Preserving Cross-Linguality of Pre-trained Models via Continual Learning

Zihan Liu, Genta Indra Winata, Andrea Madotto, Pascale Fung
Center for Artificial Intelligence Research (CAiRE)
Department of Electronic and Computer Engineering
The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong
zihan.liu@connect.ust.hk

Abstract

Recently, fine-tuning pre-trained language models (e.g., multilingual BERT) to downstream cross-lingual tasks has shown promising results. However, the fine-tuning process inevitably changes the parameters of the pre-trained model and weakens its cross-lingual ability, which leads to sub-optimal performance. To alleviate this problem, we leverage continual learning to preserve the original cross-lingual ability of the pre-trained model when we fine-tune it to downstream tasks. The experimental result shows that our fine-tuning methods can better preserve the cross-lingual ability of the pre-trained model in a sentence retrieval task. Our methods also achieve better performance than other fine-tuning baselines on the zero-shot cross-lingual part-of-speech tagging and named entity recognition tasks.

1 Introduction

Recently, multilingual language models (Devlin et al., 2019; Conneau and Lample, 2019), pre-trained on extensive monolingual or bilingual resources across numerous languages, have been shown to enjoy surprising cross-lingual adaptation abilities, and fine-tuning them to downstream cross-lingual tasks has achieved promising results (Pires et al., 2019; Wu and Dredze, 2019). Taking this further, better pre-trained language models have been proposed to improve the cross-lingual performance, such as using larger amounts of pre-trained data with larger pre-trained models (Conneau et al., 2019; Liang et al., 2020), and utilizing more tasks in the pre-training stage (Huang et al., 2019).

However, we observe that multilingual BERT (mBERT) (Devlin et al., 2019), a pre-trained language model, forgets the masked language model (MLM) task that has been learned and partially loses the cross-lingual ability (from a cross-lingual sentence retrieval (XSR)[1] experiment) after being fine-tuned to the downstream task in English, as shown in Figure 1, which results in sub-optimal cross-lingual performance to target languages.

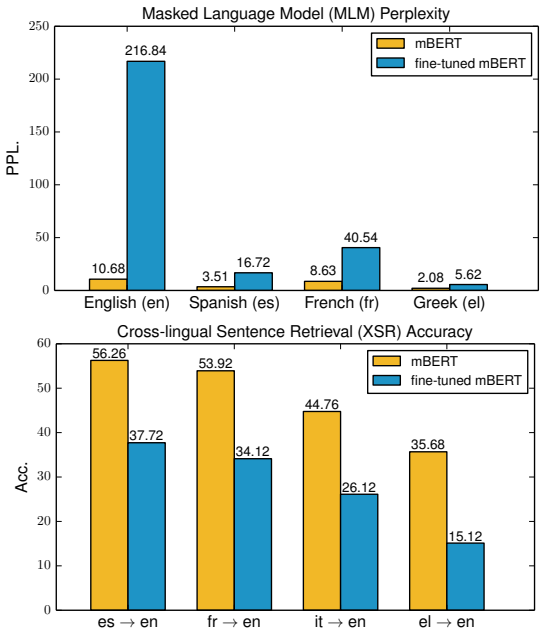

Figure 1: Masked language model and cross-lingual sentence retrieval results before and after fine-tuning mBERT to the English part-of-speech tagging task.

In this paper, we consider a new direction to improve the cross-lingual performance, which is to preserve the cross-lingual ability of pre-trained multilingual models in the fine-tuning stage. Motivated by the continual learning (Ring, 1994; Rebuffi et al., 2017; Kirkpatrick et al., 2017; Lopez-Paz and Ranzato, 2017) that aims to learn a new task without forgetting the previous learned tasks, we adopt a continual learning framework to constrain the parameter learning in the pre-trained multilingual model when we fine-tune it to downstream

[1]This task is to find the correct translation sentence from the target corpus given a source language sentence.

Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021), pages 64–71
Bangkok, Thailand (Online), August 6, 2021. ©2021 Association for Computational Linguistics

tasks in the source language. Specifically, based on the results in Figure 1, we aim to maintain the cross-linguality of pre-trained multilingual models by utilizing MLM and XSR tasks to constrain the parameter learning in the fine-tuning stage.

Experiments show that our methods help pre-trained models better preserve the cross-lingual ability. Additionally, our methods surpass other fine-tuning baselines on the strong multilingual model mBERT and XLMR (Conneau et al., 2019) on zero-shot cross-lingual part-of-speech tagging (POS) and named entity recognition (NER) tasks.

2 Related Work

Cross-lingual methods, which alleviate the need for obtaining large amounts of annotated data in target languages, have been applied to multiple NLP tasks, such as task-oriented dialogue systems (Chen et al., 2018; Liu et al., 2019), part-of-speech tagging (Wisniewski et al., 2014; Zhang et al., 2016; Kim et al., 2017), named entity recognition (Mayhew et al., 2017; Ni et al., 2017; Xie et al., 2018; Liu et al., 2021), abstractive summarization (Duan et al., 2019; Zhu et al., 2019), and dependency parsing (Schuster et al., 2019; Ahmad et al., 2019). Recently, multilingual language models (Devlin et al., 2019; Conneau and Lample, 2019; Huang et al., 2019; Conneau et al., 2019), pre-trained on a large-scale data corpus across a great many languages, have significantly improved the cross-lingual performance. However, the corresponding fine-tuning techniques have been less studied. Wu and Dredze (2019) investigated the effectiveness of fine-tuning mBERT by freezing its partial bottom layers, and Muller et al. (2021) further analyzed the fine-tuning of mBERT.

3 Methodology

In this section, we first describe the gradient episodic memory (GEM) (Lopez-Paz and Ranzato, 2017), a continual learning framework, which we adopt to constrain the fine-tuning process. Then, we introduce how we fine-tune the pre-trained multilingual model with GEM.

3.1 Gradient Episodic Memory (GEM)

We consider a scenario where the model has already learned $n-1$ tasks and needs to learn the n-th task. The main feature of GEM is an episodic memory $\mathcal{M}_k$ that stores a subset of the observed examples from task k ($k \in [1, n]$). The loss at the memories from the k-th task can be defined as

$$\mathcal{L}(f_\theta, \mathcal{M}_k) = \frac{1}{|\mathcal{M}_k|} \sum_{(x_i,k,y_i)\in\mathcal{M}_k} \mathcal{L}(f_\theta(x_i, k), y_i), \quad (1)$$

where the model f_θ is parameterized by θ. In order to maintain the performance of the model in the previous $n-1$ tasks while learning the n-th task, GEM utilizes the losses for the previous $n-1$ tasks in Eq. (1) as inequality constraints, avoiding their increase but allowing their decrease. Concretely, when observing the training samples (x, y) from the n-th task, GEM solves the following problem:

$$\begin{aligned} &\text{minimize}_\theta \;\; \mathcal{L}(f_\theta(x, n), y) \\ &\text{subject to} \\ &\mathcal{L}(f_\theta, \mathcal{M}_k) \le \mathcal{L}(f_\theta^{n-1}, \mathcal{M}_k) \text{ for all } k < n, \end{aligned} \quad (2)$$

where f_θ^{n-1} is the model before learning task n.

3.2 Fine-tuning with GEM

We consider two tasks ($n = 2$) in total by applying GEM to the fine-tuning of pre-trained multilingual models, namely, mBERT and XLMR. The first task is either what the pre-trained models have already learned (MLM) or the ability that they already possess (XSR), and the second task is the fine-tuning task. We follow Eq. (2) when we fine-tune the pre-trained models:

$$\begin{aligned} &\text{minimize}_\theta \;\; \mathcal{L}(f_\theta(x, \mathcal{T}_2), y) \\ &\text{subject to} \;\; \mathcal{L}(f_\theta, \mathcal{T}_1) \le \mathcal{L}(f_\theta^*, \mathcal{T}_1), \end{aligned} \quad (3)$$

where $\mathcal{T}_1$ and $\mathcal{T}_2$ denote the first and second tasks, respectively, and f_θ^* represents the original pre-trained model. When the MLM task is considered as the first task, we constrain the fine-tuning process of the pre-trained model by preventing it from forgetting its original task after fine-tuning so as to better preserve the original cross-lingual ability. When the XSR task is considered as the first task, on the other hand, we prevent the pre-trained model from losing its cross-lingual ability after fine-tuning. We also consider incorporating both MLM and XSR as the first task.

4 Experiments

4.1 Dataset

For the POS task, we use Universal Dependencies 2.0 (Nivre et al., 2017) and select English (en), French (fr), Spanish (es), Greek (el) and Russian (ru) to evaluate our methods. For the NER task,

Model	MLM					XSR (Spanish to English)			XSR (Italian to English)		
	en	es	fr	el	ru	P@1	P@5	P@10	P@1	P@5	P@10
mBERT	10.68	3.51	8.63	2.08	2.70	56.26	68.80	73.92	44.76	61.32	66.70
Naive Fine-tune	216.80	16.72	40.54	5.62	8.61	37.72	52.20	58.43	26.12	37.46	46.69
w/ frozen layers	95.17	9.33	30.04	3.44	5.34	38.16	53.92	59.16	28.69	42.74	48.76
Multi-Task Learning											
MTF w/ MLM	9.50	5.10	8.62	2.56	3.47	35.93	50.41	56.20	24.79	37.18	45.46
MTF w/ XSR	121.50	100.10	96.50	773.00	180.80	75.40	80.88	85.76	75.94	85.44	88.29
MTF w/ Both	9.89	9.45	11.30	3.80	4.16	77.84	82.57	87.97	74.38	83.29	86.95
Continual Learning											
GEM w/ MLM	12.99	6.62	11.39	2.87	4.22	**42.90**	**57.26**	**63.58**	**31.66**	**44.16**	**50.16**
GEM w/ XSR	252.9	26.73	55.95	11.84	16.46	63.65	75.45	80.56	63.56	78.18	83.42
GEM w/ Both	12.16	6.40	10.62	3.40	4.30	64.34	76.23	81.42	64.12	79.35	84.59

Table 1: Experiments on MLM and XSR tasks based on mBERT. Models other than mBERT are fine-tuned to the English POS task. The underlined numbers in the MLM task denote that the performance is close to mBERT's. The bold numbers in the XSR task denote the best performance after fine-tuning without using the XSR supervision.

we use CoNLL 2002 (Tjong Kim Sang, 2002) and CoNLL 2003 (Sang and De Meulder, 2003), which contain English (en), German (de), Spanish (es) and Dutch (nl), to evaluate our methods. For both tasks, we consider English as the source language and other languages as target languages.

4.2 Baselines

We compare our methods to several baselines. **Naive Fine-tune** (Wu and Dredze, 2019) is to add one linear layer on top of the pre-trained model while fine-tuning with L2 regularization. **Fine-tune with Partial Layers Frozen** (Wu and Dredze, 2019) is to fine-tune pre-trained multilingual models by freezing the partial bottom layers. And **Multi-Task Fine-tune (MTF)** is to fine-tune pre-trained multilingual models on both the fine-tuning task and additional tasks (MLM and XSR).

4.3 Training Details

We conduct the MLM task with two settings. First, we only utilize the English Wikipedia corpus (**MLM (en)**) since we observe the catastrophic forgetting in the English MLM task as in Figure 1. Second, we utilize both the source and target languages Wikipedia corpus (**MLM (all)**). The first setting is used in our main experiments. Note that we do not use all pre-trained languages in mBERT for the MLM task because it would make the fine-tuning process very time-consuming. For the XSR task, we leverage the sentence pairs between the source and target languages from the Europarl parallel corpus (Koehn, 2005).[2]

[2]More training details are in the appendix.

5 Results & Analysis

Does GEM preserve the cross-lingual ability? From Table 1, we can see that naive fine-tuning mBERT significantly decreases the MLM performance, especially in English. Since mBERT is fine-tuned to the English task, the English subword embeddings are fine-tuned, which makes mBERT lose more MLM task information in English. Naive fine-tuning also makes the XSR performance of mBERT drop significantly. We observe that fine-tuning with partial layers frozen is able to somewhat prevent the MLM performance from getting worse, while fine-tuning with GEM based on that task almost preserves the original MLM performance of mBERT. Although we only use English data in the MLM task, using GEM based on the MLM task still preserves the task-related parameters that are useful for other languages. Correspondingly, we can see that *GEM w/ MLM* achieves better XSR performance than *Naive Fine-tune w/ frozen layers*, which shows that GEM helps better preserve the cross-lingual ability of mBERT.

In addition, although *GEM w/ XSR* aggravates the catastrophic forgetting in the MLM task, it is able to significantly improve the XSR performance due to the usage of the XSR supervision. Furthermore, incorporating both the MLM and XSR tasks can better preserve the performance in both tasks.

Does GEM improve the cross-lingual performance? From Table 2, we can see that our methods consistently surpass the fine-tuning baselines on all target languages in the POS and NER tasks. In terms of the average performance, our methods outperform the baselines by an around or more

Model	POS						NER				
	en	es	fr	el	ru	avg†	en	es	de	nl	avg†
Naive Fine-tune	96.23	82.95	89.12	84.21	85.45	85.43	**91.97**	74.96	69.56	77.57	74.03
w/ frozen layers	96.07	83.41	89.41	85.54	85.17	85.88	91.90	75.27	70.23	77.89	74.46
Multi-Task Learning											
MTF w/ MLM	94.47	83.01	88.08	84.48	80.46	84.01	91.82	71.47	67.90	74.91	71.43
MTF w/ XSR	96.39	82.41	87.05	72.51	86.09	82.01	91.85	74.02	68.55	75.67	72.75
MTF w/ Both	95.63	83.52	89.07	85.21	83.10	85.28	91.74	71.87	68.12	74.86	71.62
Continual Learning											
GEM w/ MLM	**97.39**	84.65	89.74	86.04	**86.93**	86.84‡	91.93	**76.45**	70.48	78.61	75.18‡
GEM w/ XSR	96.97	84.53	89.83	**86.53**	86.36	86.81‡	91.89	76.29	70.74	78.77	75.27‡
GEM w/ Both	97.04	**84.91**	**90.32**	86.44	86.13	**86.95**‡	91.45	76.20	**70.98**	**79.19**	**75.46**‡

Table 2: Zero-shot results on POS and NER tasks based on mBERT. †The average scores excluding en. ‡The results are statistically significant compared to all baselines with $p < 0.01$ by t-test.

Task	Models	en	es	fr	el	ru	avg
MLM	mBERT	10.7	**3.51**	8.63	**2.08**	**2.70**	5.52
	MTF w/ MLM (en)	9.50	5.10	8.62	2.56	3.47	5.85
	MTF w/ MLM (all)	**9.33**	4.19	**4.89**	2.34	3.04	**4.76**
	GEM w/ MLM (en)	13.0	6.62	11.4	2.87	4.22	7.62
	GEM w/ MLM (all)	11.8	4.18	6.83	2.29	2.99	5.62
POS	Naive Fine-tune	96.2	82.9	89.1	84.2	85.5	85.4
	MTF w/ MLM (en)	94.5	83.0	88.1	84.5	80.5	84.0
	MTF w/ MLM (all)	94.7	77.5	83.3	81.9	77.0	79.9
	GEM w/ MLM (en)	97.4	**84.7**	**89.7**	**86.0**	86.9	**86.8**
	GEM w/ MLM (all)	97.2	83.9	89.2	85.9	**87.1**	86.5

Table 3: Ablation study on the two settings of using the MLM task based on mBERT.

than 1% improvement.[3] In addition, constraining mBERT fine-tuning on the MLM task shows similar performance to constraining it on the XSR task. We conjecture that the effectiveness of both methods is similar, although they come from different angles. When the information of both tasks is utilized, GEM is able to slightly improve the performance. We find that the experimental results on XLMR are consistent with mBERT.

GEM vs. MTF From Table 1, we notice that using the MLM task, MTF achieves lower perplexity than GEM since it aggressively trains mBERT on this task. However, we observe that *MTF w/ MLM* makes the performance of the XSR, POS and NER tasks worse than *Naive Fine-tune*, and we speculate that MTF pushes mBERT to be overfit on the MLM task, instead of preserving its cross-lingual ability. Meanwhile, we can see that GEM regularizes the loss of the training on the MLM task to avoid catastrophic forgetting of previously trained languages, and conserve the cross-linguality of the pre-trained multilingual models.

In addition, we observe that adding XSR objective to the training cause the MLM performance worse. Although MTF achieves the best performance in the XSR task since it directly fine-tunes mBERT on that task, we can see from Table 2 that *GEM w/ XSR* boosts the cross-lingual performance of downstream tasks, while *MTF w/ XSR* has the opposite effect. We speculate that brutally fine-tuning mBERT on the XSR task (*MTF w/ XSR*) just makes mBERT learn the XSR task, while using GEM to constrain the fine-tuning on the XSR task can preserve its cross-lingual ability of mBERT. Incorporating both the MLM and XSR tasks further improves the performance for GEM, while MTF still performs worse than *Naive Fine-tune*.

Ablation Study From Table 3, we can see that using GEM to constrain fine-tuning on MLM with all languages (*GEM w/ MLM (all)*) achieves better performance than it does with only English (*GEM w/ MLM (en)*) on the MLM task since more MLM supervision signals are provided, while their performances in the POS task are similar. Intuitively, since *GEM w/ MLM* is able to improve the cross-lingual performance, constraining on more languages should give better performance. We conjecture, however, that the constraint with all languages could be too aggressive, so mBERT might tend to be overfit to the monolingual MLM task in all languages instead of preserving its original cross-lingual ability. In addition, we observe that fine-tuning mBERT on the MLM task (MTF) would get worse when more languages are utilized.

6 Conclusion

In this paper, we propose to preserve the cross-linguality of pre-trained language models in the fine-tuning stage. To do so, we adopt a continual

[3]The results of XLMR are included in the appendix.

learning framework, GEM, to constrain the parameter learning in pre-trained multilingual models based on the MLM and XSR tasks when we fine-tune them to downstream tasks. Experiments on the MLM and XSR tasks illustrate that our methods can better preserve the cross-lingual ability of pre-trained models. Furthermore, our methods achieve better performance than fine-tuning baselines for the strong multilingual models mBERT and XLMR on the zero-shot cross-lingual POS and NER tasks.

Acknowledgement

We want to say thanks to the anonymous reviewers for the insightful reviews and constructive feedback. This work is partially funded by ITF/319/16FP and MRP/055/18 of the Innovation Technology Commission, the Hong Kong SAR Government.

References

Wasi Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard Hovy, Kai-Wei Chang, and Nanyun Peng. 2019. On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing. 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 2440–2452.

Wenhu Chen, Jianshu Chen, Yu Su, Xin Wang, Dong Yu, Xifeng Yan, and William Yang Wang. 2018. Xl-nbt: A cross-lingual neural belief tracking framework. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 414–424.

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. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32*, pages 7059–7069. Curran Associates, Inc.

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.

Xiangyu Duan, Mingming Yin, Min Zhang, Boxing Chen, and Weihua Luo. 2019. Zero-shot cross-lingual abstractive sentence summarization through teaching generation and attention. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3162–3172.

Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2485–2494.

Joo-Kyung Kim, Young-Bum Kim, Ruhi Sarikaya, and Eric Fosler-Lussier. 2017. Cross-lingual transfer learning for pos tagging without cross-lingual resources. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2832–2838.

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.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. Citeseer.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, et al. 2020. Xglue: A new benchmark dataset for cross-lingual pretraining, understanding and generation. *arXiv preprint arXiv:2004.01401*.

Zihan Liu, Jamin Shin, Yan Xu, Genta Indra Winata, Peng Xu, Andrea Madotto, and Pascale Fung. 2019. Zero-shot cross-lingual dialogue systems with transferable latent variables. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1297–1303.

Zihan Liu, Genta I Winata, Samuel Cahyawijaya, Andrea Madotto, Zhaojiang Lin, and Pascale Fung. 2021. On the importance of word order information in cross-lingual sequence labeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13461–13469.

David Lopez-Paz and Marc'Aurelio Ranzato. 2017. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, pages 6467–6476.

Stephen Mayhew, Chen-Tse Tsai, and Dan Roth. 2017. Cheap translation for cross-lingual named entity recognition. In *Proceedings of the 2017 conference*

on empirical methods in natural language processing, pages 2536–2545.

Benjamin Muller, Yanai Elazar, Benoît Sagot, and Djamé Seddah. 2021. First align, then predict: Understanding the cross-lingual ability of multilingual bert. *arXiv preprint arXiv:2101.11109*.

Jian Ni, Georgiana Dinu, and Radu Florian. 2017. Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1470–1480.

Joakim Nivre, Željko Agić, Lars Ahrenberg, et al. 2017. Universal dependencies 2.0. lindat/clarin digital library at the institute of formal and applied linguistics, charles university, prague.

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

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. 2017. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010.

Mark Bishop Ring. 1994. *Continual learning in reinforcement environments*. Ph.D. thesis, University of Texas at Austin Austin, Texas 78712.

Erik 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*, pages 142–147.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. 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 1599–1613.

Erik F Tjong Kim Sang. 2002. Introduction to the conll-2002 shared task: language-independent named entity recognition. In *proceedings of the 6th conference on Natural language learning-Volume 20*, pages 1–4.

Guillaume Wisniewski, Nicolas Pécheux, Souhir Gahbiche-Braham, and François Yvon. 2014. Cross-lingual part-of-speech tagging through ambiguous learning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1779–1785.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844.

Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A Smith, and Jaime G Carbonell. 2018. Neural cross-lingual named entity recognition with minimal resources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 369–379.

Yuan Zhang, David Gaddy, Regina Barzilay, and Tommi Jaakkola. 2016. Ten pairs to tag–multilingual pos tagging via coarse mapping between embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1307–1317.

Junnan Zhu, Qian Wang, Yining Wang, Yu Zhou, Jiajun Zhang, Shaonan Wang, and Chengqing Zong. 2019. Ncls: Neural cross-lingual summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3045–3055.

A Training Details

We utilize the Wikipedia corpus for the MLM task. Given that using all the Wikipedia corpus will greatly lower the training speed, we randomly sample 1M sentences for each language for the training of *MTF w/ MLM* and *GEM w/ MLM*, and we use another 100K sentences for each language to evaluate the model performance on the MLM task. We take the English-Spanish (en-es), English-Italian (en-it), English-French (en-fr), English-Greek (en-el), English-German (en-de), and English-Dutch (en-nl) parallel datasets from the Europarl parallel corpus. We ramdomly select 90% of them for the training of *GEM w/ MLM* and *GEM W/ XSR*, and the rest 10% of them are used for evaluating the model performance on the XSR task. We use accuracy for evaluating the POS task, BIO-based F1-score for evaluating the NER task, perplexity for evaluating the MLM task, and P@k for evaluating the XSR task. Concretely, P@k (k=1,5,10) accounts for the fraction of pairs for which the correct translation of the source language sentence is in the k-th nearest neighbors. We use an early stop strategy which is based on the average performance over the target languages to select the model. We use the Adam optimizer with a learning of 1e-5. We use batch size 16 for the all tasks, namely, POS, NER, MLM and XSR. In each iteration, we use GEM to constrain the fine-tuning on a batch of data samples from the MLM and XSR tasks. Our models are trained on V100. The number of parameters for the mBERT-based model is around 178.6 million and for the XLMR-based model is around 278.9 million.

# samples	en	es	de	nl
Train	14,040	8,319	12,152	15,802
Validation	3,249	1,914	2,867	2,895
Test	3,452	1,516	3,005	5,194

Table 4: Number of samples for each language in the CoNLL 2002 and CoNLL 2003 NER datasets.

# samples	en	es	fr	el	ru
Train	12,543	14,187	14,450	1,662	3,850
Validation	2,002	1,400	1,476	403	579
Test	2,007	426	416	456	601

Table 5: Number of samples for each language in the Universal Dependencies 2.0 dataset for the POS task.

B Data Statistics

The data statistics of the NER and POS datasets are shown in Table 4 and Table 5, respectively.

C Results

C.1 XLMR Experiments

Experiments on POS and NER tasks for $\text{XLMR}_{\text{base}}$ are illustrated in Table 6 (in the next page). The results on XLMR are consistent with mBERT.

C.2 XSR Experiments

Experiments on more language pairs are illustrated in Table 7 (in the next page). The results on French to English, Greek to English, German to English and Dutch to English are consistent with the XSR results shown in the main paper (i.e., Spanish to English and Italian to English).

Model	POS						NER				
	en	es	fr	el	ru	avg†	en	es	de	nl	avg†
Naive Fine-tune	96.55	84.61	90.37	87.23	89.32	87.88	91.95	75.86	69.59	77.83	74.42
w/ frozen layers	96.40	84.63	90.33	86.27	89.44	87.67	91.53	76.12	68.79	78.26	74.39
Multi-Task Learning											
MTF w/ MLM	96.43	82.37	89.70	83.90	86.73	85.68	91.90	74.55	67.70	78.13	73.46
MTF w/ XSR	96.93	84.94	89.08	86.93	89.27	87.55	91.93	75.35	70.58	77.65	74.53
MTF w/ Both	96.31	83.55	89.90	87.01	84.94	86.35	91.67	75.45	67.80	77.91	73.72
Continual Learning											
GEM w/ MLM	96.87	**85.90**	90.57	87.25	89.43	88.29	91.93	76.43	70.98	78.77	75.39
GEM w/ XSR	96.86	85.01	89.87	88.14	89.90	88.23	91.94	**76.61**	**71.19**	79.28	**75.69**
GEM w/ Both	96.10	85.63	**90.99**	**89.02**	**91.36**	**89.25**	91.91	76.48	70.53	**79.86**	75.62

Table 6: Zero-shot results on POS and NER tasks based on XLMR. †The average scores excluding en.

Model	XSR (French to English)			XSR (Greek to English)			XSR (German to English)			XSR (Dutch to English)		
	P@1	P@5	P@10	P@1	P@5	P@10	P@1	P@5	P@10	P@1	P@5	P@10
mBERT	53.92	65.44	72.12	35.68	59.40	65.31	52.10	64.71	69.43	54.56	66.69	72.54
Naive Fine-tune	34.12	50.03	57.90	15.12	33.35	42.69	33.68	49.23	56.45	34.79	51.13	58.01
w/ frozen layers	35.50	52.23	59.87	16.98	35.63	44.74	34.20	50.97	58.11	35.29	53.24	59.77
Multi-Task Learning												
MTF w/ MLM	32.49	48.67	56.23	14.67	32.29	40.64	32.37	47.45	55.48	32.86	50.35	56.55
MTF w/ XSR	74.20	78.65	83.69	73.94	77.59	83.47	75.48	80.67	85.44	75.83	85.28	88.35
MTF w/ Both	75.30	79.34	84.86	74.25	78.39	84.63	77.93	82.67	87.86	74.42	83.57	86.68
Continual Learning												
GEM w/ MLM	**39.79**	**55.62**	**63.34**	**21.33**	**39.60**	**47.36**	**37.70**	**53.44**	**60.53**	**38.35**	**54.89**	**63.06**
GEM w/ XSR	63.11	67.81	71.92	61.79	65.37	70.43	63.14	75.52	80.85	63.90	78.33	83.46
GEM w/ Both	63.84	68.50	72.05	61.54	64.38	69.50	64.41	76.39	81.70	64.36	79.65	84.72

Table 7: Experiments on XSR tasks based on mBERT. Models other than mBERT are fine-tuned to the English POS task. The bold numbers in the XSR task denote the best performance after fine-tuning without using the XSR supervision.