MANER: Mask Augmented Named Entity Recognition for Extreme Low-Resource Languages

Shashank Sonkar
Rice University
ss164@rice.edu

Zichao Wang
Rice University
jzwang@rice.edu

Richard G. Baraniuk
Rice University
richb@rice.edu

Abstract

This paper investigates the problem of Named Entity Recognition (NER) for extreme low-resource languages with only a few hundred tagged data samples. A critical enabler of most of the progress in NER is the readily available, large-scale training data for languages such as English and French. However, NER for low-resource languages remains relatively under-explored, leaving much room for improvement. We propose *Mask Augmented Named Entity Recognition* (MANER), a simple yet effective method that leverages the distributional hypothesis of pre-trained masked language models (MLMs) to improve NER performance for low-resource languages significantly. MANER repurposes the `[mask]` token in MLMs, which encodes valuable semantic contextual information, for NER prediction. Specifically, we prepend a `[mask]` token to every word in a sentence and predict the named entity for each word from its preceding `[mask]` token. We demonstrate that MANER is well-suited for NER in low-resource languages; our experiments show that for 100 languages with as few as 100 training examples, it improves on the state-of-the-art by up to 48% and by 12% on average on F1 score. We also perform detailed analyses and ablation studies to understand the scenarios that are best suited to MANER.

1 Introduction

Named Entity Recognition (NER) is a fundamental problem in natural language processing (NLP) (Nadeau and Sekine, 2007). Given an unstructured text, NER aims to label the named entity of each word, be it a person, a location, an organization, and so on. NER is widely employed as an important first step in many downstream NLP applications, such as scientific information retrieval (Krallinger and Valencia, 2005; Krallinger et al., 2017), question answering (Mollá et al., 2006), document classification (Guo et al., 2009), and recommender systems (Jannach et al., 2022).

Recent advances in NER have mainly been driven by deep learning-based approaches, whose training relies heavily on large-scale datasets (Rosenfeld, 2021). As a result, the most significant progress in NER is for resource-rich languages such as English (Wang et al., 2021), French (Tedeschi et al., 2021), German (Schweter and Akbik, 2020), and Chinese (Zhu and Li, 2022). This reliance on large training datasets makes it challenging to apply deep learning-based NER approaches to low-resource languages where training data is scarce. To illustrate the ubiquity of low-resource languages, WikiANN (Rahimi et al., 2019), one of the largest NER datasets, has NER-labeled data for 176 languages, but 100 of these languages have only 100 training examples.

Providing NER for low-resource languages is critical to ensure the equitable, fair, and democratized utilization of NLP technologies that are required to achieve the goal of making such technologies universally available for all (Magueresse et al., 2020; King, 2015). Several research efforts are pushing the frontiers of NER for low-resource languages in two orthogonal and complementary directions. The first direction aims to obtain larger NER datasets to solve the data scarcity problem, via either data collection or augmentation (Malmasi et al., 2022; Al-Rfou et al., 2014; Meng et al., 2021a). The second direction aims to develop new model architectures and training algorithms capable of handling scarce data. For example, ideas from meta-learning (de Lichy et al., 2021), distant supervision (Meng et al., 2021b), and transfer learning (Lee et al., 2017) leverage the few-shot generalizability of language models for NER in data-scarce settings.

Our Contributions. In this work, we propose *Mask Augmented Named Entity Recognition* (MANER), a new NER approach for low-resource languages that does not rely on additional data and does not require modifications to existing, off-

Proceedings of The Fourth Workshop on Simple and Efficient Natural Language Processing (SustaiNLP), pages 219–226

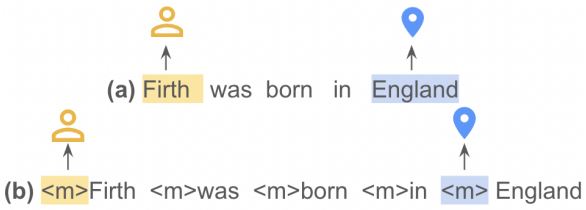

Figure 1: How MANER **(b)** differs from a standard NER model **(a)**. MANER 1) modifies the input to add a `[mask]` token before each word and 2) predicts the NER tag for a word from its preceding `[mask]` token.

the-shelf pre-trained models. The key intuition of MANER is to exploit the semantic information encoded in a pre-trained masked language model (MLM), in particular, in the `[mask]` token. Specifically, we reformat the input to the MLM by prepending a `[mask]` token to every token in the text to be annotated with NER tags. This reformatted input is then used to fine-tune the MLM with a randomly initialized NER prediction head on top of the prepended mask tokens. Extensive experiments on 100 extremely low-resource languages (each with only 100 training examples) demonstrate that MANER improves over state-of-the-art approaches by up to 48% and by 12% on average on F1 score. Detailed ablation and analyses of MANER demonstrate the importance of using the encoded semantic information and suggest scenarios in which MANER is most applicable.

2 Methodology

We now introduce MANER in detail and describe how it functions differently from a standard NER model (henceforth referred to as SNER).

SNER takes a sentence as input, passes the sentence through a transformer encoder model to obtain contextualized word embeddings, and applies a NER classifier layer on top of each word embedding to get the word's NER class.

MANER, in contrast, repurposes the `[mask]` token for the NER task. Two key differences that MANER implements as compared to SNER are 1) instead of giving the model the input sentence as is, MANER modifies the input sentence to append a `[mask]` token in front of each word and passes this modified sentence through the transformer encoder; and 2) instead of predicting the NER tag directly from the word embedding itself, MANER predicts the NER tag from the `[mask]` token embedding prepended to each word in the modified input sentence. These differences are illustrated in figure 1. We hypothesize that in such a setting, MANER will be better able to use the `[mask]` token to weigh the relative relevance of the neighboring word vs. the rest of the context when determining the label to assign to the neighboring word. Below, we expand on the above differences and introduce the two key components in MANER.

Modified input sentence. Let the set of NER labels be denoted by $\mathcal{N}$. Let the sequence of NER labels for a sentence $S = \{w_0, w_1, .., w_{n-1}\}$ consisting of n words be $L = \{c_0, c_1, .., c_{n-1}\}$ where $c_i \in \mathcal{N}$, $0 \leq i < n$. To obtain the input that MANER requires we *append* a `[mask]` token to the beginning of each word in sentence S. The new sentence is $S' = \{m, w_0, m, w_1, .., m, w_{n-1}\}$ where m is the `[mask]` token. The modified labels L' are $\{c_0, \emptyset, c_1, \emptyset, .., c_{n-1}, \emptyset\}$. The original NER label of each word is assigned to the `[mask]` token to the immediate left of the word.

MANER's classifier design. MANER uses a pre-trained, masked language model as the backbone with an NER classifier head on top. The transformer model takes a sentence as input and outputs embeddings for each token in the sentence. The NER classifier uses the token embeddings to output the most probable NER class for each token.

Denote the MANER model by $\mathcal{M}$. The transformer model is given by T, $T(S') = T(\{m, w_0, m, w_1, .., m, w_{n-1}\}) = \{\mathbf{e}_0, \mathbf{e}_1, .., \mathbf{e}_{2n-1}\}$, where $\mathbf{e}_i \in \mathbb{R}^D$, $0 \leq i < 2n-1$ is the token embedding, and D is its dimension. The NER classifier is modeled using a weight matrix $\mathbf{M} \in \mathbb{R}^{D \times |\mathcal{N}|}$ that takes the computed token embeddings as input. Using these token embeddings, the classifier outputs scores for all NER labels for each token in the sentence. Passing these scores through a softmax nonlinearity provides probabilities $\mathbf{p}_i \in \mathbb{R}^{|\mathcal{N}|}$ for all NER classes in $\mathcal{N}$ for a given token i in S:

$$\mathbf{p}_i = \text{softmax}\Big(\mathbf{M}\big(\mathbf{e}_i\big)\Big).$$

Summing up, we have

$$\mathcal{M}(T, \mathbf{M}, S', i) = \mathbf{p}_i,\ 0 \leq i < 2n.$$

MANER training and inference. During training, the weights of $\mathbf{M}$ and T are learned/fine-tuned by minimizing cross-entropy loss. Note that the loss is not calculated for labels marked $\emptyset$ in the modified label set L'. The NER label of the word is given by the NER label of the `[mask]` token preceding it. During inference, each word in the sentence is prepended with the `[mask]` token, and the NER class of each word is the most probable NER class of its prepended `[mask]` token.

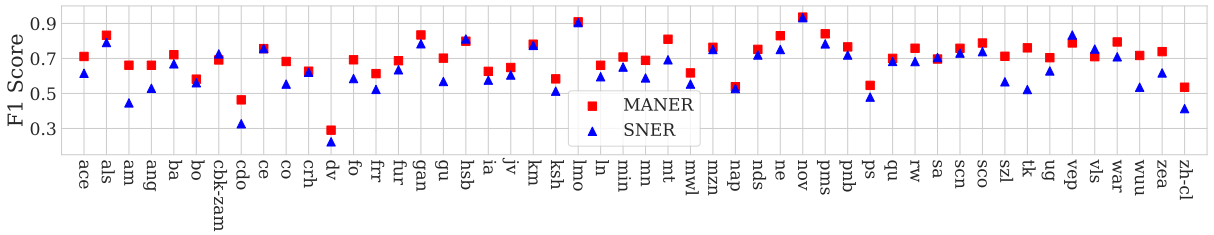

Figure 2: F1 scores comparing MANER against SNER for a subset of 50 low-resource languages in the WikiANN dataset that only have 100 training samples. The x-axis represents the languages. Overall, MANER improves NER performance for 88 of the 100 languages over SNER; performance improvement is up to 48% and on average 12%.

3 Experiments

We perform various empirical studies on MANER to 1) demonstrate its superior performance in low-resource language NER tasks and 2) provide insights into its performance and scenarios in which it will work well.

Dataset. We use the WikiANN multilingual NER dataset (Pan et al., 2017; Rahimi et al., 2019), which provides three named entities (person, location, organization) for Wikipedia articles across 176 languages. Therefore, our NER tag set $\mathcal{N}$ has four elements, with an additional null tag. To the best of our knowledge, WikiANN is by far the most comprehensive dataset for multilingual NER. Other multilingual datasets exist but they cover a few popular languages. For example, CoNLL (Kim Sang and De Meulder, 2003) contains only English, German, Dutch, and Spanish. We focus our main study on *extreme low-resource* settings by experimenting on the 100 languages in WikiANN that each has only **100 examples** for train and test splits. Further details on the WikiANN dataset are in Appendix B.

MANER implementation and baselines. We use XLM-RoBERTa-large (Conneau et al., 2019) as the backbone model for MANER and all baselines. XLM-RoBERTa-large is a multilingual version of the RoBERTa (Liu et al., 2019) model, pre-trained with the MLM objective on 2.5TB of filtered CommonCrawl data containing 100 languages (Conneau et al., 2020). We compare MANER against two baselines 1) SNER , which stands for a standard NER model and 2) MLM-NER , which is another strategy to use the `[mask]` token for NER inspired by the masked language modeling (MLM) loss. MLM-NER masks a small percentage of words and predicts the NER tag for both the masked and unmasked words, thus leveraging the `[mask]` token. More details on the above models and their training setup are in Appendix A.

3.1 Main results

Metric	SNER	MLM-NER	MANER
F1	0.649	0.643 (-0.5%)	0.715 (12%)

Table 1: Average F1 scores for the 100 languages in the WikiANN dataset with only 100 samples comparing MANER to baselines. In this extreme low-resource setting, MANER achieves an average improvement of **12%** over baselines.

In Table 1, we report the average of F1 score for the 100 languages in WikiANN that we consider. MANER provides a significant **12%** average improvement in our low-resource language settings. The MLM inspired NER-model MLM-NER , in contrast, performs only similarly to SNER. We also plot, in Figure 2, the F1 score of 50 randomly sampled low-resource languages comparing SNER against MANER (the plot for the remaining languages is in Appendix D). MANER offers up to 48% performance improvements compared to SNER, and there are only a few languages (12 out of 100) in which the SNER outperforms MANER.

We believe the reason that MANER outperforms MLM-NER is that MANER uses the `[mask]` token for NER prediction in both training and inference, whereas MLM-NER does not. Therefore, MANER learns to give more importance to the context in the case of out-of-distribution test labels using the `[mask]` token during inference. We will revisit and empirically support the above reasoning in Section 3.2. Additionally, in MLM-NER training, we mask out certain words with the `[mask]` token, which introduces noise and makes training and the NER task more difficult.

3.2 Analysis: Importance of the `[mask]` token

We now conduct two analyses to demonstrate the importance of using the `[mask]` token in MANER. Intuitively, the `[mask]` token can be helpful because it learns to encode the semantics of the context during pre-training and, thereby, the

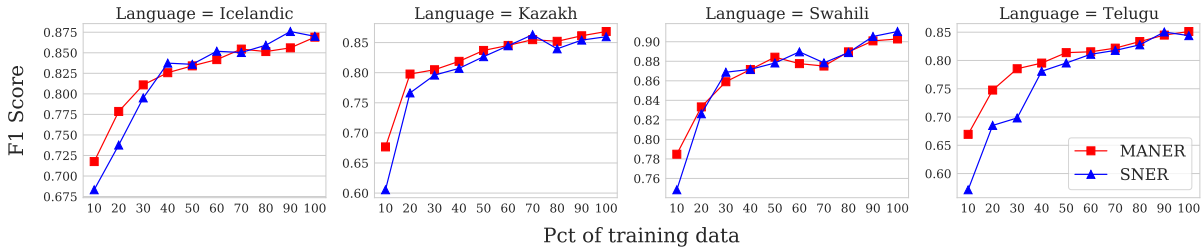

Figure 3: Measure effect of training samples to performance in MANER. MANER can give a boost in performance till 400 samples and then both MANER and SNER model perform similarly. This demonstrates that MANER is best suited for extreme low resource languages and rapid prototyping since it is easy and cost-effective to obtain very few human annotations to achieve large performance improvements (just 100 annotations are required).

Metric	SNER	MANER (w/ `[mask]`)	MANER (w/ `[rand]`)
F1	0.649	0.715 (12%)	0.679 (6%)

Table 2: Average F1 scores for 100 languages for MANER using the `[mask]` token and the `[rand]` token. Replacing the `[mask]` token with the `[rand]` token diminishes the improvements.

word that needs to be tagged (by distributional hypothesis in Harris (1954) which states the meaning of a word can be inferred from its context).

In the first analysis, we replace the `[mask]` token in MANER with a control token, namely, the random token `[rand]`. Note that the `[rand]` token is not learned during the XLM-RoBERTa model pre-training; thus, it will not encode any contextual information. As we see in Table 2, if we replace the `[mask]` token with `[rand]` , MANER achieves only a 6% improvement in F1 performance over the SNER baseline. This result illustrates the power of the context: even when the `[rand]` token does not contain contextual information during pre-training, MANER can still use the `[rand]` token to predict how much weight to assign the context and the word immediately adjacent to it depending on the test sample.

In the second analysis, we report in Table 3 the averaged F1 score of only those languages on which the XLM-RoBERTa model was pre-trained with at least 0.5GB of training data per language. The rationale behind this experiment design is that the `[mask]` token will encode the context semantics of a language only if the language was seen during the pre-training stage of XLM-RoBERTa model. As we see in Table 3, in this case, MANER provides a whooping 18% improvement in F1 score (as compared to the 12% gain in Table 1) if the language was seen in the pre-training stage. This experiment again highlights the importance of using `[mask]` token in MANER.

Metric	SNER	MANER
F1	0.603	0.705 (18%)

Table 3: F1 scores comparing MANER against SNER, averaged on a subset of languages on which XLM-RoBERTa was pre-trained. The improvement over SNER is 18% compared to 12% improvement in the previous study that included all 100 languages.

3.3 Analysis: Effect of training set size

We measure the effectiveness of MANER in situations where more training data is available. For this purpose, we select 4 languages from WikiANN dataset that have 1000 training data samples each. From Figure 3, we see that MANER boosts F1 performance over the SNER baseline until about 400 samples and then both methods perform similarly. This result demonstrates that MANER is best suited for *extreme low resource languages and rapid prototyping* because it is *easy and cost-effective to obtain very few human annotations to achieve large performance improvements* (e.g., just 100 annotations are required).

4 Conclusions

In this paper, we have proposed Mask Augmented Named Entity Recognition (MANER) for NER in extreme low-resource language settings. MANER exploits the information encoded in pre-trained masked language models (inside `[mask]` token specifically) and outperforms existing approaches for extreme low-resource languages with as few as only 100 training examples by up to 48% and by 12% on average on F1 score. Analyses and ablation studies show that using semantics encoded in `[mask]` token is integral to MANER. Future work will exploit MANER's effectiveness for highly resource-constraint and human-in-the-loop settings, such as rapid prototyping in an active learning setup and few-shot learning with human annotators.

5 Limitations

Our proposed method MANER for improving NER is best suited for low-resource settings. As discussed in Section 3.3, we measured the effectiveness of MANER in situations where more training data is available and found that MANER boosts F1 performance over the SNER baseline until about 400 training examples, and then both methods perform similarly. The result demonstrated that MANER is best suited for extreme low-resource languages and rapid prototyping because it is easy and cost-effective to obtain very few human annotations to achieve significant performance improvements.

We base the experiments in this paper on a widely adopted model, XLM-RoBERTa, pre-trained on multiple languages. It is possible that the empirical conclusions we draw from the observations do not generalize to other pre-trained models.

6 Ethics Statement

We believe providing NER for low-resource languages is critical to ensure the equitable, fair, and democratized utilization of NLP technologies that are required to achieve the goal of making such technologies universally available for all. Our work contributes to this direction by proposing MANER, which boosts performance for 100 languages with only 100 training samples each.

Acknowledgement

This work was supported by NSF grants 1842378, ONR grant N0014-20-1-2534, AFOSR grant FA9550-22-1-0060, and a Vannevar Bush Faculty Fellowship, ONR grant N00014-18-1-2047.

References

Rami Al-Rfou, Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. 2014. Polyglot-ner: Massive multilingual named entity recognition.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

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.

Cyprien de Lichy, Hadrien Glaude, and William Campbell. 2021. Meta-learning for few-shot named entity recognition. In *Proceedings of the 1st Workshop on Meta Learning and Its Applications to Natural Language Processing*. Association for Computational Linguistics.

Jiafeng Guo, Gu Xu, Xueqi Cheng, and Hang Li. 2009. Named entity recognition in query. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval - SIGIR '09*. ACM Press.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2022. A survey on conversational recommender systems. *ACM Computing Surveys*, 54(5):1–36.

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

Benjamin Philip King. 2015. *Practical Natural Language Processing for Low-Resource Languages.* Ph.D. thesis, University of Michigan.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Martin Krallinger, Obdulia Rabal, Anália Lourenço, Julen Oyarzabal, and Alfonso Valencia. 2017. Information retrieval and text mining technologies for chemistry. *Chemical Reviews*, 117(12):7673–7761.

Martin Krallinger and Alfonso Valencia. 2005. Text-mining and information-retrieval services for molecular biology. *Genome biology*, 6(7):1–8.

Ji Young Lee, Franck Dernoncourt, and Peter Szolovits. 2017. Transfer learning for named-entity recognition with neural networks.

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. *CoRR*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022. MultiCoNER: A large-scale multilingual dataset for complex named entity recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3798–3809, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Tao Meng, Anjie Fang, Oleg Rokhlenko, and Shervin Malmasi. 2021a. Gemnet: Effective gated gazetteer representations for recognizing complex entities in low-context input. In *NAACL 2021*.

Yu Meng, Yunyi Zhang, Jiaxin Huang, Xuan Wang, Yu Zhang, Heng Ji, and Jiawei Han. 2021b. Distantly-supervised named entity recognition with noise-robust learning and language model augmented self-training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Diego Mollá, Menno van Zaanen, and Daniel Smith. 2006. Named entity recognition for question answering. In *Proceedings of the Australasian Language Technology Workshop 2006*, pages 51–58, Sydney, Australia.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticæ Investigationes. International Journal of Linguistics and Language Resources*, 30(1):3–26.

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.

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

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Jonathan S. Rosenfeld. 2021. Scaling laws for deep learning. *CoRR*, abs/2108.07686.

Stefan Schweter and Alan Akbik. 2020. Flert: Document-level features for named entity recognition.

Simone Tedeschi, Valentino Maiorca, Niccolò Campolungo, Francesco Cecconi, and Roberto Navigli. 2021. WikiNEuRal: Combined neural and knowledge-based silver data creation for multilingual NER. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics.

Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu. 2021. Automated Concatenation of Embeddings for Structured Prediction. In *the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (**ACL-IJCNLP 2021**)*. Association for Computational Linguistics.

Enwei Zhu and Jinpeng Li. 2022. Boundary smoothing for named entity recognition. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

A Implementation details on baselines

A.1 SNER

Similar to MANER design, current standard NER systems (SNER) built upon transformer models also simply add a NER classifier to the top of a transformer model. The classifier predicts the NER class of each token of an unmodified sentence S:

$$\mathcal{M}_{\text{base}}(T, \mathbf{M}, S, i) = \text{softmax}\Big(\mathbf{M}\big(\mathbf{e}_{\text{i}}\big)\Big) = \mathbf{p}_{\text{i}}, \quad 0 \leq i < n,$$

where $\mathcal{M}_{\text{base}}$ is an NER model built on a transformer model T using classifier weight matrix $\mathbf{M}$. This baseline method remains the de-facto method for training NER models for most languages (especially low-resource languages) to the best of our knowledge, though specialized models have been built for popular languages like English.

Inference: Similar to training during inference, the NER class of each word in the sentence is the most probable NER tag assigned to the classified word embedding.

A.2 MLM-NER

Our MANER methodology in Section 2 is one way to change the input phrase using the mask token. In this baseline, we introduce yet another way to repurpose the `[mask]` token for NER that is inspired by the masked language modeling (MLM) framework that is used for pre-training transformer models which we refer as MLM-NER. In MLM, a word is predicted using the words surrounding it in the sentence. Since the NER category of a word is also a semantic property of the word, we use the philosophy of MLM for NER fine-tuning.

In MLM pre-training, the dataset is prepared by *masking* random words in a sentence with a `[mask]` token with a fixed probability p_{mlm}. Then, the masked words are predicted using the context information.

Analogous to MLM pre-training, for NER fine-tuning, we randomly replace words in sentence S with the `[mask]` token with the fixed probability p_{ner}. However, instead of predicting the missing words, as with MLM, we predict the NER labels L for each word w in S irrespective of whether the word was replaced by a `[mask]` token or not. In the case the word was replaced with the `[mask]` token, the transformer outputs the `[mask]` token embedding for that word.

Thus the modified input to the transformer is $S' = \text{mask(S)}$, where

$$\text{mask}(\text{w}_\text{i}) = \begin{cases} \texttt{[mask]}\ , & \text{if } p_i \leq p_{ner}, \\ w_i, & \text{otherwise} \end{cases} \quad (1)$$

with p_i a random number between 0 and 1 generated for w_i. Then, we use the first baseline NER model design $\mathcal{M}_{base}$ for training, but now it is is fine-tuned on S' and L (note we predict the label of the `[mask]` tokens as well). Inference with this model remains same as the first baseline model.

A.3 MANER Classifier Input Embedding

The NER prediction for each word in MANER is based on the embedding of the *first token of the word*. This is a common practice in NER with Transformer-based models where a word may be tokenized into multiple tokens.

A.4 Training procedures

For each language in our experiment, we train MANER and baselines for 30 epochs with a learning rate of 5e^{-6} and the loss optimized using Adam (Loshchilov and Hutter, 2019). Training takes 3 minutes on a single 11 GB GeForce GTX 1080 Ti GPU for a single language. We run MANER for following five random seeds for each language - 12345, 23451, 34512, 45123, 51234. The standard deviation in performance for SNER averaged over 100 languages for 5 runs is 0.649 ± 0.005 and for MANER is 0.715 ± 0.007.

B WikiANN dataset details

The NER labels in WikiANN are in IOB2 (Inside–outside–beginning) format (Ramshaw and Marcus, 1995) comprising PER (person), LOC (location), and ORG (organization) tags. An instance of NER tagged sentence: *UNICEF*(B-ORG) is a nonprofit organization, founded by *Ludwik*(B-PER) *Rajchman*(I-PER) headquartered at *New*(B-LOC) *York*(I-LOC), *United*(B-LOC) *States*(I-LOC).

In addition, language names corresponding to abbreviations used in figure 2 can be found in the Appendix section of Conneau et al. (2020).

C Comments on catastrophic forgetting in MANER

The *catastrophic forgetting* (Kirkpatrick et al., 2017) phenomenon that masked language models undergo during any kind of fine-tuning is one of the

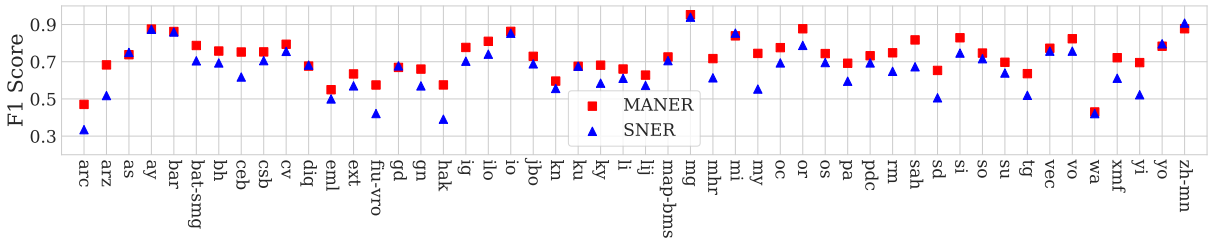

Figure 4: F1 scores comparing MANER against SNER for the remaining 50 low-resource languages in the WikiANN dataset that only have 100 training samples each. Similar to the results in Figure 2, MANER gives a significant improvement of **12%** on F1 score compared to SNER .

.

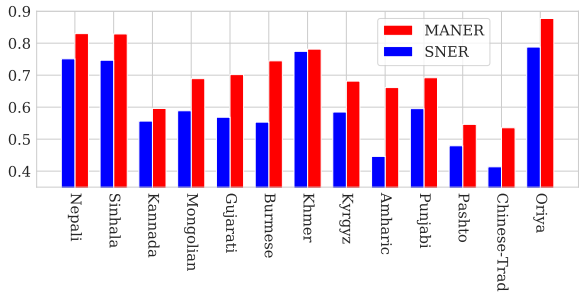

Figure 5: F1 score comparing MANER against SNER on a subset of languages on which the backbone of both models, XLM-RoBERTa-large, has been pre-trained. MANER improves upon SNER for each of these languages, with F1 score improvement of up to 22% and 18% on average.

reasons we think MANER does not provide gains when more training data is available (of course more training data also implies less reliance on specialized techniques like ours). Catastrophic forgetting causes the loss of useful context semantics encoded in the `[mask]` token during the fine-tuning stage that MANER heavily relies on. Adding an additional masked language modeling loss to the NER loss during fine-tuning may help to circumvent catastrophic forgetting; we leave this investigation as a valuable venue for future work.

D Additional experiment results

Figure 4 shows the performance comparing MANER and SNER on the remaining 50 low-resource languages in the WikiANN dataset. The results here align with that in the main text: MANER provides performance improvement, sometimes significantly, over SNER.

Figure 5 shows the performance comparing MANER and SNER on a subset of languages on which the backbone of both models, XLM-RoBERTa-large, has been pre-trained. The results corroborate with those in the main text: MANER improves upon SNER for each of these languages, with F1 score improvement of up to 22% and 18% on average.