SLICER: Sliced Fine-Tuning for Low-Resource Cross-Lingual Transfer for Named Entity Recognition

Fabian David Schmidt[1], Ivan Vulić[2], Goran Glavaš[1]
[1] Center For Artificial Intelligence and Data Science, University of Würzburg, Germany
[2] Language Technology Lab, University of Cambridge, UK
{fabian.schmidt, goran.glavas}@uni-wuerzburg.de
iv250@cam.ac.uk

Abstract

Large multilingual language models generally demonstrate impressive results in zero-shot cross-lingual transfer, yet often fail to successfully transfer to low-resource languages, even for token-level prediction tasks like named entity recognition (NER). In this work, we introduce a simple yet highly effective approach for improving zero-shot transfer for NER to low-resource languages. We observe that NER fine-tuning in the source language decontextualizes token representations, i.e., tokens increasingly attend to themselves. This increased reliance on token information itself, we hypothesize, triggers a type of overfitting to properties that NE tokens within the source languages share, but are generally *not present* in NE mentions of target languages. As a remedy, we propose a simple yet very effective *sliced fine-tuning* for NER (SLICER) that forces stronger token contextualization in the Transformer: we divide the transformed token representations and classifier into disjoint slices that are then independently classified during training. We evaluate SLICER on two standard benchmarks for NER that involve low-resource languages, WikiANN and MasakhaNER, and show that it (i) indeed reduces decontextualization (i.e., extent to which NE tokens attend to themselves), consequently (ii) yielding consistent transfer gains, especially prominent for low-resource target languages distant from the source language.

1 Introduction

In recent years, massively multilingual transformers (MMTs) have become the backbone of multilingual NLP. MMTs like mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), and mT5 (Xue et al., 2021), pretrained on corpora spanning 100+ languages, have become the main vehicle of cross-lingual transfer in NLP: fine-tuned using task annotations in the source language, an MMT can, conceptually, directly make predictions for all target languages seen in pretraining, for which little or no annotated task data exists (Pires et al., 2019; Wu and Dredze, 2019; Dufter and Schütze, 2020).

Successful zero-shot transfer, however, has been shown to critically hinge on linguistic proximity between source and target languages as well the quality of target language representations, determined by the size of target language corpora used in the MMT pretraining (Lauscher et al., 2020; Zhao et al., 2021). Unfortunately, the transfer fails where it is needed the most – for low-resource languages linguistically distant from high-resource languages with annotated task data (Ebrahimi et al., 2021; Adelani et al., 2021; Ruder et al., 2021b). Zero-shot transfer of named entity recognition (NER) models to low-resource languages suffers from a particularly profound performance drop (Adelani et al., 2021; Lauscher et al., 2020). In this work, we identify the cause and propose an effective remedy.

Contributions. (1) We analyze the representation space of an MMT, before and after source-language NER fine-tuning, and discover that it decreases token contextualization in higher Transformer layers: after fine-tuning, tokens generally learn to put much more *attention* to themselves. Put differently, monolingual NER benefits from limiting higher-layer contextualization, which, we believe results with encoding more information from the token itself and less from the context. While this may be beneficial for monolingual NER, where NE tokens share features (e.g., capitalization, morphemes), we believe it has a negative effect in cross-lingual transfer, given that NE mentions in target languages generally do not exhibit the same features. **(2)** We devise a novel *sliced* fine-tuning Transformer-based NER (SLICER): we split the transformed token vectors into disjoint segments and classify each independently with a different subset of classification parameters. We show that SLICER leads to increased contextualization in higher Transformer layers. This, as our empirical evaluation on two established multilingual NER benchmarks shows

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 10775 - 10785

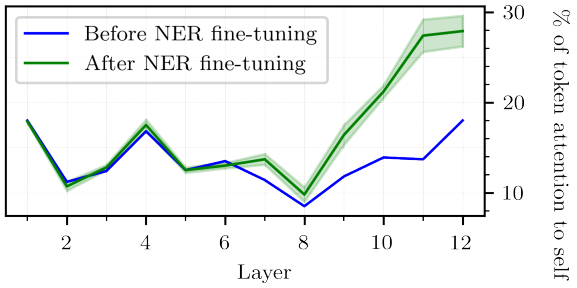

Figure 1: Average proportion of the attention that tokens put on themselves, across all Transformer layers, before and after monolingual NER fine-tuning of XLM-R on the English portion of WikiANN. For the latter, we display mean and standard deviation across ten fine-tuning runs. The proportions are averaged across sentences from the WikiANN English test set.

(Pan et al., 2017; Adelani et al., 2021), leads to substantially better transfer performance, especially for low-resource languages distant from the source.

Our work shows that, despite the task-agnostic nature of the current MMT-based paradigm for cross-lingual transfer, task-specific traits can be exploited to yield substantial performance gains. We hope it catalyzes more work on *task-specific* approaches to cross-lingual transfer with MMTs.

2 Token Contextualization in NER

Decontextualization in Monolingual NER. We first fine-tune one of the most widely used MMT models, XLM-R (Conneau et al., 2020),[1] on the English training portion of WikiANN (Pan et al., 2017), a widely used multilingual NER dataset. Figure 1 shows the average proportion of the attention that tokens in each Transformer layer place on themselves, before (i.e., for vanilla XLM-R) and after standard NER fine-tuning.[2] The proportion of attention that tokens put on themselves in vanilla XLM-R varies between 10 and 20% across the layers, with largest proportions in the first and last layer. The behavior of the corresponding XLM-R fine-tuned on English NER closely mirrors this behavior in the lower Transformer layers. In the higher layers – parameters of which change the most through NER fine-tuning – tokens start placing much more attention to themselves than in the vanilla XLM-R. The gap is particularly pronounced from the 9th layer onwards and amounts to roughly 10% more attending-to-self in the last layer, compared to the pretrained XLM-R. This suggests that monolingual NER favors (or, more precisely, requires) reduced contextualization in higher Transformer layers. This effectively means that the Transformer places more focus on token information itself, which, we hypothesize, leads to more similar representations for tokens with similar properties, regardless of their context. In monolingual NER, this is arguably beneficial because, within the same language, NE tokens generally share many token-level properties (e.g., morpho-syntax and capitalization). Because of this, the same decontextualization effect, we argue, should have a detrimental effect in cross-lingual transfer to target languages in which NE tokens generally do not share token-level properties with NE tokens of the source language.

Sliced Fine-Tuning for Cross-Lingual NER. We next devise a novel fine-tuning approach that forces the Transformer to retain more contextualization, especially in its higher layers. Given a sequence of input tokens $t_{i=1,\ldots,N}$, let $v_{t_i} \in \mathbb{R}^d$ be the contextualized representation of the i-th token, output of the last Transformer layer. In standard fine-tuning for token-level tasks, contextualized token representations are forwarded into a classifier parameterized by a linear layer $W \in \mathbb{R}^{d \times |C|}$ and a bias $b \in \mathbb{R}^{|C|}$, which project v_{t_i} into a vector of log-probabilities, one for each class $c \in C$.

Due to the observed decontextualization (see again Figure 1), we believe that NER fine-tuning on source language data leads to representations that are mutually more correlated (across tokens of same NE classes) for combinations of features that predominantly encode token-level information, and not contextual information. This, as discussed, is beneficial for monolingual NER performance, but we suspect is detrimental to cross-lingual NER transfer. Aiming to decorrelate token representations, we propose *sliced fine-tuning* for NER (SLICER), in which we slice transformed token vectors v_t into d/h disjoint subsequences (i.e., smaller vectors) of size h *during training*, where h can be any integer divisor of d: $\{v_t^{(i)}\}_{i=1}^{d/h}$ (cf. Figure 2).[3] We then accordingly slice the classifier's matrix W along the primary dimension, resulting in the classification tensor $W_S \in$

[1] We use `xlm-roberta-base` weights from the HuggingFace Transformers library.

[2] We first average the attention probability for each token on itself across attention heads. We then average the token-level scores across all subwords in a sequence, as attention probabilities depend on the sequence length. We lastly average the sequence-level scores over all sequences in the test set.

[3] Note that $h = d$ reduces SLICER to standard fine-tuning.

$\mathbb{R}^{\frac{d}{h} \times h \times |C|} = \{W_S^{(i)} \in \mathbb{R}^{h \times |C|}\}_{i=1}^{d/h}$. Each token vector slice is then independently classified by the corresponding slice of the classification tensor: $y^{(i)} = softmax(v_t^{(i)} \cdot W_S^{(i)} + b)$. We then compute the standard cross-entropy loss per slice and update both the classifier's and Transformer's parameters by minimizing the average of slice losses.[4]

Consider $d = 768$ with $h = 2$ as illustrated in Figure 2: SLICER learns to pool 768-dimensional token embeddings, output of the last attention layer, to 384 (i.e., $\frac{d}{h}$) *slices* $\in \mathbb{R}^2$ in the last feed-forward layer. SLICER then computes the loss independently by slice (i.e., on subsequent pairs of features) and averages those slice losses $\{\mathcal{L}\}_{i=1}^{384}$ into the final loss. Such training forces the Transformer to self-sufficiently compress the information to classify NE into 2 features (for $h = 1$ into 1 feature, for $h = 8$ into 8 features, etc.) as correlations between features across slices cannot be exploited by design. The low capacity of slices further disables the model to retain simplest token-level cues (e.g., casing, suffixes) that discern between different NE classes within a language. We hypothesize that the model is thereby coerced to embed more contextual information in token vectors v_t: since SLICER erodes token-idiosyncratic features, the class-independent dissimilarity between token representations decreases. This effect then propagates further backwards through the Transformer and materializes in increased contextualization via higher attention over surrounding tokens. Such improved contextualization then results in token representations that are more robust to distributional shift arising from language and domain transfer.

3 Evaluation

Experimental Setup. Unless stated differently, we train on the English training portion of WikiANN (Pan et al., 2017). We then evaluate SLICER against standard fine-tuning (Standard FT) in zero-shot transfer to (i) 23 target languages from WikiANN and (ii) 10 African languages from MasakhaNER (Adelani et al., 2021).[5] We fine-tune XLM-R (Base) with mixed precision, using AdamW with 0.05 weight decay (Loshchilov and Hutter, 2019) for optimization. We train with three different learning rates $\{5e^{-6}, 1e^{-5}, 2e^{-5}\}$ with 10% linear warmup and subsequent decay, for each setup of which we execute 10 fine-tuning runs with different random seeds. We apply 10% dropout and train in batches of size 32 for 10 epochs.

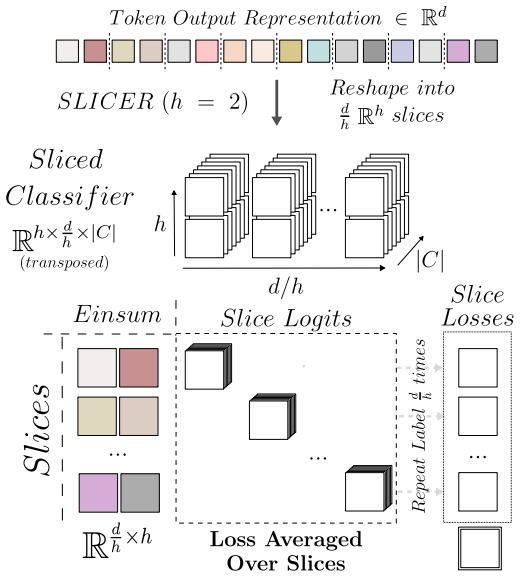

Figure 2: SLICER illustrated on a single token: During training, the token output representation $v_t \in \mathbb{R}^d$ of the transformer and $|C|$ class column vectors for each class of the classification head $W \in \mathbb{R}^{d \times |C|}$ are first reshaped from $\mathbb{R}^d$ to $\mathbb{R}^{\frac{d}{h} \times h}$, i.e. $\frac{d}{h}$ h-dimensional slices. SLICER then computes a loss $\{\mathcal{L}\}_{i=1}^{\frac{d}{h}}$ for each token slice and averages slices losses to a joint loss.

We compare SLICER against Standard FT under two common evaluation procedures for cross-lingual transfer: (1) TRUE zero-shot transfer strictly assumes that there are no labeled instances in the target language; (2) ORACLE transfer assumes that a small validation set in the target language is available for model selection: in this setting, we select the model checkpoint that yields the best target language validation performance. For SLICER, we report the results for three different values of slice size, $h \in \{1, 2, 8\}$.

Results. Table 1 displays the performance of SLICER against standard fine-tuning, for three different learning rates, on MasakhaNER and WikiANN (mean and std. deviation aggregates across all 10 and 23 target languages, respectively; we present detailed per-language results in the Appendix §A.3). Our sliced fine-tuning outperforms standard source language training across the board (for both benchmarks, both evaluation protocols,

[4] We only slice the token representations and classifier matrix W during training. At inference, we we use the whole token vector v_t and the whole matrix W, which can be viewed as ensembling over slices.

[5] We follow Ansell et al. (2022) and remap both B-DATE and I-DATE of MasakhaNER to O (0) to harmonize NER tags with WikiANN.

		WIKIANN				MASAKHANER				TOTAL			
		TRUE		ORACLE		TRUE		ORACLE		TRUE		ORACLE	
LR	Model	ø	σ	ø	σ	ø	σ	ø	σ	ø	σ	ø	σ
$5e^{-6}$	Standard FT	56.8	±1.6	58.6	±1.5	23.4	±2.0	28.7	±3.5	47.8	±1.7	49.8	±2.1
	SLICER, $h = 1$	57.2	±1.5	58.7	±1.5	**32.6**	±3.3	34.8	±2.9	50.7	±2.0	51.7	±1.9
	SLICER, $h = 2$	57.2	±1.4	58.8	±1.3	32.4	±2.8	35.1	±2.8	50.6	±1.8	51.9	±1.8
	SLICER, $h = 8$	57.3	±1.5	58.9	±1.5	32.3	±3.2	34.1	±3.5	50.7	±2.0	51.6	±2.1
$1e^{-5}$	Standard FT	56.1	±1.9	59.2	±1.7	23.1	±2.4	28.3	±3.2	47.2	±2.0	50.1	±2.1
	SLICER, $h = 1$	**57.7**	±1.7	59.8	±1.5	31.7	±3.3	35.9	±3.6	50.8	±2.1	52.8	±2.1
	SLICER, $h = 2$	57.6	±1.6	**59.9**	±1.9	32.2	±3.0	36.0	±3.4	50.9	±2.0	**52.9**	±2.4
	SLICER, $h = 8$	**57.7**	±1.7	59.8	±2.0	32.2	±3.6	35.5	±4.2	**51.0**	±2.2	52.6	±2.7
$2e^{-5}$	Standard FT	54.6	±2.2	58.8	±1.6	22.0	±2.1	26.7	±3.3	45.8	±2.1	49.3	±2.1
	SLICER, $h = 1$	56.8	±2.2	59.5	±2.6	29.4	±3.0	36.1	±4.4	49.6	±2.4	52.6	±3.1
	SLICER, $h = 2$	56.6	±2.2	59.5	±2.6	29.8	±3.9	**37.1**	±3.5	49.5	±2.7	**52.9**	±2.9
	SLICER, $h = 8$	56.6	±2.5	59.6	±2.1	28.4	±4.0	35.9	±5.1	49.1	±2.9	52.6	±3.0

Table 1: Zero-shot cross-lingual transfer performance (micro-averaged F1) for NER with English as the source language. We report averages and standard deviations across all evaluated languages of MasakhaNER and WikiANN.

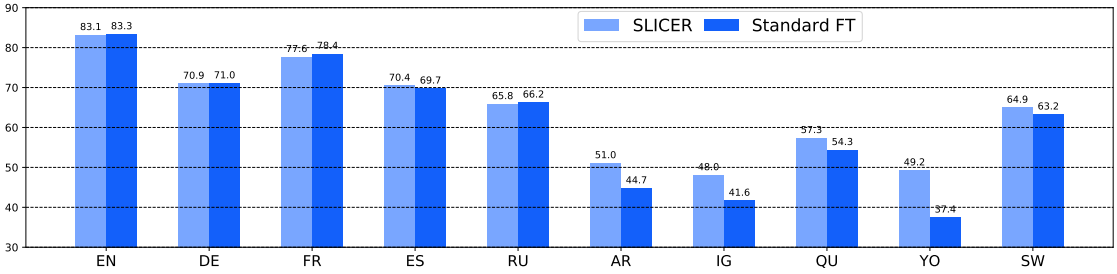

Figure 3: Comparison of cross-lingual transfer performance of SLICER ($h = 8$, $lr = 1e^{-5}$) against Standard FT for 5 high-resource Indo-European languages (EN, DE, FR, ES, and RU) and 5 non-Indo-European languages (AR, IG, QU, YO, and SW) from WikiANN. SLICER yields prominent gains on the latter group.

and all three learning rates): the gains are substantially more modest for WikiANN (between 0.5% for the smaller learning rate and 2% for the largest). On MasakhaNER, SLICER consistently yields impressive gains of around 8 F_1 points.

Delving deeper into per-language WikiANN results in Figure 3 reveals that SLICER performs on a par with Standard FT for English (source) and high-resource Indo-European target languages (e.g., German, French), but in most cases substantially outperforms Standard FT for low-resource languages from other language families. Figure 3 illustrates this, showing results on five high-resource Indo-European languages (English itself, German, French, Spanish, and Russian) and five non-Indo-European target languages (mostly low-resource: Arabic, Igbo, Quechua, Yoruba, and Swahili).

Given that we train the models on the English training portion of WikiANN (i.e., Wikipedia texts) and that MasakhaNER consists of sentences from newswire texts, transferring to MasakhaNER test sets represents not only language but also domain transfer. We believe such a setup exacerbates even more the differences between train and test distributions of token-level information of NE tokens (i.e., test sentences are even more out-of-distribution for the model) then in the case of language shift alone. The positive effect of SLICER w.r.t. to this additional domain shift becomes obvious when comparing the gap in performance (SLICER vs. Standard FT) for languages present in both MasakhaNER and WikiANN:[6] e.g., for Igbo (IG, IBO; see the Appendix A.3), the moderate edge of 6-8% F_1 points that SLICER has over Standard FT on WikiANN widens to enormous 16-19% advantage on MasakhaNER. [7] The fact that SLICER achieves the largest gains exactly in chal-

[6] WikiANN and MasakhaNER overlap in the following languages: Amharic (AM), Igbo (IG), Kinyarwanda (RW), Swahili (SW), and Yoruba (YO).

[7] It is also worth noting that WikiANN probably overestimates the absolute zero-shot cross-lingual transfer performance for NER, considering that the target-language portions were obtained by linking mentions from other languages to an English knowledge base (Lignos et al., 2022).

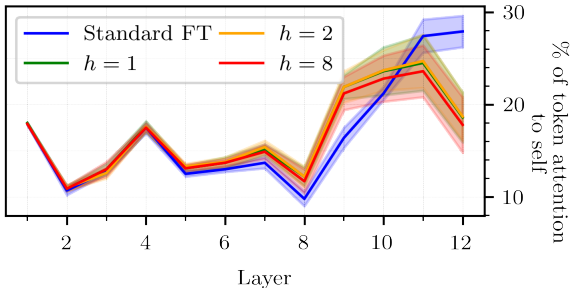

Figure 4: Average weights of tokens attending to themselves (mean and deviation across ten different fine-tuning runs), across all Transformer layers, after sliced source language fine-tuning (SLICER, shown for three different configurations, $h \in \{1, 2, 8\}$), compared against standard fine-tuning (Standard FT).

	WIKIANN TRUE		MASAKHANER TRUE	
Model	ø	σ	ø	σ
Standard FT	52.5	±1.7	32.8	±2.6
SLICER, $h = 1.0$	54.0	±1.9	37.7	±3.2

Table 2: Zero-shot cross-lingual transfer performance (micro-averaged F1) for NER with Russian as the source language for $lr = 1e^{-5}$. Numbers denote averages and standard deviations across all evaluated languages.

lenging transfer to hand-labeled MasakhaNER test sets corroborates our hypothesis that SLICER reduces reliance on context-independent token-level information and forces the Transformer to encode more information from the context.

Further Analyses. We next test the hypothesis that SLICER prevents token decontextualization in Transformer layers, which is present with Standard FT (cf., Figure 1). Figure 4 displays the average proportion of attention mass that tokens place on themselves after *sliced* fine-tuning. We observe that SLICER indeed reduces the amount of attending-to-self in higher Transformer layers: this means that more attention is placed on other tokens, corresponding to stronger contextualization.

Finally, to verify that our findings are not limited to English as the source language, we rerun all our experiments with another source language: Russian. Table 2 summarizes the aggregate cross-lingual transfer results with RU as the source (detailed results in the Appendix). We note the same trends as before (EN as source, cf. Table 1): SLICER outperforms Standard FT on both datasets, with substantially larger gains on MasakhaNER.

4 Conclusion

In this focused research effort, we show that (monolingual) fine-tuning for NER introduces token decontextualization in higher Transformer layers which, we hypothesize, has a negative effect on (zero-shot) cross-lingual NER transfer with MMTs. We devise a novel sliced fine-tuning approach, dubbed SLICER, that reduces this decontextualization effect by splitting transformed token vectors into disjoint slices which are then independently classified. We demonstrate on WikiANN and MasakhaNER that this yields substantial transfer gains, especially when transferring to low-resource languages. We additionally show that gains do not stem from a particular choice of source language. Our work shows that, despite the task-agnostic nature of the predominant MMT-based cross-lingual transfer paradigm, task specificities can still be leveraged to improve the cross-lingual transfer.

Limitations and Ethical Considerations

The main limitation of our work stems from the fact that the benefits of its methodological proposal are limited to a single task: Named Entity Recognition. This is in contrast with the vast majority of existing work that aims to improve the multilingual representation spaces and consequently boost downstream transfer performance across a wide range of tasks (Ruder et al., 2021a). We have preliminarily investigated the effects of sliced fine-tuning in cross-lingual transfer for other sequence labeling tasks, namely part of speech tagging and event trigger extraction. For those tasks, however, we observed (i) much less decontextualization (i.e., smaller increase in average attention-to-self proportions) after source language fine-tuning. and (ii) its presence in fewer Transformer layers (last or last two layers). Our sliced fine-tuning thus does not bring any substantial gains compared to Standard FT on those tasks.

Acknowledgements

We thank the state of Baden-Württemberg for its support through access to the bwHPC. Fabian David Schmidt and Goran Glavaš were supported by the EUINACTION grant from NORFACE Governance (462-19-010, GL950/2-1). Ivan Vulić is supported by a personal Royal Society University Research Fellowship (no 221137; 2022-2027) as well as a Huawei research donation to the Language Technology Lab.

References

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 2021. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

Alan Ansell, Edoardo Ponti, Anna Korhonen, and Ivan Vulić. 2022. Composable sparse fine-tuning for cross-lingual transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1778–1796, Dublin, Ireland. 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.

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.

Philipp Dufter and Hinrich Schütze. 2020. Identifying elements essential for BERT's multilinguality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4423–4437, Online. Association for Computational Linguistics.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir, Gustavo A. Gim énez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando A. Coto Solano, Ngoc Thang Vu, and Katharina Kann. 2021. Americasnli: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. *CoRR*, abs/2104.08726.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Constantine Lignos, Nolan Holley, Chester Palen-Michel, and Jonne Sälevä. 2022. Toward more meaningful resources for lower-resourced languages. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 523–532, Dublin, Ireland. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie

Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021a. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, et al. 2021b. Xtreme-r: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness 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 833–844.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Mengjie Zhao, Yi Zhu, Ehsan Shareghi, Ivan Vulić, Roi Reichart, Anna Korhonen, and Hinrich Schütze. 2021. A closer look at few-shot crosslingual transfer: The choice of shots matters. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5751–5767, Online. Association for Computational Linguistics.

A Appendix

A.1 Further Reproducibility Details

Hardware & Infrastructure. We train our models on a cluster that provides virtual machines on which each model was trained on a single NVIDIA Tesla V100 32GB GPU. Each model (incl. evaluation) requires a runtime of c.1.5 hrs, on average.

Additional Hyperparameters. We train on 10 random seeds ($\{42, ..., 51\}$) as set by Pytorch Lightning's `seed_everything`. For other hyperparameters, please refer to §3.

Code. Our implementation is publicly available at `https://github.com/fdschmidt93/SLICER`.

Language	ISO code	Validation	Test
Afrikaans	af	1000	1000
Amharic	am	100	100
Aymara	ay	100	100
Bulgarian	bg	10000	10000
German	de	10000	10000
Greek	el	10000	10000
English	en	10000	10000
French	fr	10000	10000
Hebrew	he	10000	10000
Hindi	hi	1000	1000
Japanese	ia	100	100
Igbo	ig	100	100
Japanese	ja	10000	10000
Quechua	qu	100	100
Russian	ru	10000	10000
Rwanda	rw	100	100
Swahili	sw	1000	1000
Tamil	ta	1000	1000
Telegu	te	1000	1000
Turkish	tr	10000	10000
Urdu	ur	1000	1000
Vietnamese	vi	10000	10000
Yoruba	yo	100	100
Chinese	zh	10000	10000

Table 3: WikiANN: list of languages included in our experiments.

A.2 List of Target Languages

We access both WikiANN and MasakhaNER via the Huggingface `datasets` library (Lhoest et al., 2021). Table 3 and 4 list the number of sentences for validation and testing by language.

Language	ISO code	Validation	Test
Amharic	am	250	500
Hausa	hau	272	545
Igbo	ibo	319	638
Kinyarwanda	kin	301	604
Luganda	lug	200	401
Luo	luo	92	185
Nigerian-Pidgin	pcm	300	600
Swahili	kin	300	602
Wolof	wol	267	536
Yoruba	yo	303	608

Table 4: MasakhaNER: list of languages included in our experiments.

A.3 Full Results By Target Language

A.3.1 MasakhaNER

		AMH				HAU				IBO				KIN				LUG			
		TRUE		ORACLE		TRUE		ORACLE		TRUE		ORACLE		TRUE		ORACLE		TRUE		ORACLE	
LR	Model	ø	σ	ø	σ	ø	σ	ø	σ	ø	σ	ø	σ	ø	σ	ø	σ	ø	σ	ø	σ
$5e^{-6}$	Standard FT	32.4	±1.5	35.7	±2.6	38.7	±1.6	40.3	±2.6	14.5	±2.9	25.9	±4.7	11.4	±1.6	16.0	±4.0	14.0	±2.0	19.3	±2.9
	SLICER, $h=1$	33.6	±1.9	36.8	±2.3	44.9	±2.4	46.7	±2.8	33.5	±3.4	36.7	±3.6	20.6	±4.2	22.7	±3.3	22.3	±3.6	23.9	±2.7
	SLICER, $h=2$	33.7	±2.0	37.5	±2.2	45.1	±2.4	46.5	±2.6	33.5	±3.3	38.3	±3.2	20.3	±3.5	23.1	±2.9	21.9	±3.0	24.6	±2.1
	SLICER, $h=8$	33.6	±2.6	37.0	±3.7	45.4	±2.6	47.0	±1.9	33.4	±3.7	37.0	±2.6	20.8	±3.5	22.6	±2.2	22.2	±3.0	23.2	±2.2
$1e^{-5}$	Standard FT	32.9	±1.6	35.9	±2.5	39.4	±1.8	41.3	±2.2	13.7	±3.8	25.2	±5.2	11.4	±2.6	15.9	±2.7	12.7	±3.0	18.7	±3.5
	SLICER, $h=1$	32.0	±2.3	36.3	±3.1	46.0	±3.3	47.9	±2.4	31.8	±3.8	37.2	±3.3	21.0	±3.2	25.7	±4.6	21.7	±3.4	27.0	±4.5
	SLICER, $h=2$	31.2	±1.8	36.5	±2.7	46.2	±2.5	47.6	±3.2	33.1	±3.1	38.1	±2.6	21.2	±3.6	25.2	±4.3	22.4	±4.0	27.0	±5.3
	SLICER, $h=8$	31.8	±2.5	39.0	±3.2	45.6	±2.2	47.1	±2.9	32.3	±4.1	37.1	±4.9	20.6	±4.7	23.4	±5.2	21.3	±3.4	24.1	±5.1
$2e^{-5}$	Standard FT	29.9	±3.1	33.6	±1.9	38.6	±2.1	41.8	±1.9	12.4	±2.7	20.1	±5.2	10.0	±1.6	14.7	±3.8	11.6	±1.7	18.3	±3.9
	SLICER, $h=1$	29.2	±2.5	33.9	±4.3	44.2	±1.6	48.4	±3.0	28.1	±4.1	37.4	±5.2	18.4	±2.8	24.1	±4.7	20.4	±2.8	26.2	±3.1
	SLICER, $h=2$	28.3	±3.0	34.8	±3.0	44.4	±2.2	49.0	±2.5	27.8	±6.2	38.8	±3.6	18.3	±3.1	26.3	±3.8	20.0	±3.1	26.5	±2.7
	SLICER, $h=8$	28.1	±2.4	33.5	±3.5	43.7	±2.5	47.1	±1.8	25.5	±5.7	38.0	±4.7	16.6	±3.8	25.0	±6.6	18.6	±3.8	26.0	±6.1

		LUO				PCM				SWA				WOL				YOR			
		TRUE		ORACLE		TRUE		ORACLE		TRUE		ORACLE		TRUE		ORACLE		TRUE		ORACLE	
LR	Model	ø	σ	ø	σ	ø	σ	ø	σ	ø	σ	ø	σ	ø	σ	ø	σ	ø	σ	ø	σ
$5e^{-6}$	Standard FT	0.3	±1.8	14.7	±2.0	39.1	±1.2	41.7	±2.2	48.3	±1.4	51.2	±1.8	10.6	±2.1	19.0	±5.5	14.4	±3.5	23.6	±6.1
	SLICER, $h=1$	7.1	±2.5	17.7	±2.2	41.5	±1.9	42.4	±2.8	53.0	±2.1	54.8	±2.9	25.9	±5.1	30.3	±3.1	33.4	±5.8	35.8	±3.7
	SLICER, $h=2$	7.5	±1.3	18.0	±1.9	41.3	±1.6	42.7	±1.9	53.5	±1.5	55.4	±2.0	25.1	±4.4	30.0	±4.7	32.2	±5.2	35.0	±4.2
	SLICER, $h=8$	7.3	±1.2	17.5	±3.7	41.8	±1.6	43.1	±2.1	53.2	±2.3	55.4	±2.7	25.1	±5.5	26.9	±7.0	30.5	±6.0	31.3	±6.9
$1e^{-5}$	Standard FT	0.0	±1.6	13.1	±1.1	39.9	±1.8	42.1	±2.7	48.6	±2.1	51.5	±2.7	9.7	±1.7	17.2	±2.9	12.8	±4.1	21.6	±6.1
	SLICER, $h=1$	5.8	±3.0	19.1	±4.3	42.1	±1.3	44.5	±4.0	52.7	±2.0	55.8	±1.9	21.9	±4.8	31.6	±5.3	32.4	±5.4	33.9	±3.1
	SLICER, $h=2$	6.5	±2.3	20.2	±2.6	42.8	±1.6	43.9	±2.3	52.9	±1.5	56.1	±2.4	24.1	±4.3	30.7	±3.5	31.9	±5.6	34.9	±5.4
	SLICER, $h=8$	7.1	±3.9	18.7	±3.0	42.5	±1.8	43.6	±3.0	52.6	±1.7	54.3	±2.7	24.0	±5.7	32.0	±6.0	34.2	±6.1	36.0	±5.7
$2e^{-5}$	Standard FT	9.7	±1.5	13.6	±1.8	39.3	±1.6	40.5	±3.5	48.2	±1.7	52.5	±1.9	9.2	±1.2	14.6	±3.7	11.4	±3.3	16.8	±5.4
	SLICER, $h=1$	4.7	±2.5	19.6	±4.1	42.0	±1.6	46.1	±3.8	51.7	±2.2	55.8	±2.7	17.7	±4.2	32.4	±7.4	28.1	±5.7	37.2	±5.4
	SLICER, $h=2$	5.0	±3.1	20.6	±2.9	43.0	±1.3	46.4	±3.1	51.9	±1.7	56.6	±3.0	19.8	±7.0	33.7	±6.2	29.7	±8.4	38.3	±4.3
	SLICER, $h=8$	3.2	±3.1	20.2	±4.6	42.9	±2.3	47.5	±3.1	51.0	±1.9	54.4	±2.7	16.3	±6.3	31.2	±8.8	28.2	±7.7	35.8	±9.1

A.3.2 WikiANN

		AM				AR				AY				BG			
		TRUE		ORACLE		TRUE		ORACLE		TRUE		ORACLE		TRUE		ORACLE	
LR	Model	ø	σ	ø	σ	ø	σ	ø	σ	ø	σ	ø	σ	ø	σ	ø	σ
$5e^{-6}$	Standard FT	43.9	±2.3	44.8	±3.3	49.1	±1.6	51.5	±2.3	40.8	±0.9	41.3	±0.5	78.4	±0.5	78.5	±0.5
	SLICER, $h=1$	46.8	±4.1	48.0	±3.7	51.7	±2.3	53.9	±2.2	41.7	±0.4	40.8	±1.6	78.4	±0.4	78.5	±0.4
	SLICER, $h=2$	46.9	±3.8	47.9	±2.5	52.6	±1.6	55.3	±1.7	41.6	±0.3	41.6	±0.9	78.5	±0.5	78.6	±0.6
	SLICER, $h=8$	47.9	±1.9	48.7	±1.7	52.0	±3.3	54.6	±3.5	41.4	±0.8	41.4	±1.1	78.5	±0.4	78.6	±0.4
$1e^{-5}$	Standard FT	43.0	±3.0	44.5	±2.7	44.7	±1.5	50.3	±2.0	37.2	±2.1	41.6	±4.7	78.9	±0.3	79.1	±0.4
	SLICER, $h=1$	43.4	±3.0	47.2	±2.2	51.3	±2.0	54.7	±2.8	40.2	±1.7	40.9	±1.6	79.2	±0.4	79.2	±0.4
	SLICER, $h=2$	42.6	±2.2	47.8	±2.8	50.6	±3.0	54.9	±2.7	41.8	±1.0	45.6	±12.3	78.9	±0.5	79.3	±0.4
	SLICER, $h=8$	44.4	±2.2	45.4	±3.5	51.0	±3.8	54.2	±3.0	41.2	±1.2	44.5	±9.9	79.1	±0.4	79.3	±0.3
$2e^{-5}$	Standard FT	40.6	±3.1	43.4	±2.3	42.0	±2.3	46.8	±3.3	32.4	±4.3	40.4	±0.8	78.2	±0.6	79.0	±0.5
	SLICER, $h=1$	41.9	±3.4	44.6	±4.6	46.6	±4.0	51.1	±4.7	38.9	±2.6	45.3	±15.1	79.0	±0.6	79.3	±0.5
	SLICER, $h=2$	41.9	±3.4	45.2	±4.2	46.8	±4.1	52.3	±4.5	39.6	±2.3	46.0	±13.7	79.0	±0.8	79.4	±0.6
	SLICER, $h=8$	41.8	±4.5	45.0	±3.9	47.5	±3.5	54.0	±2.4	38.9	±3.8	40.4	±2.9	78.8	±0.5	79.1	±0.5

		DE				EL				EN				ES			
		TRUE		ORACLE		TRUE		ORACLE		TRUE		ORACLE		TRUE		ORACLE	
LR	Model	ø	σ	ø	σ	ø	σ	ø	σ	ø	σ	ø	σ	ø	σ	ø	σ
$5e^{-6}$	Standard FT	70.4	±0.7	70.5	±0.7	76.1	±0.5	76.3	±0.5	82.3	±0.2	82.3	±0.2	68.6	±3.0	70.2	±3.4
	SLICER, $h=1$	69.9	±0.5	70.4	±0.6	75.5	±0.5	75.7	±0.5	82.0	±0.1	82.0	±0.1	66.9	±3.3	68.3	±3.4
	SLICER, $h=2$	70.1	±0.4	70.4	±0.7	75.5	±0.7	75.7	±0.7	82.1	±0.2	82.1	±0.2	66.8	±2.3	68.5	±2.1
	SLICER, $h=8$	70.1	±0.9	70.6	±0.8	75.8	±0.4	76.0	±0.3	82.1	±0.2	82.1	±0.2	67.1	±3.7	68.6	±3.9
$1e^{-5}$	Standard FT	71.0	±0.7	71.3	±0.6	75.2	±0.6	76.0	±0.5	83.3	±0.1	83.3	±0.1	69.7	±2.2	72.6	±2.2
	SLICER, $h=1$	70.7	±0.7	71.2	±0.7	75.8	±0.5	76.0	±0.6	83.1	±0.1	83.1	±0.1	71.6	±4.0	73.5	±3.0
	SLICER, $h=2$	70.9	±0.7	71.1	±0.6	75.9	±0.6	76.2	±0.5	83.2	±0.2	83.2	±0.2	70.2	±4.1	71.3	±3.1
	SLICER, $h=8$	70.9	±0.4	71.4	±0.6	75.2	±1.2	75.7	±0.8	83.1	±0.1	83.1	±0.1	70.4	±3.4	71.8	±4.0
$2e^{-5}$	Standard FT	70.6	±0.5	71.4	±0.6	72.9	±1.5	75.4	±0.6	83.5	±0.1	83.5	±0.2	69.0	±2.2	73.5	±1.6
	SLICER, $h=1$	70.8	±0.7	71.2	±0.7	74.5	±1.0	74.9	±1.0	83.8	±0.1	83.8	±0.1	73.3	±2.3	74.6	±2.2
	SLICER, $h=2$	71.2	±0.3	71.4	±0.4	74.3	±1.1	74.6	±1.0	83.8	±0.2	83.8	±0.2	72.9	±3.0	73.8	±2.8
	SLICER, $h=8$	71.0	±0.6	71.5	±0.7	73.9	±1.0	75.0	±1.4	83.8	±0.2	83.8	±0.2	71.4	±3.9	72.3	±3.3

		FR				HE				HI				IG			
		TRUE		ORACLE		TRUE		ORACLE		TRUE		ORACLE		TRUE		ORACLE	
LR	Model	ø	σ	ø	σ	ø	σ	ø	σ	ø	σ	ø	σ	ø	σ	ø	σ
$5e^{-6}$	Standard FT	77.5	±0.7	77.6	±0.8	54.6	±0.8	54.8	±0.8	68.1	±0.8	68.6	±0.7	41.5	±1.9	45.6	±1.8
	SLICER, $h = 1$	76.3	±0.8	76.5	±0.9	54.1	±0.5	54.3	±0.5	67.8	±0.9	68.2	±0.8	49.2	±1.5	50.1	±1.6
	SLICER, $h = 2$	76.3	±0.7	76.4	±0.8	54.3	±0.6	54.5	±0.7	67.7	±1.0	68.4	±1.0	49.1	±1.9	49.7	±1.0
	SLICER, $h = 8$	76.3	±0.8	76.6	±0.8	54.3	±0.7	54.5	±0.6	67.9	±0.6	68.0	±1.1	49.7	±1.9	50.3	±1.6
$1e^{-5}$	Standard FT	78.4	±0.6	78.7	±0.8	55.2	±0.9	55.6	±1.0	67.8	±0.7	68.9	±1.3	41.6	±2.1	45.8	±2.4
	SLICER, $h = 1$	77.8	±1.3	78.5	±0.9	54.7	±0.9	55.4	±0.7	66.8	±1.3	68.1	±1.4	48.8	±2.3	50.3	±1.8
	SLICER, $h = 2$	77.4	±1.2	77.6	±1.0	55.0	±0.8	55.4	±1.0	67.4	±1.1	68.2	±1.3	49.7	±2.0	51.5	±1.3
	SLICER, $h = 8$	77.6	±1.2	77.8	±1.2	55.2	±1.5	55.5	±1.2	67.7	±0.9	68.5	±1.0	48.0	±4.2	50.2	±3.1
$2e^{-5}$	Standard FT	78.1	±0.9	79.0	±0.6	53.2	±0.7	54.2	±0.9	66.3	±0.8	68.0	±1.7	40.6	±3.6	44.1	±3.2
	SLICER, $h = 1$	78.7	±1.0	78.9	±1.3	53.4	±1.4	54.2	±1.8	65.9	±1.8	67.3	±1.9	46.6	±2.1	49.8	±2.4
	SLICER, $h = 2$	78.3	±1.3	78.4	±1.3	53.3	±1.3	54.0	±1.4	66.0	±1.4	67.0	±1.5	46.4	±2.7	49.6	±2.6
	SLICER, $h = 8$	77.9	±1.1	78.2	±1.1	53.5	±1.4	54.8	±1.8	66.5	±1.3	68.2	±1.7	45.6	±3.7	49.7	±3.0

		JA				QU				RU				RW			
		TRUE		ORACLE		TRUE		ORACLE		TRUE		ORACLE		TRUE		ORACLE	
LR	Model	ø	σ	ø	σ	ø	σ	ø	σ	ø	σ	ø	σ	ø	σ	ø	σ
$5e^{-6}$	Standard FT	13.9	±1.2	14.3	±1.4	55.0	±2.4	54.1	±2.4	65.4	±0.4	65.6	±0.5	54.7	±3.2	53.9	±3.4
	SLICER, $h = 1$	13.7	±2.5	14.2	±2.5	54.8	±2.6	53.9	±3.1	64.0	±0.5	64.4	±0.7	53.9	±2.5	54.3	±2.8
	SLICER, $h = 2$	13.9	±1.9	14.4	±2.1	54.7	±2.0	54.6	±1.8	63.9	±0.9	64.1	±1.0	53.5	±2.6	54.0	±1.6
	SLICER, $h = 8$	14.4	±2.1	14.9	±2.2	54.8	±2.4	53.9	±2.1	63.9	±0.5	64.1	±0.6	52.9	±2.3	52.7	±2.5
$1e^{-5}$	Standard FT	15.4	±1.7	16.0	±1.5	54.3	±3.3	54.7	±3.8	66.2	±0.6	66.9	±0.8	57.1	±2.7	57.7	±3.1
	SLICER, $h = 1$	15.9	±1.9	16.1	±1.9	56.3	±2.2	55.8	±1.9	66.0	±0.8	66.2	±0.9	55.5	±2.5	57.3	±1.8
	SLICER, $h = 2$	15.5	±1.5	16.0	±1.7	56.8	±1.7	56.1	±2.3	65.8	±0.8	66.3	±0.9	55.3	±3.4	55.6	±1.8
	SLICER, $h = 8$	15.8	±1.9	16.5	±2.4	57.3	±1.6	56.9	±2.3	65.8	±0.8	66.1	±0.8	55.1	±2.9	55.6	±1.8
$2e^{-5}$	Standard FT	16.2	±1.1	17.7	±1.5	55.3	±2.1	55.6	±2.2	65.6	±1.1	66.2	±1.1	58.6	±2.4	58.1	±3.0
	SLICER, $h = 1$	16.9	±2.5	18.2	±2.8	55.3	±3.0	55.2	±2.5	66.1	±0.9	66.3	±0.8	58.4	±3.0	56.8	±4.3
	SLICER, $h = 2$	17.4	±2.0	18.4	±2.0	55.4	±2.1	55.1	±2.6	66.0	±1.2	66.3	±1.1	54.2	±2.3	55.4	±3.3
	SLICER, $h = 8$	18.1	±2.2	19.5	±2.1	55.6	±2.4	55.9	±2.3	65.9	±0.8	66.6	±1.2	56.9	±4.8	57.8	±4.7

		SW				TA				TE				TR			
		TRUE		ORACLE		TRUE		ORACLE		TRUE		ORACLE		TRUE		ORACLE	
LR	Model	ø	σ	ø	σ	ø	σ	ø	σ	ø	σ	ø	σ	ø	σ	ø	σ
$5e^{-6}$	Standard FT	64.7	±0.5	64.8	±0.6	60.1	±0.7	60.2	±1.1	54.1	±1.2	55.5	±1.6	69.2	±1.4	69.5	±1.3
	SLICER, $h = 1$	64.7	±0.8	64.9	±0.8	57.9	±0.6	58.3	±0.7	52.8	±1.1	54.0	±1.2	68.9	±0.8	69.8	±0.6
	SLICER, $h = 2$	64.7	±1.1	64.6	±0.8	57.9	±0.5	58.2	±0.8	52.6	±1.0	54.1	±1.7	69.0	±1.1	69.7	±1.0
	SLICER, $h = 8$	64.7	±1.5	64.8	±1.1	57.9	±0.8	58.2	±1.3	52.6	±0.9	54.7	±2.2	69.2	±1.2	70.1	±0.8
$1e^{-5}$	Standard FT	63.2	±1.6	64.5	±0.9	59.7	±0.7	60.6	±1.4	53.2	±1.0	55.6	±1.3	68.8	±1.9	69.7	±1.8
	SLICER, $h = 1$	63.9	±2.3	64.6	±1.4	57.8	±1.5	58.6	±1.8	52.0	±1.3	53.5	±2.0	69.0	±1.9	69.8	±1.3
	SLICER, $h = 2$	64.6	±1.7	65.0	±1.0	58.2	±0.8	58.9	±0.9	52.5	±1.4	54.3	±1.8	68.3	±1.7	69.3	±1.6
	SLICER, $h = 8$	64.9	±1.1	65.0	±1.0	57.6	±1.4	58.3	±1.6	52.0	±1.3	53.5	±1.6	68.7	±1.3	69.4	±1.2
$2e^{-5}$	Standard FT	60.1	±2.8	63.7	±1.9	57.3	±2.0	59.7	±1.1	50.5	±0.8	53.6	±1.6	67.4	±1.6	69.7	±1.3
	SLICER, $h = 1$	63.4	±1.7	64.2	±1.4	57.1	±2.5	58.0	±2.1	51.4	±2.2	53.3	±2.0	67.2	±2.0	68.6	±1.4
	SLICER, $h = 2$	63.6	±1.1	64.3	±1.0	56.8	±2.0	57.8	±2.5	50.4	±2.2	53.1	±1.9	68.3	±1.8	68.9	±1.5
	SLICER, $h = 8$	63.2	±2.4	64.5	±2.1	56.9	±0.8	58.9	±1.6	50.8	±1.7	52.6	±2.0	67.5	±1.8	68.2	±2.0

		UR				VI				YO				ZH			
		TRUE		ORACLE		TRUE		ORACLE		TRUE		ORACLE		TRUE		ORACLE	
LR	Model	ø	σ	ø	σ	ø	σ	ø	σ	ø	σ	ø	σ	ø	σ	ø	σ
$5e^{-6}$	Standard FT	61.7	±3.6	63.9	±2.1	68.9	±1.3	69.5	±1.0	47.8	±6.1	50.6	±3.2	22.7	±1.1	23.1	±1.2
	SLICER, $h = 1$	63.9	±2.6	65.4	±1.7	70.4	±1.5	70.7	±1.6	48.5	±1.7	48.8	±1.8	23.5	±2.7	24.2	±2.6
	SLICER, $h = 2$	64.2	±2.7	65.5	±2.8	70.1	±1.1	70.3	±1.2	47.9	±1.5	48.9	±2.1	23.7	±2.3	24.6	±2.4
	SLICER, $h = 8$	63.5	±3.5	65.4	±2.2	70.1	±1.3	70.5	±1.4	48.5	±1.6	48.7	±2.2	24.5	±2.1	25.4	±2.4
$1e^{-5}$	Standard FT	59.5	±2.3	63.8	±2.3	69.3	±1.6	70.2	±1.6	37.4	±9.1	48.7	±2.2	24.2	±2.1	25.0	±1.7
	SLICER, $h = 1$	63.7	±2.0	66.3	±2.1	71.1	±0.9	71.8	±1.0	49.4	±2.0	50.3	±2.2	25.7	±1.6	26.3	±1.4
	SLICER, $h = 2$	62.9	±1.4	66.2	±1.5	71.1	±1.3	72.4	±1.2	47.8	±3.2	49.3	±2.7	25.0	±1.1	25.7	±1.4
	SLICER, $h = 8$	64.6	±2.3	66.9	±2.0	70.9	±0.9	71.6	±1.2	49.2	±2.1	50.5	±2.6	25.5	±1.8	26.6	±2.0
$2e^{-5}$	Standard FT	54.1	±4.5	61.2	±3.4	69.3	±1.1	70.8	±1.6	31.4	±8.0	48.1	±2.0	25.1	±1.8	27.3	±1.4
	SLICER, $h = 1$	59.0	±4.1	62.9	±2.3	71.2	±1.0	72.2	±1.0	45.7	±4.6	50.1	±2.3	26.0	±3.1	27.0	±3.2
	SLICER, $h = 2$	58.8	±2.9	63.4	±2.3	72.1	±1.7	72.9	±1.7	42.5	±8.2	48.0	±6.5	26.0	±2.2	27.7	±2.1
	SLICER, $h = 8$	60.4	±3.0	64.4	±2.1	71.5	±1.1	72.9	±1.9	41.4	±10.0	48.1	±3.8	27.0	±2.0	28.6	±1.6

A.4 Russian as Source Language

		WIKIANN				MASAKHANER				TOTAL			
		TRUE		ORACLE		TRUE		ORACLE		TRUE		ORACLE	
LR	Model	ø	σ	ø	σ	ø	σ	ø	σ	ø	σ	ø	σ
$5e^{-6}$	Standard FT	53.8	±1.7	56.0	±2.0	33.9	±3.1	37.9	±3.7	47.9	±2.1	50.7	±2.5
	SLICER, $h = 1.0$	54.9	±1.9	56.3	±2.0	38.7	±2.9	41.1	±3.2	50.2	±2.2	51.8	±2.3
$1e^{-5}$	Standard FT	52.5	±1.7	56.5	±2.2	32.8	±2.6	38.3	±3.9	46.7	±2.0	51.1	±2.7
	SLICER, $h = 1.0$	54.0	±1.9	56.4	±2.1	37.7	±3.2	40.9	±3.8	49.2	±2.3	51.8	±2.6
$2e^{-5}$	Standard FT	51.3	±2.2	56.8	±2.6	31.1	±3.1	36.9	±4.2	45.3	±2.5	51.0	±3.1
	SLICER, $h = 1.0$	51.6	±2.5	55.7	±2.7	33.7	±4.2	40.1	±4.2	46.3	±3.0	51.1	±3.1

Table 5: Zero-shot cross-lingual transfer performance (micro-averaged F1) for NER with Russian as the source language. We report averages and st. deviations across all languages of MasakhaNER and WikiANN.