

Overcoming Catastrophic Forgetting in Massively Multilingual Continual Learning

Genta Indra Winata^{1*}, Lingjue Xie^{1*}, Karthik Radhakrishnan^{1*}, Shijie Wu^{1*},
Xisen Jin^{2†}, Pengxiang Cheng¹, Mayank Kulkarni^{3†}, Daniel Preotiuc-Pietro¹

¹Bloomberg ²University of Southern California ³Amazon Alexa AI
{gwinata, lxie91, kradhakris10, swu671, pcheng134, dpreotiucpie}@bloomberg.net
maykul@amazon.com, xisenjin@usc.edu

Abstract

Real-life multilingual systems should be able to efficiently incorporate new languages as data distributions fed to the system evolve and shift over time. To do this, systems need to handle the issue of catastrophic forgetting, where the model performance drops for languages or tasks seen further in its past. In this paper, we study catastrophic forgetting, as well as methods to minimize this, in a massively multilingual continual learning framework involving up to 51 languages and covering both classification and sequence labeling tasks. We present LR ADJUST, a learning rate scheduling method that is simple, yet effective in preserving new information without strongly overwriting past knowledge. Furthermore, we show that this method is effective across multiple continual learning approaches. Finally, we provide further insights into the dynamics of catastrophic forgetting in this massively multilingual setup.

1 Introduction

Standard supervised NLP methods perform well when training on enough data from a uniform distribution. However, they fail to retain knowledge learnt in the past when sudden shifts occur in training data distributions. This effect of dropping performance on data from past distributions is commonly referred to as *catastrophic forgetting* (McCloskey and Cohen, 1989; de Masson D’Autume et al., 2019; Biesialska et al., 2020a), where stability or preservation of knowledge is traded off for increased plasticity or the ability to acquire new knowledge. To tackle this issue, continual learning (CL) methods were proposed under various settings, such as limited compute or ability to store past data (Lopez-Paz and Ranzato, 2017; de Masson D’Autume et al., 2019). The data shifts commonly studied are obtained by training over a sequence of non-iid partitions (Chaudhry et al.,

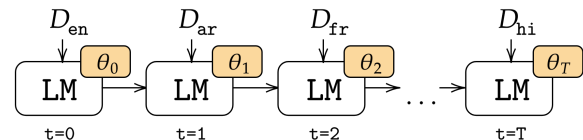


Figure 1: Multilingual Continual Learning: The model θ is trained sequentially with data from T different languages.

2018), different tasks (Jin et al., 2021), or by training on various domains such as in task-oriented dialogue (Madotto et al., 2021), named entity recognition (Monaikul et al., 2021), part-of-speech (Liu et al., 2020), and intent detection (Wu et al., 2021).

Lifelong learning is key to the success of deployed multilingual systems, enabling the system to incorporate annotated data for new languages as they become available without costly retraining and redeployment of the entire system. This sequential availability of data for new languages is a common case of training data shift (see Figure 1 for the task setup). Yet, the effect of catastrophic forgetting was not yet systematically studied for multi-lingual models with multiple diverse languages. M’hamdi et al. (2022) study continual learning in a cross-lingual setting limited to just six languages. The cross-lingual abilities of pre-trained models were found to drop when performing fine-tuning for a target language (Liu et al., 2021), although applying continual learning approaches can effectively reduce the magnitude of the effect (Lopez-Paz and Ranzato, 2017).

In this paper, we systematically study the effect of catastrophic forgetting and mitigation strategies in a massively multilingual setting covering up to 51 languages on three different tasks. We start by quantifying the extent to which forgetting happens when languages are presented to the model in sequence, identifying an up to 16% F1 drop compared to training using all the data mixed. Next, we propose LR ADJUST, a simple, yet effective, method

* These authors contributed equally. † The work was done while at Bloomberg.

to preserve the learned knowledge by adjusting the learning rate over time to alleviate the knowledge overwriting from the new language and preserve the previous learned knowledge. This method is orthogonal to continual learning methods and thus can be handily combined with any of these. We find that across three different CL methods, LR ADJUST helps further reduce the gap between a fully trained model and the CL setup. We conduct analysis on the aspect of cross-lingual transfer in backward and forward directions to measure the influence of the CL on previous tasks and its ability in zero-shot learning respectively. Finally, we conduct analyses on the effects of catastrophic forgetting when first training on multiple languages jointly and when using a curriculum learning approach informed by language similarity.

2 Massively Multilingual Continual Learning

2.1 Task Setup

We define a curriculum of T tasks as an ordered set of data sets $\mathcal{D} = \{D_1, D_2, \dots, D_t, \dots, D_T\}$ and model θ_t , where D_t is the data set with task t . In this case, the task is a distinct language. The weights of model θ are updated continuously $\theta_{t+1} \leftarrow f(\theta_t, D_t)$ by minimizing the log-likelihood over data set D_t via gradient updates.

2.2 Inter-task Learning Rate Adjustment

We propose LR ADJUST, a simple and effective method to adjust the learning rate when we start training on a new task. Our intuition is that models are susceptible to catastrophic forgetting when we provide a higher learning rate, so the learning rate should be toned down with time to ensure the preservation of the learned knowledge and to reduce the effect of overwriting the weights with the new knowledge. Learning rate adjustments have been studied in the context of incremental learning (Cavalin et al., 2009; Khreich et al., 2012) and for efficient optimization using schedules (Ge et al., 2019). Concretely, the new learning rate is lowered every time as the following: $lr_t = \max(lr_{min}, lr_{t-1} * \gamma)$, with a weight γ , where $\gamma < 1$ and lr_{min} is the minimum learning rate. The method is detailed in Algorithm 1.

2.3 Continual Learning Method

We experiment with the following continual learning approaches:

Algorithm 1 Inter-task Learning Rate Adjustment (LR ADJUST)

Require: An ordered list of tasks $\mathcal{D} = \{D_1, D_2, \dots, D_t, \dots, D_T\}$

Require: γ : learning rate adjustment coefficient, lr_t : learning rate at time t , θ : model weights, lr_{min} : minimum learning rate

- 1: Randomly initialize the classifier on θ
 - 2: **for all** $D_t \in \mathcal{D}$ **do**
 - 3: Adjust learning rate to $lr_t = \max(lr_{min}, lr_{t-1} * \gamma)$
 - 4: Compute $\nabla_{\theta} \mathcal{L}_{D_t}(f_{\theta})$ using D_t
 - 5: $\theta_{t+1} \leftarrow \theta_t - lr_t \nabla_{\theta} \mathcal{L}_{D_t}(f_{\theta})$
 - 6: **end for**
-

- **Experienced Replay** (de Masson D’Autume et al., 2019) uses an episodic memory to store seen training data in memory and retrieve it from memory for fine-tuning. We schedule the replay step to be run every few iterations. During the replay step, we retrieve the data and fine-tune the model using the retrieved data. The number of stored data is constrained to ensure efficient memory use. And, we take a uniform distribution of samples across all labels.
- **Averaged GEM (A-GEM)** (Chaudhry et al., 2018) also utilizes an episodic memory \mathcal{M} and is a more efficient implementation of GEM (Lopez-Paz and Ranzato, 2017) that computes the gradient constraints and minimizes the loss as follows:

$$\mathcal{L}_t(\theta_t, \mathcal{M}) \leq \mathcal{L}_{t-1}(\theta_{t-1}, \mathcal{M}),$$

where loss \mathcal{L}_t is constrained to be lower or equal to the loss \mathcal{L}_{t-1} .

- **Elastic Weight Consolidation (EWC)** (Kirkpatrick et al., 2017) minimizes the following loss:

$$\mathcal{L}_t(\theta) = \mathcal{L}_t(\theta) + \sum_i \frac{\lambda}{2} F_i (\theta_t - \theta_{t-1}^*)^2,$$

where F_i is the Fisher information matrix, λ is a coefficient that sets how important the old task is compared to the new one, and θ_{t-1}^* is the previous learned weights. The F_i is pre-computed after each task is completed and we incorporate the loss to the training on each gradient update.

3 Experimental Setup

3.1 Data sets

We use a multilingual natural language understanding data set, MASSIVE (FitzGerald et al., 2022)

and a multilingual named entity recognition (NER) data set, WikiAnn (Rahimi et al., 2019). The MASSIVE data set consists of two tasks: intent classification and slot filling with 51 languages. The WikiAnn data set consists of 40 languages. We adopt the data splits from the original papers.

3.2 Methods

Our model architecture is an encoder-only multilingual model XLM-R_{BASE} (Conneau et al., 2020) with a classification layer for each task. All parameters in the model are updated during the training. The full hyper-parameters are listed in Appendix A. The language order for the experiments are listed in Appendix B. We experiment with the following approaches:

- **MULTI**: A single multilingual model is trained on data mixed from all languages. This represents an upper bound to CL methods, as there are no memory or data sequencing constraints.
- **MONO**: A separate model is trained on all the supervised data for each language and applied to all inputs in that language.
- **VANILLA**: A single model is trained by sequentially presenting data from each language. The language order is selected randomly.
- **CL Methods**: We run REPLAY, A-GEM, and EWC to train a single model on data from each language presented sequentially.
- **CL Methods + LR ADJUST**: We run the CL methods with the learning rate adjustment method described in Section 2.2.

3.3 Metrics

We measure the ability of cross-lingual transfer using CL metrics adapted from Lopez-Paz and Ranzato (2017). We define a matrix $R \in \mathbb{R}^{T \times T}$, where $R_{i,j}$ denotes the test score performance of the model on task t_j when training the last sample from task t_i . We formally define the metrics as:

3.3.1 Cross-lingual Forward Transfer (CFT)

This metric represents the ability to perform zero-shot learning by evaluating on the test data from tasks/languages that are unseen in training. We formally define the metric as:

$$CFT = \frac{1}{T-1} \sum_{i=1}^{T-1} \bar{X}_i,$$

$$\bar{X}_i = \frac{1}{T-i} \sum_{j=i+1}^T R_{i,j},$$

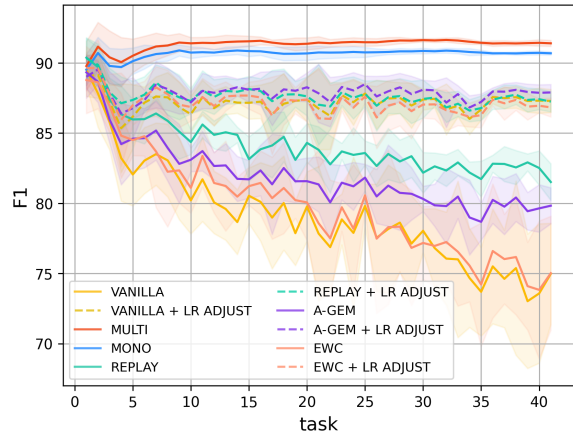


Figure 2: Average F1 scores and standard deviation over 5 runs on WikiAnn (Rahimi et al., 2019) evaluated over increasing number of languages seen in training.

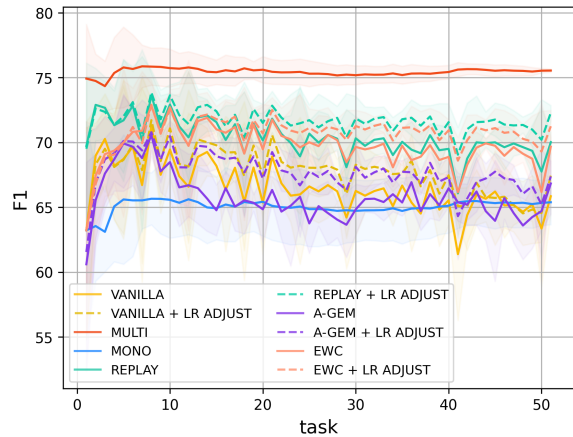


Figure 3: Average F1 scores and standard deviation over 5 runs on MASSIVE-Slot (FitzGerald et al., 2022) evaluated over increasing number of languages seen in training.

where \bar{X}_i is the average performance of the languages that will be seen in the future ($t_{>i}$).

3.3.2 Cross-lingual Backward Transfer (CBT)

This metric measures the influence of learning a task t_i on the performance of the previous tasks. We formally define the metric as the following:

$$CBT = \frac{1}{T-1} \sum_{i=1}^{T-1} R_{T-1,i} - R_{i,i}.$$

CBT practically measures the effect of catastrophic forgetting of past tasks after adding a new task to the model.

4 Results and Analysis

Figures 2, 3 and 4 show performance numbers across the languages seen to that point in train-

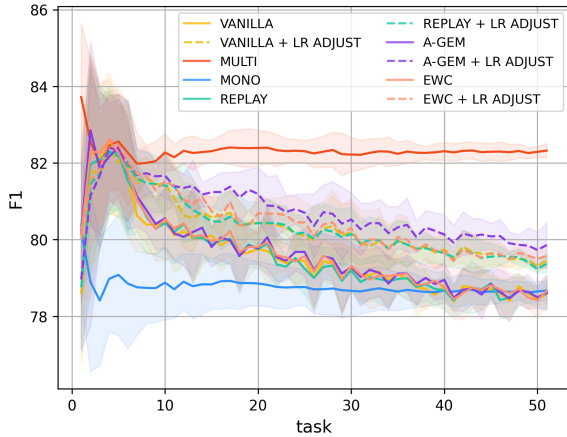


Figure 4: Average F1 scores and standard deviation over 5 runs on MASSIVE-Intent (FitzGerald et al., 2022) evaluated over increasing number of languages seen in training.

ing for the three data sets. Each point on the graph shows the average and standard deviation of the F1 scores, obtained over 5 runs with different seeds. For the CL experiments, the seed also controls the order of languages in training. This can lead to higher variance across runs – compared to mixing the same data across runs – because the forgetting effect depends on the language order. Table 1 shows the forward and backward transfer for the three data sets.

Results show the following:

Catastrophic forgetting is present in multilingual continual learning, with performance dropping quickly when training on languages sequentially (VANILLA). This method converges to between 4 – 17 F1 lower than mixing the same data and training a full model every time (MULTI). We also see that this effect is present even when the performance of monolingual models is close to that of multilingual models, as in the case of WikiAnn. Training a full model on all languages (MULTI) always performs best, outperforming training on one language at a time (MONO) sometimes substantially (MASSIVE-Slot - 10 F1, MASSIVE-Intent - 4 F1), highlighting the importance of cross-lingual transfer and preserving information from past seen languages in a continual learning setup.

Continual learning methods generally help dampen the impact of forgetting. For example, in WikiAnn, the REPLAY and A-GEM CL methods reduce the backward transfer from 16.90 to 11.87 and 10.75 respectively, albeit EWC does not substantially improve the performance relative to the VANILLA

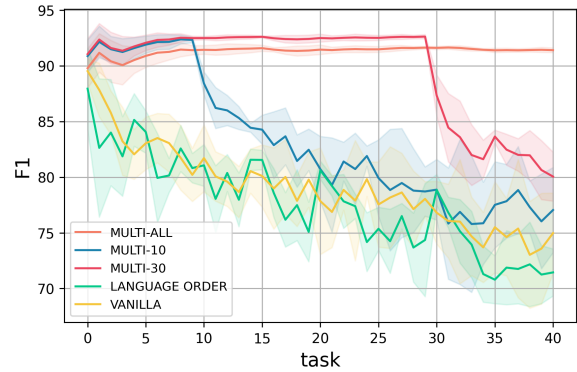


Figure 5: Multi-task strategies and training using the language order heuristics on WikiAnn. It shows the averaged F1 scores and standard deviation on 5 runs.

method.

LR ADJUST, the learning rate adjustment scheme, **further reduces the gap** to the multi-task model significantly and consistently across languages when combined with any of the CL methods. For example, on the WikiAnn dataset, the backward transfer is reduced from 16.90 to just 3.59 and 3.79 for the A-GEM and REPLAY methods respectively, making multilingual CL feasible. Further, we see that using CL methods alone results in a continuous drop in performance as more languages are added, while adding LR ADJUST stabilizes average performance after the first few languages, resulting in a flatter curve.

Finally, we see that the **patterns of improvement hold when studying cross-lingual forward transfer**, which quantifies the zero-shot model performance on languages unseen in training. The continual learning approaches improve over sequential training (e.g. +4.45 on WikiAnn) and using LR ADJUST in addition further boosts performance (e.g. +9.25 for VANILLA, +5.83 for REPLAY on WikiAnn). This shows that the resulting models were able to retain essential and generalizable information for the task that is more universal across all languages.

4.1 Multi-task Training vs. Catastrophic Forgetting

We conduct additional experiments to understand whether initially training on multiple languages at once can reduce the severity of catastrophic forgetting in the CL setup when new languages are added. We run a multi-task training on the first k languages, where k is 10 or 30, and then, we run the remaining languages sequentially on the WikiAnn data set. As shown in Figure 5, the model is more

	WikiAnn		MASSIVE-Slot		MASSIVE-Intent	
	CFT	CBT	CFT	CBT	CFT	CBT
VANILLA	67.77	-16.90	57.27	-7.92	79.60	-3.15
+ LR ADJUST	77.02	-4.19	59.17	-5.20	80.32	-1.56
EWC	68.58	-16.83	60.79	-5.36	79.67	-2.99
+ LR ADJUST	76.78	-4.50	62.30	-3.36	80.46	-1.53
A-GEM	69.54	-11.87	57.63	-6.22	79.68	-3.13
+ LR ADJUST	77.26	-3.59	57.93	-2.48	80.73	-1.65
REPLAY	72.22	-10.75	60.38	-6.04	79.23	-3.18
+ LR ADJUST	78.05	-3.79	62.54	-3.26	80.32	-1.65

Table 1: Cross-lingual Forward Transfer (CFT) and Cross-lingual Backward Transfer (CBT). Higher scores are better.

robust to forgetting when it is exposed to multi-task training with more languages with higher final average scores at the final task, but the graphs shows the performance still drops dramatically after being exposed to the first new language fed sequentially.

4.2 The Role of Language Order

To investigate the role of the language order on CL, we decide to reorder the language list by using heuristics. We start with two languages from the same family, as listed in Ethnologue (Eberhard and Gary, 2019), and add all languages from the same family one by one, then switch to a new language family and continue the same process. We conjecture that seeing a similar language at an interval will allow a more effective cross-lingual transfer. Figure 5 (LANGUAGE ORDER) displays the results, which indicate that performance does not improve after we manually select the languages, and its performance is similar to random ordering (VANILLA).

5 Related Work

CL aims to learn effectively over iterations by leveraging information from previously learned tasks (McCloskey and Cohen, 1989). CL has been applied towards mitigating catastrophic forgetting in many computer vision tasks (Lopez-Paz and Ranzato, 2017). Broadly speaking, CL methods can be classified (Biesialska et al., 2020b) into rehearsal (re-using training examples from prior task) (Rolnick et al., 2019; de Masson D’Autume et al., 2019), regularization (guide consolidation via additional loss) (Kirkpatrick et al., 2017), memory (Lopez-Paz and Ranzato, 2017; Chaudhry et al., 2018), and architectural methods (using task-specific parameters) (Rusu et al., 2016; Madotto et al., 2021).

6 Conclusion

We present the first study of catastrophic forgetting in a massively multilingual setting involving up to 51 languages on named entity recognition and natural language understanding tasks. We investigate continual learning methods and present a learning rate scheduling method that is simple yet effective in reducing the effects of catastrophic forgetting. Furthermore, we show that this method is effective across multiple continual learning methods. Finally, we provide analysis and further insights into the dynamics of catastrophic forgetting.

Acknowledgments

We are grateful to Abhinav Singh and Shuyi Wang for feedback on a draft of this manuscript.

Limitations

The experiment in this paper is limited to three common CL methods: Replay, A-GEM, and EWC. Due to compute resources, we experiment with XLM-R_{BASE} and do not compare with other variants or larger base models. We acknowledge that the MASSIVE data set varies in coverage across language groups and evaluation could over-represent certain linguistic families (Pikuliak and Simko, 2022).

Ethics Statement

In our experiments, we use publicly available data sets with permissive licenses for research experiments. We do not release new data or annotations as part of this work. There are no potential risks.

References

- Magdalena Biesialska, Katarzyna Biesialska, and Marta R Costa-jussà. 2020a. Continual lifelong learning in natural language processing: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6523–6541.
- Magdalena Biesialska, Katarzyna Biesialska, and Marta R. Costa-jussà. 2020b. [Continual lifelong learning in natural language processing: A survey](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6523–6541, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Paulo R Cavalin, Robert Sabourin, Ching Y Suen, and Alceu S Britto Jr. 2009. Evaluation of incremental learning algorithms for hmm in the recognition of alphanumeric characters. *Pattern Recognition*, 42(12):3241–3253.
- Arslan Chaudhry, Marc’Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. 2018. Efficient lifelong learning with a-gem. *arXiv preprint arXiv:1812.00420*.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard 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.
- Eberhard David, Gary Simons, and Charles Fennig. 2019. *Ethnologue: Languages of the world. Twenty-second Edition, SIL International, Dallas*.
- Cyprien de Masson D’Autume, Sebastian Ruder, Lingpeng Kong, and Dani Yogatama. 2019. Episodic memory in lifelong language learning. *Advances in Neural Information Processing Systems*, 32.
- David M Eberhard and F Gary. 2019. Simons, and charles d. fennig (eds.). 2019. *Ethnologue: Languages of the world*, 22.
- Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, et al. 2022. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages. *arXiv preprint arXiv:2204.08582*.
- Rong Ge, Sham M Kakade, Rahul Kidambi, and Pra-neeth Netrapalli. 2019. The step decay schedule: A near optimal, geometrically decaying learning rate procedure for least squares. *Advances in neural information processing systems*, 32.
- Xisen Jin, Bill Yuchen Lin, Mohammad Rostami, and Xiang Ren. 2021. [Learn continually, generalize rapidly: Lifelong knowledge accumulation for few-shot learning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 714–729, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Wael Khreich, Eric Granger, Ali Miri, and Robert Sabourin. 2012. A survey of techniques for incremental learning of hmm parameters. *Information Sciences*, 197:105–130.
- 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.
- Zihan Liu, Genta Indra Winata, Andrea Madotto, and Pascale Fung. 2020. Exploring fine-tuning techniques for pre-trained cross-lingual models via continual learning. *arXiv preprint arXiv:2004.14218*.
- Zihan Liu, Genta Indra Winata, Andrea Madotto, and Pascale Fung. 2021. Preserving cross-linguality of pre-trained models via continual learning. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 64–71.
- David Lopez-Paz and Marc’Aurelio Ranzato. 2017. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30.
- Andrea Madotto, Zhaojiang Lin, Zhenpeng Zhou, Seungwhan Moon, Paul A Crook, Bing Liu, Zhou Yu, Eunjoon Cho, Pascale Fung, and Zhiguang Wang. 2021. Continual learning in task-oriented dialogue systems. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7452–7467.
- Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of Learning and Motivation*, volume 24, pages 109–165. Elsevier.
- Meryem M’hamdi, Xiang Ren, and Jonathan May. 2022. Cross-lingual lifelong learning. *arXiv preprint arXiv:2205.11152*.
- Natawut Monaikul, Giuseppe Castellucci, Simone Filice, and Oleg Rokhlenko. 2021. Continual learning for named entity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13570–13577.
- Matúš Pikuliak and Marian Simko. 2022. [Average is not enough: Caveats of multilingual evaluation](#). In *Proceedings of the The 2nd Workshop on Multilingual Representation Learning (MRL)*, pages 125–133, Abu Dhabi, United Arab Emirates (Hybrid). Association for 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.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. 2019. Experience replay for continual learning. *Advances in Neural Information Processing Systems*, 32.

Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. 2016. Progressive neural networks. *arXiv preprint arXiv:1606.04671*.

Tongtong Wu, Massimo Caccia, Zhuang Li, Yuan-Fang Li, Guilin Qi, and Gholamreza Haffari. 2021. Pre-trained language model in continual learning: A comparative study. In *International Conference on Learning Representations*.

A Hyper-parameters

In all experiments, we run with five different seeds {42, 52, 62, 72, 82} using a V100 32GB GPU and each run takes up to a week to finish.

A.1 WikiAnn

Table 2 shows the hyper-parameters used in the experiments using WikiAnn dataset.

	MULTI	VANILLA	MONO	MULTI-10/30	REPLAY	A-GEM	EWC
LR	1e-4	5e-5	5e-5 (all except hi)	5e-5	-	-	-
LR Decay	0.9997	0.999	1e-5 (hi)	0.9997	-	-	-
Batch Size	32	32	32	32	-	-	-
Max epochs	20	20	20	20	-	-	-
Early stopping	5	5	5	5	-	-	-
EWC_Reg	-	-	-	-	-	-	10000
store_memory_prob	-	-	-	-	0.0005	0.01	-
max_store_num_samples	-	-	-	-	1E+7	1E+5	-
retrieve_num_samples	-	-	-	-	100	-	-
run_per_step	-	-	-	-	5000	2000	-

Table 2: Hyper-parameters for WikiAnn data set.

A.2 MASSIVE-Slot

Table 3 shows the hyper-parameters used in the experiments using MASSIVE-Slot dataset.

	MULTI	VANILLA	MONO	REPLAY	A-GEM	EWC
LR	2e-5	1e-5	1e-5	-	-	-
LR Decay	0.9999	0.9999	0.9999	-	-	-
Batch Size	32	8	8	-	-	-
Max epochs	20	20	20	20	-	-
Early stopping	5	5	5	5	-	-
EWC_Reg	-	-	-	-	-	10000
store_memory_prob	-	-	-	0.0005	0.01	-
max_store_num_samples	-	-	-	1E+7	1E+5	-
retrieve_num_samples	-	-	-	100	-	-
run_per_step	-	-	-	5000	2000	-

Table 3: Hyper-parameters for MASSIVE-Slot data set.

A.3 MASSIVE-Intent

Additionally, we run the CL setup on MASSIVE-Intent dataset and the results are shown in Figure 4. Table 4 shows the hyper-parameters used in the experiments using MASSIVE-Intent dataset.

	MULTI	VANILLA	MONO	REPLAY	A-GEM	EWC
LR	2e-5	2e-5	2e-5	-	-	-
LR Decay	0.999	0.995	0.995	-	-	-
Batch Size	32	32	32	-	-	-
Max epochs	20	20	20	-	-	-
Early stopping	5	5	5	-	-	-
EWC_Reg	-	-	-	-	-	10000
store_memory_prob	-	-	-	0.0005	0.01	-
max_store_num_samples	-	-	-	1E+7	1E+5	-
retrieve_num_samples	-	-	-	100	-	-
run_per_step	-	-	-	5000	2000	-

Table 4: Hyper-parameters for MASSIVE-Intent data set.

B Language Order

We randomly shuffle the language order for each seed. Tables 5 and 6 show the language order we use in the experiments for WikiANN and MASSIVE datasets, respectively.

seed	languages
42	el, bn, en, ta, ms, mk, ro, es, bs, sk, it, pl, lv, hr, et, sq, sv, nl, fa, lt, id, ru, tl, pt, hu, he, uk, sl, bg, af, tr, no, ca, cs, de, fi, fr, hi, ar, da, vi
52	el, sl, hr, he, fa, it, lt, tl, mk, cs, pl, hu, bs, tr, uk, fr, ta, pt, sq, da, ms, no, et, vi, ar, af, id, fi, es, ca, ru, sv, en, de, bg, nl, lv, ro, sk, bn, hi
62	et, fi, ar, sv, ms, fa, sq, tr, it, ru, no, el, da, pl, hi, bg, cs, nl, hr, sl, mk, he, lv, tl, vi, bn, ro, id, de, af, ca, uk, sk, en, lt, hu, pt, fr, bs, es, ta
72	fr, uk, mk, hr, ar, sl, sk, ta, bn, hi, ca, ro, pt, cs, fa, nl, en, he, pl, el, bg, sv, no, ru, da, ms, tl, af, id, vi, et, fi, it, de, hu, lv, sq, lt, es, tr, bs
82	de, fi, ar, pl, pt, da, ms, hu, et, lv, ca, lt, af, fa, sq, mk, id, it, ta, sl, tr, ro, uk, bs, hi, vi, cs, bn, nl, tl, fr, no, bg, sv, he, en, es, hr, sk, ru, el

Table 5: Language Order for Experiments with WikiAnn.

C Geographical Information of Languages

Table 7 shows all languages’ language families and subgroups on NusaX and MASSIVE datasets.

seed	languages
42	kn-IN, ka-GE, is-IS, fa-IR, bn-BD, tl-PH, ko-KR, en-US, mn-MN, hu-HU, my-MM, ja-JP, fi-FI, az-AZ, sq-AL, sl-SL, es-ES, km-KH, pt-PT, af-ZA, te-IN, id-ID, nl-NL, zh-CN, sw-KE, ms-MY, ml-IN, it-IT, jv-ID, ta-IN, tr-TR, ro-RO, nb-NO, th-TH, fr-FR, zh-TW, vi-VN, ar-SA, lv-LV, ru-RU, cy-GB, pl-PL, da-DK, el-GR, he-IL, hi-IN, hy-AM, ur-PK, am-ET, de-DE, sv-SE
52	[nl-NL, is-IS, bn-BD, id-ID, en-US, my-MM, kn-IN, he-IL, ja-JP, da-DK, sq-AL, hu-HU, tl-PH, lv-LV, sw-KE, zh-TW, mn-MN, fi-FI, am-ET, zh-CN, fr-FR, sl-SL, sv-SE, ta-IN, it-IT, vi-VN, hi-IN, ur-PK, cy-GB, pt-PT, de-DE, ro-RO, ru-RU, km-KH, pl-PL, te-IN, af-ZA, ml-IN, jv-ID, fa-IR, th-TH, es-ES, el-GR, ar-SA, ko-KR, ka-GE, ms-MY, nb-NO, tr-TR, az-AZ, hy-AM
62	sv-SE, az-AZ, ko-KR, ja-JP, el-GR, ru-RU, my-MM, ka-GE, ur-PK, vi-VN, tl-PH, pt-PT, fr-FR, kn-IN, tr-TR, en-US, fi-FI, sl-SL, he-IL, hy-AM, ml-IN, ar-SA, sw-KE, da-DK, te-IN, cy-GB, it-IT, id-ID, zh-TW, lv-LV, km-KH, pl-PL, nl-NL, ms-MY, am-ET, de-DE, sq-AL, hu-HU, af-ZA, th-TH, zh-CN, nb-NO, es-ES, jv-ID, ta-IN, is-IS, mn-MN, hi-IN, bn-BD, fa-IR, ro-RO
72	fi-FI, tl-PH, tr-TR, da-DK, zh-TW, hi-IN, sw-KE, ko-KR, ms-MY, lv-LV, cy-GB, az-AZ, ml-IN, kn-IN, sv-SE, hy-AM, de-DE, id-ID, vi-VN, it-IT, te-IN, fr-FR, my-MM, ta-IN, es-ES, hu-HU, nb-NO, pt-PT, ro-RO, ar-SA, nl-NL, af-ZA, mn-MN, ru-RU, am-ET, en-US, km-KH, he-IL, ja-JP, el-GR, zh-CN, is-IS, ka-GE, sq-AL, pl-PL, th-TH, jv-ID, fa-IR, ur-PK, sl-SL, bn-BD
82	az-AZ, he-IL, am-ET, fr-FR, ta-IN, ka-GE, ja-JP, hy-AM, bn-BD, ml-IN, ro-RO, pl-PL, jv-ID, pt-PT, nl-NL, tr-TR, mn-MN, zh-TW, ko-KR, ur-PK, af-ZA, cy-GB, sq-AL, vi-VN, hi-IN, km-KH, tl-PH, kn-IN, sw-KE, it-IT, sv-SE, sl-SL, de-DE, el-GR, is-IS, fi-FI, da-DK, ru-RU, ms-MY, lv-LV, ar-SA, th-TH, hu-HU, te-IN, es-ES, fa-IR, id-ID, nb-NO, my-MM, zh-CN, en-US

Table 6: Language Order for Experiments with MAS-SIVE.

Language Code	Name	Family	Subgroup
af / af-ZA	Afrikaans	Indo-European	Germanic
am-ET	Amharic	Afro-Asiatic	Semitic
ar / ar-SA	Arabic	Afro-Asiatic	Semitic
az-AZ	Azerbaijani	Turkic	Southern
bn / bn-BD	Bengali	Indo-European	Indo-Iranian
bg	Bulgarian	Indo-European	Balto-Slavic
bs	Bosnian	Indo-European	Balto-Slavic
ca	Catalan	Indo-European	Italic
cs	Czech	Indo-European	Balto-Slavic
cy-GB	Welsh	Indo-European	Celtic
da / da-DK	Danish	Indo-European	Germanic
de / de-DE	German	Indo-European	Germanic
en / en-US	English	Indo-European	Germanic
el / el-GR	Greek	Indo-European	Greek
es / es-ES	Spanish	Indo-European	Italic
et	Estonian	Uralic	Finnic
fa	Persian	Indo-European	Indo-Iranian
fi / fi-FI	Finnish	Uralic	Finnic
fr / fr-FR	French	Indo-European	Italic
id / id-ID	Indonesian	Austronesian	Malayo-Polynesian
is-IS	Icelandic	Indo-European	Germanic
it / it-IT	Italian	Indo-European	Italic
ja-JP	Japanese	Japonic	Japanese
jv-ID	Javanese	Austronesian	Malayo-Polynesian
he / he-IL	Hebrew	Afro-Asiatic	Semitic
hi / hi-IN	Hindi	Indo-European	Indo-Iranian
hr	Croatian	Indo-European	Balto-Slavic
hu / hu-HU	Hungarian	Uralic	Hungarian
hy-AM	Armenian	Indo-European	Armenian
ka-GE	Georgian	Kartvelian	Georgian
km-KH	Khmer	Austro-Asiatic	Mon-Khmer
kn-IN	Kannada	Dravidian	Southern
ko-KR	Korean	Koreanic	Korean
lt	Lithuanian	Indo-European	Balto-Slavic
lv / lv-LV	Latvian	Indo-European	Balto-Slavic
ml-IN	Malayalam	Dravidian	Southern
mn-MN	Mongolian	Mongolic	Eastern
ms / ms-MY	Malay	Austronesian	Malayo-Polynesian
mk	Macedonian	Indo-European	Balto-Slavic
my-MM	Burmese	Sino-Tibetan	Tibeto-Burman
nb-NO	Norwegian	Indo-European	Germanic
nl / nl-NL	Dutch	Indo-European	Germanic
no	Norwegian	Indo-European	Germanic
pl / pl-PL	Polish	Indo-European	Balto-Slavic
pt / pt-PT	Portuguese	Indo-European	Italic
ro / ro-RO	Romanian	Indo-European	Italic
ru / ru-RU	Russian	Indo-European	Balto-Slavic
sl / sl-SL	Slovenian	Indo-European	Balto-Slavic
sk	Slovak	Indo-European	Balto-Slavic
sq / sq-AL	Albanian	Indo-European	Albanian
sw-KE	Swahili	Niger-Congo	Atlantic-Congo
sv / sv-SE	Swedish	Indo-European	Germanic
ta / ta-IN	Tamil	Dravidian	Southern
te-IN	Telugu	Dravidian	South-Central
th-TH	Thai	Kra-Dai	Kam-Tai
tl / tl-PH	Tagalog	Austronesian	Malayo-Polynesian
tr / tr-TR	Turkish	Turkic	Southern
ur-PK	Urdu	Indo-European	Indo-Iranian
uk	Ukrainian	Indo-European	Balto-Slavic
vi / vi-VN	Vietnamese	Austro-Asiatic	Mon-Khmer
zh-CN / zh-TW	Chinese	Sino-Tibetan	Chinese

Table 7: Geographical information of languages under study. The language family is based on Ethnologue (David et al., 2019).

ACL 2023 Responsible NLP Checklist

A For every submission:

- A1. Did you describe the limitations of your work?
Page 5. Section Limitations
- A2. Did you discuss any potential risks of your work?
Not applicable. No potential risks
- A3. Do the abstract and introduction summarize the paper’s main claims?
Yes, on the first page (abstract and section 1 introduction)
- A4. Have you used AI writing assistants when working on this paper?
Left blank.

B Did you use or create scientific artifacts?

Left blank.

- B1. Did you cite the creators of artifacts you used?
No response.
- B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.
- 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)?
No response.
- 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?
No response.
- B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.
- 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.
No response.

C Did you run computational experiments?

Left blank.

- C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
We put the computing infrastructure and GPU hours on Appendix A.

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?

Section 3 Experimental Setup and Appendix A

- 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?

Figure 2 and Figure 3, we provide the average and std

- 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.)?

Not applicable. Left blank.

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.