Sliced at SemEval-2022 Task 11: Bigger, Better? Massively Multilingual LMs for Multilingual Complex NER on an Academic GPU Budget

Barbara Plank
Center for Information and Language Processing (CIS), LMU Munich, Germany
Department of Computer Science, ITU Copenhagen, Denmark
bplank@itu.dk

Abstract

Massively multilingual language models (MMLMs) have become a widely-used language representation method, and multiple large MMLMs were proposed in recent years. A trend is to train MMLMs on larger text corpora or with more layers. In this paper we set out to evaluate recent popular MMLMs on detecting semantically ambiguous and complex named entities with an academic GPU budget. Our submission of a single model for 11 languages on the SemEval Task 11 MultiCoNER shows that a fine-tuned XLM-R$_{large}$ outperforms the more recent RemBERT, ranking 9th from 26 submissions in the multilingual track. Compared to RemBERT, the XLM-R model has the additional advantage to fit on a slice of a multi-instance GPU. As contrary to expectations and recent findings, we found RemBERT to not be the best MMLM, we further set out to investigate this discrepancy with additional experiments on multilingual Wikipedia NER data. While we expected RemBERT to have an edge on that dataset as it is closer to its pre-training data, surprisingly, our results show that this is not the case.

1 Introduction

Pre-trained language models have revolutionized the field of Natural Language Processing (NLP) in recent years (Peters et al., 2018; Devlin et al., 2019; Zhuang et al., 2021). Especially for cross-lingual transfer learning or creating a single multilingual model, pre-trained massively multilingual language models (MMLMs) have become a de-facto standard (Conneau et al., 2020; Chung et al., 2021).

MMLMs such as BERT and XLM-R share the same underlying idea: multilingual representations are obtained by learning from large text collections in multiple languages and are trained using a language modeling objective. MMLMs, in contrast to alternative cross-lingual transfer strategies, thus do not rely on explicit alignment via parallel data and explicit transfer via translations with e.g. annotation projection. MMLMs, together with the pre-training and fine-tuning paradigm, have enabled impressive results (Conneau et al., 2020; Lauscher et al., 2020; Müller-Eberstein et al., 2021).

This paper describes our submission to Task 11 on Multilingual Complex Named Entity Recognition (MultiCoNER) (Malmasi et al., 2022b,a). We evaluate several recent MMLMs in a fine-tuning regime to answer the following main research question (RQ1): *To what extent are more recent larger LMs outperforming earlier MMLMs for the task of multilingual complex NER?* To do so, we test four MMLMs (mBERT, XLM-R base and large and the most recently proposed RemBERT). As other NER datasets exists, albeit with different labels, we further explore multi-task learning (RQ2): *To what extent can we improve upon MultiCoNER by using cross-lingual cross-domain NER data as auxiliary data?* Finally, as the MMLMs were pre-trained on different kinds of data, we include experiments on a second NER dataset, to answer RQ3: *To what extent does RemBERT outperform XLM-R when the pre-training data is a closer match?*

Our contributions are:

- We train a single multilingual model on MultiCoNER, which to our knowledge, is the largest multilingual NER evaluation campaign to date in terms of manually-annotated multilingual training data availability. We compare four MMLMs for the task and examine task performance and GPU budget (here: 20gb).
- Surprisingly, we find that RemBERT does not work well. To shed more light on this, we run additional experiments on a NER benchmark which is closer to RemBERT's pre-training data and prior work. Overall we find XLM-R$_{large}$ to provide the best performance and a good space/training-time trade-off.

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 1494 - 1500

Model	Model Name	Language Variety	Languages	Vocab	H_dim	Layers	Params
mBERT	bert-base-multilingual-cased	Wikipedia	104	120k	768	12	110M
XLM-R$_{base}$	xlm-roberta-base	CommonCrawl	100	250k	768	12	270M
XLM-R$_{large}$	xlm-roberta-large	CommonCrawl	100	250k	1024	24	550M
RemBERT	google/rembert	CommonCrawl+Wikipedia	110	250k	1152	32	559M

Table 1: Overview of pre-trained massively multilingual language models (MMLMs) used in this work. All 12 languages used in MultiCoNer are part of the pre-training data of all MMLMs.

2 Experimental Setup

This section describes the model, all data sets and the multilingual language models used in this work.

2.1 Transformer-CRF

We use a transformer-CRF model for NER, implemented in the MaChAmp toolkit (van der Goot et al., 2021) (v0.3 beta). The toolkit enables easy exchange of pre-trained LM for fine-tuning as well as multi-task learning. Our model is a single MMLM fine-tuned with a single CRF decoder.

To train MaChAmp, we use the proposed default parameters (van der Goot et al., 2021), which have shown to work well across tasks with a learning rate of $lr = 0.0001$. We train the model for 20 epochs and select the best checkpoint using the provided dev data (`multi_dev`).

2.2 MMLMs

As our main RQ is to test the effect of using different pre-trained massively multilingual language models, we opted for four well-known variants: multilingual BERT (Devlin et al., 2019), XLM-R with both the $base$ (XLM-R$_b$) and $large$ (XLM-R$_l$) version (Conneau et al., 2020), and the more recently introduced RemBERT (Chung et al., 2021).

An overview of the MMLMs is provided in Table 1, and summarized as follows:

- mBERT: trained using both Masked-Language Model (MLM) and next-sentence prediction tasks on Wikipedia data and trained with an exponentially decaying smoothing distribution over languages with $\alpha = 0.7$, i.e., to down-scale high-resource languages and up-scale lower-resource languages; 12 layers.

- XLM-R: trained using only MLM on CommonCrawl data and trained with an exponentially decaying smoothing distribution with $\alpha = 0.3$ (more aggressive smoothing); with 12 (base) or 24 (large) layers.

- RemBERT: trained using only MLM[1] on Wikipedia and CommonCrawl data, with an exponentially decaying smoothing distribution $\alpha = 0.5$; trained with decoupled input and output embeddings and parameters re-distributed over 32 layers.

We observe that the MMLMs are trained on 100, 104 and 110 languages, where RemBERT is the MMLM with the largest amount of language coverage (110). We note that for MultiCoNER, all 11 target languages are included in the the pre-training material of all four MMLMs. What differs amongst others is the vocabulary size, the number of layers, the pre-training method, the pre-training data and the number of parameters. XLM$_{large}$ and RemBERT are closest in terms of total number of parameters, with RemBERT having an additional 9M parameters over XLM-R$_{large}$; mBERT and XLM-R$_{base}$ are a fifth and half of the number of parameters, respectively. What stands out is the type of pre-training data, as the language variety that these MMLMs were trained on differs. While the ones created by researchers at Google opted mainly for Wikipedia data (mBERT, RemBERT), the XLM-R models from FAIR research are trained on a cleaned CommonCrawl dump (Wenzek et al., 2020).

2.3 MultiCoNER Data

The training dataset provided by MultiCoNER organizers is one of the largest and most diverse multilingual NER datasets available to date in terms of training/dev/test sizes and language/domain coverage (presumably manually annotated).[2]

In contrast, the Panx (WikiAnn) dataset (Pan et al., 2017; Rahimi et al., 2019) covers a much larger span of languages (over 200) but its coverage is limited to Wikipedia data, and it contains fewer entity types and semi-automatic annotation. Moreover, MultiCoNER is set up for complex NER,

[1] shorturl.at/pwyFQ

[2] At the time of writing this system paper we did not have further information on how the data was annotated.

Multilingual	Train	dev	test
sentences	168.3k	8,800	472k
token	2,752,814	144,641	4,565,160
word types	360,679	47,570	556,102
type/token	0.13	0.33	0.12
entities	237,212	12,513	n/a
PER	43,953	2,342	n/a
PROD	25,001	1,848	n/a
CORP	32,890	1,738	n/a
CW	38,373	2,015	n/a
LOC	54,030	2,932	n/a
GRP	32,846	1,638	n/a

Table 2: Overview of the MultiCoNER dataset. Every sentence in train and dev contains at least one entity.

Model	micro F1	train h	GPU m	fits?
$lr = 0.00001$				
mBERT	81.31	4h20	8gb	✓
RemBERT	83.89	7h40	27gb	✗
$lr = 0.0001$				
RemBERT	85.03	7h30	21gb	✗
XLM-R$_b$	83.89	7h13	10gb	✓
XLM-R$_l$	**86.32**	10h30	**16gb**	✓
RemBERT+aux	85.00	9h30	23gb	✗
XLM$_l$+aux	85.89	11h30	19gb	✓

Table 3: Result of training the transformer-CRF with different MMLMs on the multilingual development set. Training time (in hours), GPU max memory usage (approximate) and whether the training fits a single slice of a A100 GPU (the GPU was partitioned into two 20gb slices using NVIDIA's mig mode). XLM$_l$ takes more time but less memory, which enables parallel training of two XLM$_l$ models on a single sliced GPU.

which include further challenges such as detecting named entities on search queries and code-mixed data (Meng et al., 2021; Fetahu et al., 2021). MultiCoNER contains 6 entity types (person, location, product, corporation, groups but also complex entities) for the following 11 languages: English (en), Spanish (es), Dutch (nl), Russian (ru), Turkish (tr), Korean (ko), Persian/Farsi (fa), German (de), Chinese (zh), Hindi (hi), and Bangla (bn). While we focus on the multilingual track, and the data provided for it (`multi`), the shared task further features monolingual tracks and a code-mixed track containing data of some of these languages.

Table 2 provides an overview of the MultiCoNER multilingual data. It is a very large training set of 168.3k sentences, 2.7M tokens and over 237k entities spanning 11 languages. Table 6 in Table 2 shows the size of the dev and test portions (note the very large test data with 472k sentences) and the distribution over the 6 entity types in the training and dev data. Location (LOC) and person names (PER) constitutes the largest portion of entities, followed by CW, corporate names (CORP) and group names (GRP) and the least frequent entity type is product names (PROD). The appendix lists sizes of individual language test files.

2.4 Auxiliary or Matching Data?

For RQ2, we use a multi-task learning setup, and model the MultiCoNER task as the main task with a CRF output decoder, and add a second CRF-decoder that predicts NER types from the union of three cross-lingual cross-domain datasets. In particular, we use German data (Benikova et al., 2014), and two recently proposed derivatives annotated on top of Danish (Plank et al., 2020) and EWT-NNER on top of the English Web Treebank (Plank, 2021). While all three data sources were annotated for nested NERs (a two-level annotation scheme, which annotates e.g., a location for "Birmingham" inside "University of Birmingham"), we here use only their inner layer entities. In addition, these data contain slightly different entity types with finer-grained subsets: location, person, organization and miscellaneous entities with two additional suffixes that add derivations (e.g., adjectival forms like "Brasilian" and partial NERs like "Nintendo-based"). The total auxiliary training data (we take the union of English, Danish and German data) consists of a total of 21.4k sentences (roughly 8% of the MultiCoNER main task data) with 42.6k entity annotations. We train the multi-task model jointly by full-parameter sharing, no loss weighting and selecting the best checkpoint using the sum over both main and auxiliary task development span-F1.

For RQ3, we compare the two best MMLMs on a NER benchmark derived from Wikipedia (Pan et al., 2017) (WikiAnn/Panx). We follow the setup of Lauscher et al. (2020) and use 12 languages: Indian (in), English (en), Arabic (ar), Finnish (fi), Hindi (hi), Japanese (ja), Russian (ru), Turkish (tr), Basque (eu), Hebrew (he), Italian (it), Korean (ko), Swedish (sv) and Chinese (zh), which use the same splits as RemBERT (Chung et al., 2021), which were provided by Rahimi et al. (2019).

2.5 NER evaluation: macro-F1 vs micro-F1

For evaluation and model selection, we use the CoNLL span-F1 which is a micro-F1 over span-based entity F1 scores. We note that some earlier work report Accuracy, which can be misleading for NER due to the high number of non-entity tokens typical for the task. The MultiCoNER organizers opted for span-based *macro* F1. While both micro and macro F1 consider entities correct only if both the entity boundaries and the labels match, the MultiCoNER macro entity-based F1 is typically lower and hence a more conservative measure, particularly when entity types are unbalanced, which is the case for MultiCoNER. Hence, we adapted a Python-version of the Conlleval scorer to include macro-F1.[3] We report macro-F1 for the aggregated evaluation measures.

3 Results

Table 3 provides the main results of a single model trained on the MultiCoNER `multi_train` data and evaluated on the `multi_dev` portion.

Bigger is not always better From the MMLM comparison in Table 3 we first observe that mBERT is outperformed substantially by more recent MMLMs. As expected, XLM-R$_l$ outperforms XLM-R$_b$. However, regarding RQ1 our results show that bigger is not always better: XLM-R$_l$ outperforms the more recent and even larger RemBERT model. Surprisingly, this is consistently the case over all target languages.

XLM-R$_l$ is more space efficient While XLM-R$_l$ is the best MMLM in terms of F1 on multilingual complex NER, for an academic GPU budget XLM-R$_l$ is also more efficient: it still just fits a single 20gb GPU slice (NVIDIA A100 40GB GPU split into two slices), at a cost of slightly longer training duration. This means we can fine-tune 2 XLM-R large models in parallel, while only one RemBERT. Further details are given in Table 3.

Test set results Table 4 provides the results on the MultiCoNER test set (last column, `multi_test`). The results confirm the findings from dev: XLM-R$_l$ results in the best model, and substantially outperforms RemBERT. This model ranks 9th in the multilingual track out of 26 participating systems. The auxiliary task (RQ2) fails to provide additional signal, which we hypothesize might be due to the high-resource setup (already a high amount of training data exists), but this would need further investigation. While `multi_test` already contains test data from all 11 languages, the shared task provides further monolingual (and code-mixed) test sets, which is not equal to the sum in `multi_test`. We submitted runs of the best models on these individual test sets which confirm that XLM-R$_l$ remains the strongest MMLM for NER on MultiCoNER.

[3] Available at https://github.com/bplank/conlleval

4 Discussion

In this section, we provide a deeper analysis of our results, adding an additional experiment on data outside of the shared task. First, we examine the per-class F1 score on MultiCoNER, to rule out that the strong results of XLM-R$_l$ are purely due to strong results on a few frequent entity types. Second, we compare RemBERT vs XLM-R$_l$ on a Wikipedia NER dataset to answer RQ3. Previous findings suggest that RemBERT is a stronger model for multilingual NER (Chung et al., 2021) by testing it on *on Wikipedia* data, which is also closer to its pre-training data. So we test whether this is the case with our model as well.

Test set results per entity Figure 1 provides a breakdown of the three best models, showing the per-class F1 scores for RemBERT, XLM-R$_l$ and XLM-R$_l$+aux. The results show that XLM-R$_l$ outperforms RemBERT over all six entity classes, and that the multi-task setup consistently hurts over all entity types.

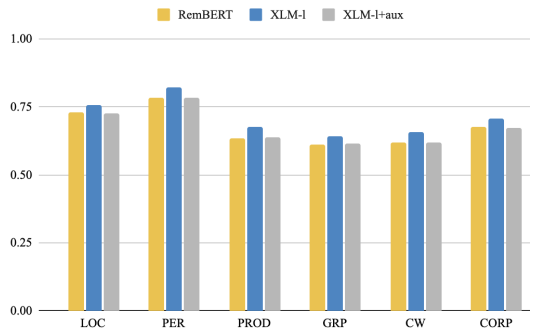

Figure 1: Per-class F1 score on the multilingual test set.

Bigger and better because of better pre-training match? Contrary to our findings, previous work suggests that RemBERT outperforms XLM-R on NER (Chung et al., 2021). We hypothesize that this is not the case on MultiCoNER due to the more

macro-F1	bn	de	en	es	fa	hi	ko	nl	ru	tr	zh	mix	multi
RemBERT	57.14	77.94	73.39	73.62	61.49	60.77	64.59	76.24	70.99	64.97	62.46	70.99	67.61
RemBERT+aux	58.06	77.19	73.27	73.48	62.39	60.09	65.03	75.55	71.10	65.42	62.61	70.79	67.66
XLM-R$_l$	63.05	78.90	74.54	75.11	68.66	67.00	70.66	77.66	73.73	68.77	65.21	72.74	**71.07**

Table 4: Results on the test set. XLM$_l$ is the model that achieved the 9th rank in the multilingual track. Evaluation of this model on all single language and the code-mixed test set included for completeness.

diverse data. A manual inspection reveals that the MulitCoNER data includes both Wikipedia-style data, but also questions without question symbols and short search queries. Therefore, we investigate whether RemBERT instead outperforms XLM-R on a Wikipedia NER benchmark, which matches RemBERT's pre-training data better, cf. Table 1.

	EN	ZH	TR	RU	AR	HI	EU
RB	86.2	82.4	94.1	90.9	90.7	90.7	92.9
XLM	85.9	83.0	93.9	91.3	91.5	92.2	93.2
	FI	HE	IT	JA	KO	SV	avg
RB	92.0	89.5	93.2	76.6	89.6	95.6	89.5
XLM	92.5	90.3	93.1	77.9	90.9	95.9	**90.1**

Table 5: RemBERT (RB) vs XLM-R$_l$ (XLM) on Wikipedia/Panx. Except for EN, TR and IT, XLM-R outperforms RB. Macro-averaged span micro F1 of 90.1.

Table 5 show the results on 13 languages (averaged over 3 runs). The results show that RemBERT has a slight advantage on 3 out of the 13 languages (EN, TR, IT), but overall and on 9 out of the 13 languages XLM-R$_l$ performs substantially better. This additional experiment surprisingly dis-confirms our hypothesis that RemBERT would have an advantage on this Wikipedia/Panx NER data due to the better pre-training data match. While this is in contrast to previous findings on Panx (Chung et al., 2021), the reason is less clear. We studied the literature and found a study on another task (quality estimation) that similarly reports negative results: replacing the XLM-R decoder with a RemBERT decoder performs better only for 1 language pair out of 4 in quality estimation (Treviso et al., 2021). We conclude that XLM-R$_l$ remains the best MMLM for multlingual NER, as tested across two benchmarks (MultiCoNER and WikiAnn).

5 Limitations

We provided a study on four MMLMs for transformer-based multilingual complex NER to shed lights on MMLMs and GPU usage. Our study is limited in number of MMLMs tested, and the fact that we provide only approximate GPU memory consumption figures.

6 Conclusions

We test four massively multlingual language models as encoders for multilingual complex NER. Our results show that XLM-R$_l$ results in the overall best model, and surprisingly outperforms the more recently proposed RemBERT, also in terms of GPU memory consumption. While auxiliary-task training did not further prove promising, we additionally studied the discrepancy between RemBERT and XLM-R on a second benchmark (PANX/WikiAnn data). While we hypothesized that RemBERT would outperform XLM-R, our results show that this is not the case. Overall, a bigger model might not be the best choice, especially not for an academic GPU budget.

Code, scripts and shared task prediction files (labels only) available at: https://github.com/bplank/multiconer2022

Acknowledgements

I would like to thank Rob van der Goot for feedback on earlier drafts of this paper. Thanks to my DFF project number 9063-00077B for providing the GPUs used in this work.

References

Darina Benikova, Chris Biemann, and Marc Reznicek. 2014. NoSta-D named entity annotation for German: Guidelines and dataset. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2524–2531, Reykjavik, Iceland. European Language Resources Association (ELRA).

Hyung Won Chung, Thibault Févry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2021. Rethinking embedding coupling in pre-trained language models. In *ICLR*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised

cross-lingual representation learning at scale. In *Proceedings of 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.

Besnik Fetahu, Anjie Fang, Oleg Rokhlenko, and Shervin Malmasi. 2021. Gazetteer Enhanced Named Entity Recognition for Code-Mixed Web Queries. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1677–1681.

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.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022a. MultiCoNER: a Large-scale Multilingual dataset for Complex Named Entity Recognition.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022b. SemEval-2022 Task 11: Multilingual Complex Named Entity Recognition (MultiCoNER). In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Tao Meng, Anjie Fang, Oleg Rokhlenko, and Shervin Malmasi. 2021. GEMNET: Effective gated gazetteer representations for recognizing complex entities in low-context input. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1499–1512.

Max Müller-Eberstein, Rob van der Goot, and Barbara Plank. 2021. Genre as weak supervision for cross-lingual dependency parsing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4786–4802, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Barbara Plank. 2021. Cross-lingual cross-domain nested named entity evaluation on English web texts. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1808–1815, Online. Association for Computational Linguistics.

Barbara Plank, Kristian Nørgaard Jensen, and Rob van der Goot. 2020. DaN+: Danish nested named entities and lexical normalization. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6649–6662, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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.

Marcos Treviso, Nuno M. Guerreiro, Ricardo Rei, and André F. T. Martins. 2021. IST-unbabel 2021 submission for the explainable quality estimation shared task. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 133–145, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online. Association for Computational Linguistics.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

A Data statistics MultiCoNER

Lang	Sentences/entities	
	dev	test
ba	800/800	133,119/-
de	800/1,239	217,824/-
en	800/1,230	217,818/-
es	800/1,176	217,887/-
fa	800/1,213	165,702/-
hi	800/828	141,565/-
ko	800/1,302	178,249/-
nl	800/1,157	217,337/-
ru	800/1,042	217,501/-
tr	800/1,245	136,935/-
zh	800/1,281	151,661/-
multi	8,800/12,513	471,911/-

Table 6: Data statistics of dev/test of MultiCoNER.