49.00
6	52.29	53.41	53.03	62.97	55.48	50.85	57.88	55.37	57.16	49.10
8	57.88	53.49	52.47	63.73	55.87	50.96	58.25	55.83	57.05	50.09
10	60.22	53.28	52.36	64.02	55.79	51.38	57.90	56.11	57.47	49.30

Table 9: Detailed per-language few-shot language results with XLM-R for different number of target-language data instances k. For low-level tasks, we report results with RAND sampling.



(a) DEP

(b) XQUAD

(c) XNLI

(d) POS

(e) NER

Figure 4: Graphical illustration of few-shot transfer gains for individual languages, for XLM-R and all languages.

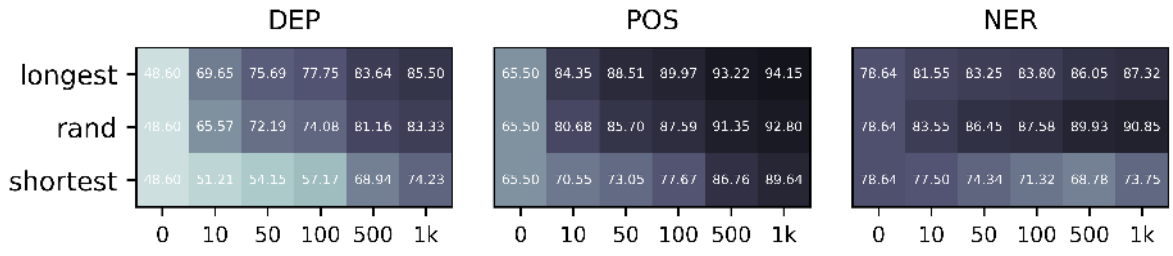


Figure 5: Heatmap of performance gains for low-level tasks from few-shot transfer with XLM-R for different sampling strategies. X-axis: number of target-language instances k; Y-axis: sampling strategy.