mPMR: A Multilingual Pre-trained Machine Reader at Scale*

Weiwen Xu[1,2,†] **Xin Li**[2,‡] **Wai Lam**[1] **Lidong Bing**[2]
[1]The Chinese University of Hong Kong
[2]DAMO Academy, Alibaba Group
{wwxu,wlam}@se.cuhk.edu.hk {xinting.lx,l.bing}@alibaba-inc.com

Abstract

We present multilingual Pre-trained Machine Reader (mPMR), a novel method for multilingual machine reading comprehension (MRC)-style pre-training. mPMR aims to guide multilingual pre-trained language models (mPLMs) to perform natural language understanding (NLU) including both sequence classification and span extraction in multiple languages. To achieve cross-lingual generalization when only source-language fine-tuning data is available, existing mPLMs solely transfer NLU capability from a source language to target languages. In contrast, mPMR allows the direct inheritance of multilingual NLU capability from the MRC-style pre-training to downstream tasks. Therefore, mPMR acquires better NLU capability for target languages. mPMR also provides a unified solver for tackling cross-lingual span extraction and sequence classification, thereby enabling the extraction of rationales to explain the sentence-pair classification process.[1]

1 Introduction

Multilingual pre-trained language models, acronymed as mPLMs, have demonstrated strong Natural language understanding (NLU) capability in a wide range of languages (Xue et al., 2021; Cai et al., 2021, 2022; Conneau et al., 2020a; Ding et al., 2022; Li et al., 2020a). In particular, mPLMs can maintain exceptional cross-lingual language understanding (XLU) capability on unseen *target* languages though mPLMs are only fine-tuned on resource-rich *source* languages like English.

It has been proved that optimizing cross-lingual representations of mPLMs can improve XLU capability. For example, cross-lingual supervisions, such as parallel sentences (Conneau and Lample, 2019) or bilingual dictionaries (Conneau et al., 2020b) could enhance cross-lingual representations with better language alignment. XLM-R (Conneau et al., 2020a) and mT5 (Xue et al., 2021) showed that appropriately incorporating more languages during pre-training leads to better cross-lingual representations. A few works enriched the cross-lingual representations with factual knowledge through the utilization of multilingual mentions of entities (Calixto et al., 2021; Ri et al., 2022) and relations (Liu et al., 2022; Jiang et al., 2022) annotated in knowledge graphs. Despite their differences, the above methods essentially constructed more diverse multilingual corpora for pre-training mPLMs. These mPLMs would presumably meet their saturation points and are known to suffer from *curse of multilinguality* (Conneau et al., 2020a; Pfeiffer et al., 2022; Berend, 2022). Under this situation, introducing more training data from either existing (Pfeiffer et al., 2022) or unseen (Conneau et al., 2020a) languages for enhancing mPLMs may not bring further improvement or even be detrimental to their cross-lingual representations.

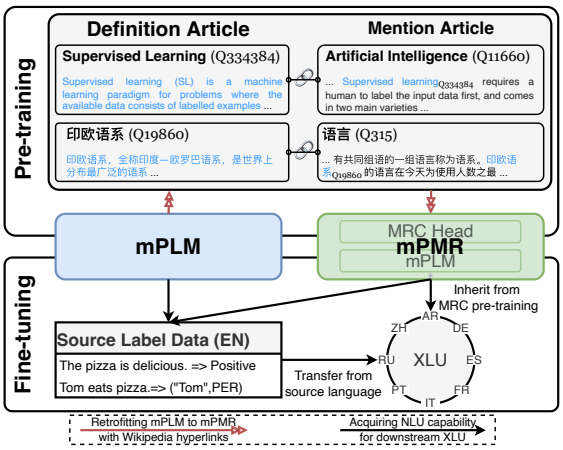

Figure 1: Pre-training and fine-tuning of mPMR.

* This work was supported by Alibaba Group through Alibaba Research Intern Program. The work described in this paper was also partially supported by a grant from the Research Grant Council of the Hong Kong Special Administrative Region, China (Project Code: 14200719). † This work was done when Weiwen Xu was an intern at Alibaba DAMO Academy. ‡ Xin Li is the corresponding author.

[1]The code, data, and checkpoints are released at https://github.com/DAMO-NLP-SG/PMR

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 1533–1546
July 9-14, 2023

In the paper, instead of training a new mPLM with better cross-lingual representations, we propose **m**ultilingual **P**re-trained **M**achine **R**eader (mPMR) to directly guide existing mPLMs to perform NLU in various languages. As shown in Figure 1, mPMR resembles PMR (Xu et al., 2022) for constructing multilingual machine reading comprehension (MRC)-style data with Wikipedia hyperlinks. These data are used to retrofit an mPLM into an mPMR through an MRC-style continual pre-training. During retrofitting process (i.e., pre-training), mPMR jointly learns the general sequence classification and span extraction capability for multiple languages. In XLU fine-tuning, mPLMs solely rely on cross-lingual representations to transfer NLU capability from a source language to target languages. By contrast, mPMR enables the direct inheritance of multilingual NLU capability from the MRC-style pre-training to downstream tasks in a unified MRC formulation, which alleviates the discrepancies between source-language fine-tuning and target-language inference (Zhou et al., 2022a,b, 2023). Therefore, mPMR shows greater potential in XLU than mPLMs.

To improve the scalability of mPMR across multiple languages, we further propose *Unified Q/C Construction* and *Stochastic answer position* strategies for refining the curation of MRC data. With these two strategies, mPMR can better generalize to low-resource languages and becomes more robust to position bias (Ko et al., 2020).

The experimental results show that mPMR obtains clear improvements over XLM-R (Conneau et al., 2020a) on span extraction, with an average improvement of up to 12.6 F1 on TyDiQA, and 8.7 F1 on WikiAnn respectively. The analysis reveals that mPMR benefits from more multilingual MRC data for pre-training. We also found that mPMR converges faster in downstream tasks and is capable of using its strong extraction capability for explaining the sequence classification process.

2 mPMR

We present the MRC model and training data of mPMR. We closely follow PMR (Xu et al., 2022) and introduce the modifications for enabling multilingual MRC-style pre-training.

2.1 Model Pre-training

Our mPMR follows the same MRC architecture of Xu et al. (2022, 2023) with an encoder and an extractor. The encoder maps input tokens X, the concatenation of the query Q, the context C, and special markers (i.e., [CLS] and [SEP]), into hidden representations H. For any two tokens X_i and X_j ($i < j$), the extractor receives their contextualized representations H_i and H_j and predicts the probability score $S_{i,j}$ indicating the probability of the token span $X_{i:j}$ being the answer to the query Q.

mPMR is guided with the Wiki Anchor Extraction (WAE) objective to train both the encoder and the extractor. WAE checks if the answer to the query exists in the context. If so, WAE would first regard the query and the context to be relevant and extracts the [CLS] token as a sequence-level relevance indicator. WAE would then extract all corresponding answers from the context.

2.2 Multilingual MRC Data

Training mPMR requires the existence of labeled (query, context, answer) triplets. To obtain such data, we collected Wikipedia articles with anchor annotations for 24 languages, which are the most widely used and cover a reasonable number of languages used in XLU tasks (Ri et al., 2022).

As shown in Figure 1, we utilized a Wikipedia anchor to obtain a pair of correlated articles. One side of the pair is the article that provides in-depth descriptions of the anchor entity, which we defined as the *definition article*. The other side of the pair is named as the *mention article*, which mentions the specific anchor text[2]. We composed an answerable MRC example in which the anchor is the answer, the surrounding text of the anchor in the mention article is the context, and the definition of the anchor entity in the definition article is the query. Additionally, we can generate an unanswerable MRC example by pairing a query with an irrelevant context without anchor association.

Unified Q/C Construction. PMR constructed the MRC query and context as valid sentences so as to keep the text coherent. However, sentence segmentation tools are usually not available for low-resource languages. To remedy this, we did not apply sentence segmentation but only preprocess Wikipedia articles with word tokenization in mPMR. For each anchor, the MRC query comprises the first Q words in the definition article. To prevent information leakage during pre-training, similar to PMR, we anonymized the anchor entity

[2]definition/mention article refers to home/reference article of Xu et al. (2022).

Model	#Params	EQA			NER		ABSA	Sentence Pair		Avg.
		XQuAD	MLQA	TyDiQA	WikiAnn	CoNLL	SemEval16	PAWS-X	XNLI	
Metrics		F1 / EM	F1 / EM	F1 / EM	F1	F1	F1	Acc.	Acc.	
XLM-R	550M	76.6 / 60.8	71.6 / 53.2	65.1 / 45.0	65.4	82.0	66.9‡	86.4	79.2	74.2
mT5	580M	67.0 / 49.0	64.6 / 45.0	57.2 / 41.2	55.7	71.0‡	62.5‡	86.4	75.4	67.5
VECO	550M	77.3 / 61.8	71.7 / 53.2	67.6 / 49.1	65.7	81.3‡	63.0‡	**88.7**	**79.9**	74.4
mLUKE-W	561M	**79.6** / -	72.7 / -	65.2 / 48.5‡	67.7‡	83.0	61.2‡	88.2‡	79.4‡	74.6
Wiki-CL	550M	72.1 / 56.9	70.8 / 50.5	73.2 / 57.3	64.7	-	-	88.4	79.2	-
KMLM	550M	77.3 / 61.7	72.1 / 53.7	67.9 / 50.4	66.7‡	83.2	66.1‡	88.0	79.2	75.1
Our MRC Formulation										
XLM-R$_{\text{base}}$	270M	70.8 / 56.9	64.4 / 47.9	50.8 / 38.2	57.9	79.2	60.0	85.0	73.3	67.7
mPMR$_{\text{base}}$	270M	74.0 / 59.5	65.3 / 48.7	63.4 / 49.0	66.6	81.7	62.1	86.1	73.6	71.6
XLM-R	550M	77.1 / 61.3	71.5 / 53.9	67.4 / 51.6	63.6	81.4	66.1	86.9	78.6	74.1
mPMR	550M	79.2 / **64.4**	**73.1 / 55.4**	**74.7 / 58.3**	**70.7**	**84.1**	**68.2**	88.0	79.3	**77.2**

Table 1: The results of all XLU tasks. We report the average results of all languages for each dataset. We also compute the overall average score among all datasets in the **Avg.** column. We reproduce the missing results with the ‡ label. Some results of Wiki-CL are left blank because they do not release their model checkpoint.

in the query to the [MASK] token. The MRC context consists of C words surrounding the anchor.

Stochastic Answer Position. As mentioned by Ko et al. (2020), the model is prone to overfitting to the position shortcut if the answer in the context exhibits a fixed position pattern. In our case, suppose that the MRC context consists of $C/2$ words on both the left and right sides of the anchor, the model may learn the shortcut that the middle part of the context is likely to be the answer. To prevent such position bias, we propose a stochastic answer position method, which allows the answer to be presented in any position within the context. Specifically, given an anchor in a Wikipedia article, the context comprises ξ words preceding the anchor and the $C - \xi$ words following the anchor, where ξ is a random integer ranging from 0 to C and varies across different contexts. In accordance with PMR, we treated all text spans identical to the anchor in the current context as valid answers.

3 Experimental Setup

Implementation Details. In mPMR, the encoder is loaded from XLM-R (Conneau et al., 2020a) and the extractor is randomly initialized. Both components are then continually pre-trained using the multilingual MRC data that we constructed. More hyper-parameters can be found in Appendix A.1.

Downstream XLU Tasks. We evaluated mPMR on a series of span extraction tasks, including Extractive Question Answering (EQA), Named Entity Recognition (NER), and Aspect-Based Sentiment Analysis (ABSA). We also evaluated our mPMR on two sequence classification tasks. We followed Xu et al. (2022) to convert all tasks into MRC formulation to effectively leverage the knowledge that is acquired during MRC-style pre-training. For EQA, we used XQuAD (Artetxe et al., 2020), MLQA (Lewis et al., 2020), and TyDiQA (Clark et al., 2020). For NER, we used WikiAnn (Pan et al., 2017) and CoNLL (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003). SemEval16 (Pontiki et al., 2016) was used for ABSA task. Regarding the sequence classification, we used XNLI (Conneau et al., 2018) and PAWS-X (Yang et al., 2019). Additional dataset information and concrete examples are provided in Appendix A.2

Baselines. We compared mPMR with recent methods on improving cross-lingual representations, including 1) models pre-trained on a large number of languages: XLM-R (Conneau et al., 2020a), mT5 (Xue et al., 2021), and VECO (Luo et al., 2021); 2) models that exploited multilingual entity information: Wiki-CL (Calixto et al., 2021), and mLUKE-W (Ri et al., 2022); and 3) Model that utilized multilingual relation information: KMLM (Liu et al., 2022). For a fair comparison, all models have approximately the same parameter size.

4 Results and Analyses

XLU Performance. Table 1 shows the results on a variety of XLU tasks. mPMR outperforms all previous methods with an absolute improvement of 2.1 F1 over the best baseline (i.e. KMLM). mPMR shows greater improvements over previ-

Index	Model	#Lang	PAWS-X	XQuAD	WikiAnn	Avg.
#1	XLM-R_{base}	0	85.0	70.8	57.9	71.2
#2	#1 + MRC data in English	1	85.2 (0.2↑)	71.0 (0.2↑)	59.5 (1.6↑)	71.9 (0.7↑)
#3	#2 + Stochastic Answer Position	1	85.5 (0.3↑)	73.0 (2.0↑)	60.0 (0.5↑)	72.8 (0.9↑)
#4	#3 + MRC data in more languages	10	85.9 (0.4↑)	73.5 (0.5↑)	64.7 (4.7↑)	74.7 (1.9↑)
#5	#4 + MRC data in even more languages (mPMR$_{base}$)	24	**86.1** (0.2↑)	**74.0** (0.5↑)	**66.6** (1.9↑)	**75.6** (0.9↑)

Table 2: The process of retrofitting XLM-R into mPMR using multilingual MRC data (English→10 languages→24 languages) and our Stochastic Answer Position method. Each row accumulates modifications from all rows above.

Label	Sentence 1	Sentence 2
Entailment	Rami Nieminen (born February 25 , 1966) is a Finnish footballer.	Rami Nieminen (born 25 February 1966) is a Finnish former footballer.
Contradiction	In 1938 he became the Government Anthropologist of the Egyptian-Anglo Sudan and conducted fieldwork with the Nuba.	In 1938 he became the government anthropologist of the anglo-Egyptian Sudan and led fieldwork with the Nuba .
Entailment	Stipsits 出生于科尔新堡，并在维也纳施塔莫斯多夫度过了他的童年。	什蒂普西奇出生于德国科恩堡，在维也纳斯塔莫斯多夫度过了他的童年。
Contradiction	纳舒厄白银骑士团队加入了夏季大学联盟，是本市的现役球队。	Nashua Silver Knights 队是当前夏季联赛的一部分，也是该市的大学体育队。
Entailment	これらの見方は、福音主義的、清教徒的、プロテスタント的な動きが出現するとともに、しばしば表明されてきました。	これらの見解は多くの場合、新教徒、清教徒、福音主義者が出現するなかで示されてきた。
Contradiction	1954 年にスリナムに戻った後、弁護士としてパラマリボに定住した。	1954 年、パラマリボに戻ると、彼はスリナムで弁護士として定住しました。

Table 3: Case study on PAWS-X. mPMR can extract rationales to explain the sequence-pair classification in multiple languages.

ous methods on span extraction tasks. In particular, mPMR achieves up to 7.3 and 7.1 F1 improvements over XLM-R on TyDiQA and WikiAnn respectively. Such significant improvements probably come from the following two facts: (1) WikiAnn comprises a larger number of target languages (i.e. 40). Therefore, existing methods may struggle to align these low-resource languages with English due to a lack of language-specific data. (2) TyDiQA is a more challenging cross-lingual EQA task with 2x less lexical overlap between the query and the answer than MLQA and XQuAD (Hu et al., 2020). Our mPMR, which acquires target-language span extraction capability from both MRC-style pre-training and English-only QA fine-tuning, achieves larger performance gains on more challenging task.

mPMR Pre-training. To reflect the impact of our MRC-style data and Stochastic Answer Position method on pre-training, we present a step-by-step analysis of the retrofitting process starting from XLM-R in Table 2. Our findings suggest that the significant improvements observed are largely due to the inclusion of multilingual MRC data. Introducing English MRC data (model #2) gives marginal improvements because model #2

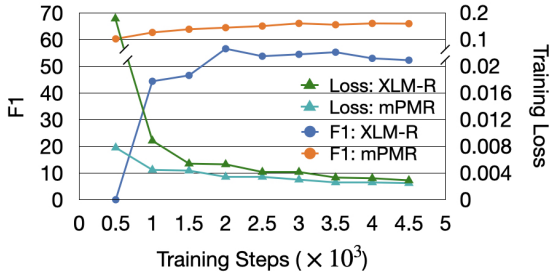

Figure 2: Convergence speed (Test set F1 and the training loss) of mPMR$_{base}$ and XLM-R$_{base}$ on WikiAnn.

can only rely on cross-lingual representations to transfer the knowledge acquired during MRC-style pre-training. When using MRC data on more languages (model #4 and #5), we can observe significant improvements on XLU tasks. This can be attributed to the NLU capability directly inherited from MRC-style pre-training in target languages. Additionally, with our Stochastic Answer Position method (model #3), mPMR becomes more robust to position bias and thus improves XLU tasks.

Explainable Sentence-pair Classification. Inspired by PMR (Xu et al., 2022), we investigated if the extraction capability of mPMR can be leveraged to explain sentence-pair classification. Note

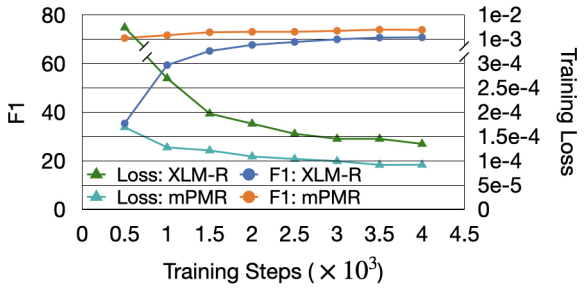

Figure 3: Convergence speed (Test set F1 and the training loss) of $\text{mPMR}_{\text{base}}$ and $\text{XLM-R}_{\text{base}}$ on XQuAD.

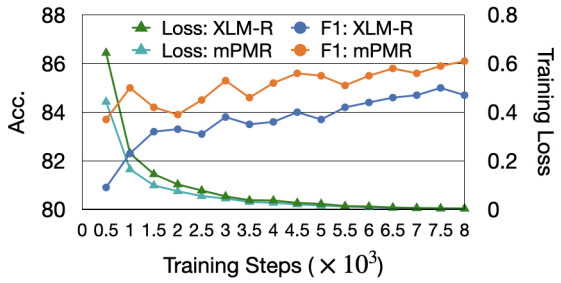

Figure 4: Convergence speed (Test set F1 and the training loss) of $\text{mPMR}_{\text{base}}$ and $\text{XLM-R}_{\text{base}}$ on PAWS-X.

that sentence-pair classification focuses on the inference between the two sentences. If we construct the query with only the task label as PMR does, such query does not solely correspond to any meaningful span in the context, and thus is hard to guide the span extraction. Therefore, we leveraged another template "`[CLS]` label Sen-1 `[SEP]` Sen-2 `[SEP]`", where the two sentences are represented separately in the query and the context. In this template, we can extract the exact span from Sen-2 that leads to a contraction or entailment relation (i.e., the task label) with Sen-1. Specifically, we passed the sentence pair to the model twice, with each sentence of the pair being designated as the Sen-2 respectively, and extract the context span with the highest probability score from both sentences.

As shown in Table 3, the extracted spans are indeed important rationales that determine the relationship between two sentences. Such a finding confirms that the extraction capability of mPMR can be appropriately used for explaining the sentence-pair classification process. While the extraction capability may affect the learning of sequence classification during fine-tuning, resulting in a 0.4 Acc. decrease on XNLI.

mPMR Fine-tuning. We investigated the effects of mPMR on XLU fine-tuning. Figure 2 shows that mPMR converges faster than XLM-R on WikiAnn with an extremely low loss value even fine-tuned for 500 steps. In terms of test set performance, mPMR outperforms XLM-R comprehensively and exhibits greater stability. As a result, mPMR provides a better starting point for addressing XLU tasks compared to XLM-R. More examples from XQuAD and PAWS-X are provided in Figure 3 and 4.

5 Conclusions

This paper presents a novel multilingual MRC-style pre-training method, namely mPMR. mPMR provides a unified solver for cross-lingual span extraction and sequence classification and enables direct transfer of NLU capability from pre-training to downstream tasks. mPMR clearly improves the previous baselines and provides a possible solution to explain the sentence-pair classification process.

Limitations

We identify the following two limitations of our work:

- Different from raw text, constructing MRC-style data from Wikipedia requires the existence of hyperlinks. This idea works well for resource-rich languages, such as English and Chinese. While such an idea is less effective for languages with few hyperlink annotations in Wikipedia because a small amount of MRC-style training data is difficult to guide the learning of NLU capability in those languages. A possible solution is to explore other data resources to automatically construct large-scale MRC data for pre-training.

- As observed in Table 1, the improvements of sequence classification tasks are less significant than those of span extraction tasks. We suggest that the existence of anchors is not a strong relevance indicator between our constructed query and context. Such a finding is also observed in Chang et al. (2020). Therefore, constructing more relevant query-context pairs for sequence classification pre-training can possibly remedy this issue.

References

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Giuseppe Attardi. 2015. Wikiextractor. https://github.com/attardi/wikiextractor.

Gábor Berend. 2022. Combating the curse of multilinguality in cross-lingual WSD by aligning sparse contextualized word representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Deng Cai, Xin Li, Jackie Chun-Sing Ho, Lidong Bing, and Wai Lam. 2021. Multilingual AMR parsing with noisy knowledge distillation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*.

Deng Cai, Xin Li, Jackie Chun-Sing Ho, Lidong Bing, and Wai Lam. 2022. Retrofitting multilingual sentence embeddings with Abstract Meaning Representation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

Iacer Calixto, Alessandro Raganato, and Tommaso Pasini. 2021. Wikipedia entities as rendezvous across languages: Grounding multilingual language models by predicting Wikipedia hyperlinks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Wei-Cheng Chang, Felix X. Yu, Yin-Wen Chang, Yiming Yang, and Sanjiv Kumar. 2020. Pre-training tasks for embedding-based large-scale retrieval. In *International Conference on Learning Representations*.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the 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. 2020a. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Bosheng Ding, Junjie Hu, Lidong Bing, Mahani Aljunied, Shafiq Joty, Luo Si, and Chunyan Miao. 2022. GlobalWoZ: Globalizing MultiWoZ to develop multilingual task-oriented dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

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

Xiaoze Jiang, Yaobo Liang, Weizhu Chen, and Nan Duan. 2022. Xlm-k: Improving cross-lingual language model pre-training with multilingual knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Miyoung Ko, Jinhyuk Lee, Hyunjae Kim, Gangwoo Kim, and Jaewoo Kang. 2020. Look at the first sentence: Position bias in question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Juntao Li, Ruidan He, Hai Ye, Hwee Tou Ng, Lidong Bing, and Rui Yan. 2020a. Unsupervised domain adaptation of a pretrained cross-lingual language model. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*.

Xin Li, Lidong Bing, Wenxuan Zhang, Zheng Li, and Wai Lam. 2020b. Unsupervised cross-lingual adaptation for sequence tagging and beyond. *arXiv preprint arXiv:2010.12405*.

Linlin Liu, Xin Li, Ruidan He, Lidong Bing, Shafiq Joty, and Luo Si. 2022. Enhancing multilingual language model with massive multilingual knowledge triples. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

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

Fuli Luo, Wei Wang, Jiahao Liu, Yijia Liu, Bin Bi, Songfang Huang, Fei Huang, and Luo Si. 2021. VECO: Variable and flexible cross-lingual pre-training for

language understanding and generation. 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)*.

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)*.

Jonas Pfeiffer, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. 2022. Lifting the curse of multilinguality by pre-training modular transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*.

Ryokan Ri, Ikuya Yamada, and Yoshimasa Tsuruoka. 2022. mLUKE: The power of entity representations in multilingual pretrained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

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

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

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

Weiwen Xu, Xin Li, Yang Deng, Wai Lam, and Lidong Bing. 2023. Peerda: Data augmentation via modeling peer relation for span identification tasks. In *The 61th Annual Meeting of the Association for Computational Linguistics.*

Weiwen Xu, Xin Li, Wenxuan Zhang, Meng Zhou, Lidong Bing, Wai Lam, and Luo Si. 2022. From clozing to comprehending: Retrofitting pre-trained language model to pre-trained machine reader. *arXiv preprint arXiv:2212.04755*.

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*.

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

Wenxuan Zhang, Ruidan He, Haiyun Peng, Lidong Bing, and Wai Lam. 2021. Cross-lingual aspect-based sentiment analysis with aspect term code-switching. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Meng Zhou, Xin Li, Yue Jiang, and Lidong Bing. 2022a. Enhancing cross-lingual prompting with mask token augmentation. *arXiv preprint arXiv:2202.07255*.

Ran Zhou, Xin Li, Lidong Bing, Erik Cambria, and Chunyan Miao. 2023. Improving self-training for cross-lingual named entity recognition with contrastive and prototype learning. In *The 61th Annual Meeting of the Association for Computational Linguistics.*

Ran Zhou, Xin Li, Lidong Bing, Erik Cambria, Luo Si, and Chunyan Miao. 2022b. ConNER: Consistency training for cross-lingual named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

A Appendix

A.1 More Implementation Details

We collect the 2022-08-01 dump[3] of Wikipedia articles for the 24 languages in consideration. The statistics of each language can be found in Table 4. Then for each article, we extract the plain text with anchors via WikiExtractor (Attardi, 2015). Word tokenization is performed using spaCy[4] if the language is supported, otherwise, we utilize PyThaiNLP[5] for Thai and Sacremoses[6] for remaining languages. For each anchor entity, we construct 10 answerable MRC examples and 10 unanswerable MRC examples as described in Sec. 2.2. Anchor entities with low frequency (below 10 occurrences for English entities and 5 occurrences for entities in other languages) were excluded.

In mPMR, we use Huggingface's implementations of XLM-R (Wolf et al., 2020). During the pre-training stage, the query length Q is set to 50 words, and the context length C is set to 200 words. Both are computed before the subword segmentation. We follow the default learning rate schedule and dropout settings used in XLM-R. We use AdamW (Loshchilov and Hutter, 2019) as our optimizer. We train both $\text{mPMR}_{\texttt{base}}$ and mPMR on 4 A100 GPU. The learning rate is set to 1e-5, and the effective batch size for each step is set to 256 and 80 for $\text{mPMR}_{\texttt{base}}$ and mPMR respectively in order to maximize the usage of the GPU memory. We use the average scores of XQuAD, CoNLL, and PAWS-X to select the best mPMR checkpoint. In fact, we continually pre-train $\text{mPMR}_{\texttt{base}}$ and mPMR for 250,000 and 100,000 steps. The training speed is around 6250 steps per hour. The hyper-parameters of $\text{mPMR}_{\texttt{large}}$ on downstream XLU tasks can be found in Table 5.

A.2 Downstream XLU Tasks

We evaluate mPMR on XLU tasks including both span extraction (EQA, NER, and ABSA) and sequence classification (sentence pair classification). We follow (Xu et al., 2022) to convert all tasks into MRC formulation and tackle them accordingly. We show concrete examples for each task in Table 6. Specifically, we evaluate the performance of EQA on three benchmarks: XQuAD (Artetxe et al., 2020), MLQA (Lewis et al., 2020), and TyDiQA (Clark et al., 2020) covering 11, 7, and 9 languages respectively. For NER evaluation, we use the WikiAnn dataset (Pan et al., 2017) restricted to the 40 languages from XTREME (Hu et al., 2020), as well as the CoNLL dataset with 4 languages (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003); We also evaluate the XLU performance of SemEval16 ABSA on 6 languages (Pontiki et al., 2016), where we collect the data from Li et al. (2020b); Zhang et al. (2021). Regarding the sequence classification task, we evaluate XNLI (Conneau et al., 2018) and PAWS-X (Yang et al., 2019) with 15 and 7 languages respectively.

A.3 mPMR Performance per Language

We show the detailed results for each language in each task in Table 7 (XQuAD), Table 8 (MLQA), Table 9 (TyDiQA), Table 10 (WikiAnn), Table 11 (CoNLL), Table 12 (SemEval16), Table 13 (PAWS-X), and Table 14 (XNLI).

[3] https://dumps.wikimedia.org/enwiki/latest
[4] https://github.com/explosion/spaCy
[5] https://github.com/PyThaiNLP/pythainlp
[6] https://github.com/alvations/sacremoses

Language	# Entities	# MRC examples	Language	# Entities	# MRC examples
ar	118,292	2,020,502	ko	94,616	1,597,076
bn	25,081	410,634	nl	251,323	4,185,913
de	864,746	14,795,826	pl	283,925	4,765,015
el	56,383	946,114	pt	216,695	3,648,603
en	966,197	19,303,940	ru	432,437	7,342,472
es	412,476	7,044,972	sv	169,030	2,808,214
fi	113,118	1,960,636	sw	4,857	65,724
fr	595,879	10,164,216	te	11,005	170,664
hi	15,350	242,078	th	31,676	522,434
id	70,960	1,164,662	tr	71,294	1,175,276
it	376,417	6,421,850	vi	68,665	1,147,772
ja	423,884	7,338,308	zh	259,785	4,438,004
			Total	5,934,091	103,680,905

Table 4: Data statistics of mPMR pre-training data. The statistics is computed after removing the low-frequency entities. The number of MRC examples includes both answerable and unanswerable examples.

Dataset	**XQuAD**	**MLQA**	**TyDiQA**	**WikiAnn**	**CoNLL**	**SemEval16**	**PAWS-X**	**XNLI**
Query Length	64	64	64	32	32	32	64	64
Input Length	384	384	384	192	192	192	192	192
Batch Size	8	8	8	16	16	32	16	32
Learning Rate	3e-5	3e-5	2e-5	1e-5	1e-5	2e-5	5e-5	3e-5
Epoch	3	3	10	10	10	20	10	3

Table 5: Hyper-parameters settings in fine-tuning XLU tasks.

Task		Example Input	Example Output
EQA (XSQuAD)	Ori.	Question: Who lost to the Broncos in the divisional round? Context: The Broncos defeated the Pittsburgh Steelers in the divisional round, 23–16, by scoring 11 points in the final three minutes of the game.	Answer: "Pittsburgh Steelers"
	PMR	[CLS] Who lost to the Broncos in the divisional round ? [SEP] [SEP] The Broncos defeated the Pittsburgh Steelers in the divisional round, 23–16 , by scoring 11 points in the final three minutes of the game . [SEP]	(17,18) - "Pittsburgh Steelers"
NER (CoNLL)	Ori.	Two goals in the last six minutes gave holders Japan an uninspiring 2-1 Asian Cup victory over Syria on Friday.	("Japan", LOC); ("Syria", LOC); ("Asian Cup", MISC)
	PMR	[CLS] "ORG" . Organization entities are limited to named corporate, governmental, or other organizational entities. [SEP] [SEP] Two goals in the last six minutes gave holders Japan an uninspiring 2-1 Asian Cup victory over Syria on Friday . [SEP]	∅
		[CLS] "PER" . Person entities are named persons or family . [SEP] [SEP] Two goals in the last six minutes gave holders Japan an uninspiring 2-1 Asian Cup victory over Syria on Friday . [SEP]	∅
		[CLS] "LOC" . Location entities are the name of politically or geographically defined locations such as cities , countries . [SEP] [SEP] Two goals in the last six minutes gave holders Japan an uninspiring 2-1 Asian Cup victory over Syria on Friday . [SEP]	(32,32) - "Japan"; (40,40) - "Syria"
		[CLS] "MISC" . Examples of miscellaneous entities include events , nationalities , products and works of art . [SEP] [SEP] Two goals in the last six minutes gave holders Japan an uninspiring 2-1 Asian Cup victory over Syria on Friday . [SEP]	(34,35) - "Asian Cup"
ABSA (SemEval16)	Ori.	Nice ambience, but highly overrated place.	("ambience", POS); ("place", NEG)
	PMR	[CLS] "POS" . For aspect terms of positive sentiment . [SEP] [SEP] Nice ambience , but highly overrated place . [SEP]	(13,13) - "ambience"
		[CLS] "NEG" . For aspect terms of negative sentiment . [SEP] [SEP] Nice ambience , but highly overrated place . [SEP]	(18,18) - "place"
		[CLS] "NEU" . For aspect terms of neutral sentiment . [SEP] [SEP] Nice ambience , but highly overrated place . [SEP]	∅
Sen. Pair Classification (PAWS-X)	Ori.	Hypothesis: The Tabaci River is a tributary of the River Leurda in Romania. Premise: The Leurda River is a tributary of the River Tabaci in Romania.	Contradiction
	PMR	[CLS] Contradiction . The hypothesis is a sentence with a contradictory meaning to the premise . [SEP] [SEP] Hypothesis : The Tabaci River is a tributary of the River Leurda in Romania . Premise : The Leurda River is a tributary of the River Tabaci in Romania . [SEP]	(0,0) - "[CLS]"
		[CLS] Entailment . The hypothesis is a sentence with a similar meaning as the premise . [SEP] [SEP] Hypothesis : The Tabaci River is a tributary of the River Leurda in Romania . Premise : The Leurda River is a tributary of the River Tabaci in Romania . [SEP]	∅

Table 6: MRC examples of XLU tasks. We use English examples here for demonstration purposes. Ori. indicates the original data format of these tasks.

Model	en	ar	de	el	es	hi	ru	th	tr	vi	zh	Avg.
XLM-R_{base}	82.2 / 72.0	65.5 / 49.9	73.9 / 59.7	71.2 / 56.3	76.3 / 59.4	66.4 / 52.0	73.7 / 58.9	64.7 / 54.6	67.0 / 52.8	73.3 / 54.7	65.0 / 55.9	70.8 / 56.9
mPMR$_{base}$	84.4 / 73.4	69.6 / 53.2	76.4 / 61.5	74.9 / 58.4	77.4 / 60.2	69.2 / 54.5	75.2 / 58.8	69.2 / 57.6	70.4 / 55.8	74.8 / 55.8	71.8 / 65.5	74.0 / 59.5
XLM-R	86.5 / 75.6	72.4 / 54.8	79.3 / 63.0	79.2 / 61.6	82.0 / 62.9	76.1 / 59.1	79.0 / 62.9	72.2 / 59.8	75.4 / 60.8	79.7 / 60.8	68.2 / 58.2	77.3 / 61.7
mPMR	87.6 / 76.5	75.9 / 60.0	81.5 / 65.0	80.8 / 63.9	82.8 / 65.1	76.5 / 60.3	80.9 / 65.3	75.5 / 65.5	76.7 / 61.3	81.5 / 62.2	71.5 / 63.4	79.2 / 64.4

Table 7: XQuAD results (F1 / EM) for each language.

Model	en	ar	de	es	hi	vi	zh	Avg.
XLM-R$_{\texttt{base}}$	79.3 / 67.2	55.4 / 38.1	62.0 / 49.1	66.8 / 50.2	59.4 / 44.8	66.1 / 46.7	61.8 / 39.5	64.4 / 47.9
mPMR$_{\texttt{base}}$	81.1 / 68.9	58.5 / 41.0	63.6 / 50.5	68.5 / 52.1	60.3 / 46.4	68.3 / 49.2	56.6 / 32.9	65.3 / 48.7
XLM-R	83.4 / 71.0	64.9 / 45.8	69.6 / 54.8	74.1 / 56.8	70.7 / 53.4	73.3 / 53.0	64.4 / 42.4	71.5 / 53.9
mPMR	84.0 / 71.4	66.4 / 47.0	70.3 / 56.2	74.5 / 57.1	71.4 / 54.1	74.7 / 54.4	70.5 / 47.3	73.1 / 55.4

Table 8: MLQA results (F1 / EM) for each language.

Model	en	ar	bn	fi	id	ko	ru	sw	te	Avg.
XLM-R$_{\texttt{base}}$	66.8 / 57.3	55.7 / 42.0	31.5 / 20.4	52.6 / 40.3	69.1 / 55.6	36.3 / 27.9	54.8 / 36.5	53.0 / 34.7	37.4 / 28.8	50.8 / 38.2
mPMR$_{\texttt{base}}$	71.1 / 61.6	66.3 / 52.6	56.5 / 41.6	65.5 / 53.1	73.9 / 63.7	50.4 / 38.8	64.4 / 37.9	57.4 / 41.1	65.3 / 50.4	63.4 / 49.0
XLM-R	71.3 / 60.7	69.3 / 52.3	66.2 / 53.1	64.3 / 51.3	76.5 / 62.5	58.3 / 46.7	64.7 / 43.4	68.6 / 53.1	67.3 / 41.1	67.4 / 51.6
mPMR	76.4 / 65.2	76.0 / 58.0	72.3 / 55.8	74.4 / 56.5	84.1 / 71.3	62.2 / 50.7	72.5 / 43.2	76.5 / 63.1	77.7 / 60.8	74.7 / 58.3

Table 9: TyDiQA-GoldP results (F1 / EM) for each language.

Model	en	af	ar	bg	bn	de	el	es	et	eu	fa	fi	fr	he	hi	hu	id	it	ja	jv
XLM-R$_{\texttt{base}}$	84.2	75.3	47.3	79.0	66.3	77.5	75.3	78.0	69.6	56.0	38.1	70.4	81.4	50.8	67.9	72.4	51.0	79.6	19.6	63.9
mPMR$_{\texttt{base}}$	85.1	80.7	57.6	80.2	71.9	81.2	77.6	79.5	79.1	71.3	49.6	80.4	82.4	65.2	71.7	82.2	58.6	83.5	43.2	72.0
XLM-R	85.4	81.1	53.9	84.0	73.8	82.3	82.8	80.4	68.8	54.8	64.2	75.9	81.4	59.3	72.9	76.4	59.3	84.6	13.2	71.2
mPMR	86.0	81.7	56.1	85.9	79.6	82.3	82.3	75.5	82.7	69.6	75.2	84.1	82.0	66.5	75.9	84.0	59.9	86.1	49.1	72.4
	ka	kk	ko	ml	mr	ms	my	nl	pt	ru	sw	ta	te	th	tl	tr	ur	vi	yo	zh
XLM-R$_{\texttt{base}}$	58.7	40.6	34.3	50.8	46.0	63.8	40.6	81.5	80.0	65.4	76.1	43.0	46.4	4.2	71.9	68.7	45.7	70.9	1.5	23.0
mPMR$_{\texttt{base}}$	72.2	45.1	52.9	62.4	59.4	68.1	57.4	83.7	81.5	71.8	77.3	50.5	57.4	3.0	74.2	80.3	55.7	75.2	31.6	49.9
XLM-R	59.9	41.7	41.3	56.8	58.2	76.7	29.6	86.1	85.2	72.2	77.6	52.3	51.6	7.1	78.8	70.9	64.0	80.0	27.2	22.4
mPMR	77.3	46.8	57.9	70.6	68.1	73.8	57.8	86.0	83.6	72.8	79.8	62.6	58.1	3.8	83.0	80.3	76.2	83.6	36.1	54.4

Table 10: WikiAnn results (F1 Score) for each language.

Model	en	de	es	nl	Avg.
XLM-R$_{\texttt{base}}$	91.3	71.0	78.7	75.7	79.2
mPMR$_{\texttt{base}}$	91.9	74.3	80.8	79.7	81.7
XLM-R	92.8	73.7	81.6	77.7	81.4
mPMR	93.5	75.0	85.0	83.1	84.1

Table 11: CoNLL results (F1 Score) for each language.

Model	en	es	fr	nl	ru	tr	Avg.
XLM-R$_{\texttt{base}}$	76.5	65.4	55.6	61.2	56.1	45.4	60.0
mPMR$_{\texttt{base}}$	77.6	68.6	56.4	62.2	59.5	48.4	62.1
XLM-R	82.4	71.3	60.3	67.4	61.2	49.1	66.1
mPMR	82.8	71.9	64.7	67.4	66.9	55.7	68.2

Table 12: SemEval16 results (F1 Score) for each language.

Model	en	de	es	fr	ja	ko	zh	Avg.
XLM-R$_{\texttt{base}}$	94.3	87.7	89.1	88.7	77.0	76.6	81.3	85.0
mPMR$_{\texttt{base}}$	94.3	88.4	90.1	88.9	79.0	79.4	82.4	86.1
XLM-R	95.2	89.3	91.0	90.9	79.6	79.9	82.5	86.9
mPMR	95.2	90.6	90.3	91.3	81.2	82.9	84.6	88.0

Table 13: PAWS-X accuracy scores (Acc.) for each language.

Model	en	ar	bg	de	el	es	fr	hi	ru	sw	th	tr	ur	vi	zh	Avg.
XLM-R$_{\texttt{base}}$	84.6	71.0	76.8	75.6	74.9	77.9	76.9	68.9	74.1	64.4	71.1	72.4	65.2	73.2	73.0	73.3
mPMR$_{\texttt{base}}$	84.2	71.5	77.2	75.5	75.5	78.6	76.9	69.5	74.7	62.5	71.4	71.6	65.5	74.3	74.0	73.6
XLM-R	88.2	77.0	81.7	81.2	81.2	84.2	81.7	74.9	78.9	70.8	75.7	77.4	70.6	78.0	77.7	78.6
mPMR	88.3	77.9	82.9	82.2	81.0	83.5	82.2	75.2	79.8	71.2	76.1	78.9	71.6	78.9	79.0	79.3

Table 14: XNLI accuracy scores (Acc.) for each language.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section Limitations

☐ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract, Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 3

☑ B1. Did you cite the creators of artifacts you used?
Section 3

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 3, Appendix A.1

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Appendix A.1, Appendix A.2

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Appendix A.1

C ☑ Did you run computational experiments?

Section 4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 3, Appendix A.1

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix A.1

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 3, Appendix A.1

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.