CLLE: A Benchmark for Continual Language Learning Evaluation in Multilingual Machine Translation

Han Zhang[1,2], **Sheng Zhang**[3], **Yang Xiang**[2], **Bin Liang**[1,4], **Jinsong Su**[5], **Zhongjian Miao**[5], **Hui Wang**[2,*], **and Ruifeng Xu**[1,2,4*]
[1] Harbin Institute of Technology, Shenzhen, China
[2] Peng Cheng Laboratory, Shenzhen, China
[3] National University of Defense Technology, Changsha, China
[4] Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies
[5] Xiamen University, Xiamen, China
hanlardresearch@gmail.com, zhangsheng@nudt.edu.cn
{xiangy, wangh06}@pcl.ac.cn, bin.liang@stu.hit.edu.cn
{jssu, miaozhongjian}@stu.xmu.edu.cn, xuruifeng@hit.edu.cn

Abstract

Continual **L**anguage **L**earning (CLL) in multilingual translation is inevitable when new languages are required to be translated. Due to the lack of unified and generalized benchmarks, the evaluation of existing methods is greatly influenced by experimental design which usually has a big gap from the industrial demands. In this work, we propose the first **C**ontinual **L**anguage **L**earning **E**valuation benchmark CLLE in multilingual translation. CLLE consists of a **C**hinese-ce**n**tric corpus — CN-25 and two CLL tasks — the close-distance language continual learning task and the language family continual learning task designed for real and disparate demands. Different from existing translation benchmarks, CLLE considers several restrictions for CLL, including domain distribution alignment, content overlap, language diversity, and the balance of corpus. Furthermore, we propose a novel framework COMETA based on **C**onstrained **O**ptimization and **META**-learning to alleviate catastrophic forgetting and dependency on historical training data by using a meta-model to retain the important parameters for old languages. Our experiments prove that CLLE is a challenging CLL benchmark and that our proposed method is effective when compared with other strong baselines. *Due to the construction of corpus, the task designing and the evaluation method are independent of the central language, we also construct and release the English-centric corpus EN-25 to facilitate academic research*[1].

1 Introduction

Training a multilingual Neural Machine Translation (NMT) model jointly in all the directions requires collecting in advance the parallel corpus of all the languages, which is less practical due to the continuously occurrence of the translation requirement of the new languages. Adding new languages to a well-trained multilingual NMT model is a resource-saving method compared with training from scratch. However, directly finetuning on new languages will result in catastrophic forgetting of historical languages. **C**ontinual **L**anguage **L**earning (CLL) methods (Berard, 2021; Garcia et al., 2021; Lyu et al., 2020; Escolano et al., 2019, 2020, 2021), focus on gradually extending the language capacity of multilingual NMT models without forgetting old languages which is the major challenge of CLL tasks.

The existing multilingual NMT evaluation benchmarks (Akhbardeh et al., 2021; Qi et al., 2018; Schwenk et al., 2021; Zhang et al., 2020) focus more on multi-task NMT or continual domain learning (Thompson et al., 2019) but put little emphasis on CLL restrictions. Hence, most of the existing CLL methods (Berard, 2021; Garcia et al., 2021) are evaluated on the traditional multilingual NMT evaluation benchmarks and conducted on a simple experiment (e.g. training a multilingual NMT model and adding a specific new language) which has a big gap from the realistic CLL. In the industrial demands (Lyu et al., 2020), there are usually more new languages and families are required to be continually learned in more continual learning stages. Due to the lack of CLL benchmark, there are no rigorous evaluations of the CLL methods for the number of languages, language family distribution, and learning order. The existed methods' (Berard, 2021; Garcia et al., 2021; Lyu et al., 2020; Escolano et al., 2019, 2020, 2021) experiment settings such as the selection of the new and

* R. Xu and H. Wang are corresponding authors.
[1]The entire corpus is released at https://github.com/HITSZ-HLT/CLLE

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 428–443
December 7-11, 2022

old languages, the availability of historical training data, and the growth of model parameters are not unified. However, the catastrophic forgetting observed is sensitive to the experimental design more than any inherent modeling limitations (Hussain et al., 2021). Hence, it is urgent and necessary to design a benchmark to unify the configurations of the CLL tasks.

In this work, we propose the first CLL benchmark — CLLE for CLL in the multilingual NMT scenario. CLLE consists of a **C**hinese-ce**n**tric and domain distribution-consistent multilingual parallel corpus — CN-25, which is collected by extracting and refining the CC-Matrix corpus (Schwenk et al., 2021). Specifically, CN-25 includes 25 languages aligned with Chinese, 23 of which have more than 650k sentence pairs. The corpus refinement is processed with the text-based filter rules and the LaBSE (Feng et al., 2022) multilingual model. The content domain distribution of each language is adjusted to be similar by adjusting the number of samples for each topic clustered by K-means.

We design the close-distance language continual learning (CLCL) task and the language family continual learning (LFCL) task associated with the CLLE benchmark to verify the CLL method on disparate experiment settings. To be specific, in the CLCL task, the new languages are from the learned languages families, and the addition of new languages is needed at only one learning stage. In the more challenging LFCL task, the new languages are from the unseen languages families, and more learning stages are introduced.

For the CLL tasks, we propose the COMETA framework based on the **C**onstrained **O**ptimization (Thompson et al., 2019; Aljundi et al., 2018) and the **META**-learning (de Masson d'Autume et al., 2019; Wang et al., 2020; Liang et al., 2022). We train a CNN-based meta-model to predict the performance of the NMT model in old languages according to its parameters. Then we use the meta-model to calculate the importance weights to retain the language-specific embeddings (Qi et al., 2018; Mathur et al., 2019; Liang et al., 2021) of old languages. Compared with standard constrained optimization methods such as EWC (Thompson et al., 2019) and MAS (Aljundi et al., 2018), COMETA retains the knowledge of the old languages without accessing the historical training data. And the importance weights can be dynamically updated, which is more flexible for the CLL process.

The main contributions of this work include:

- We introduce the first CLL benchmark CLLE which includes the CN-25 corpus and two CLL tasks, and the construction method of the CN-25 corpus can be used for any central language.
- We design two CLL tasks to verify the CLL method on disparate experiment settings, and the tasks are derived from the requirements in real scenarios.
- We propose the COMETA method based on constrained optimization and meta-learning, which outperforms existing constrained optimization methods without using the old training data.

2 Related work

Benchmarks in Multilingual Neural Machine Translation In this section, we focus on multilingual NMT benchmarks including Chinese. WMT series (Bojar et al., 2016, 2017; Neves et al., 2018; Barrault et al., 2019; Specia et al., 2020; Akhbardeh et al., 2021) corpus includes the commonly used of high-quality languages and most of them are aligned with English. The corpus updates each year, and the multilingual low-resource translation corpus is now added such as the Indo-European (Akhbardeh et al., 2021) translation. WAT series (Nakazawa et al., 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021) provide a multilingual multi-domain parallel corpus between Asian languages and English, and the patent task includes a Chinese-Japanese parallel corpus. FLORES-101 (Goyal et al., 2022) includes the multilingual parallel corpus of 101 languages, in which 3001 sentences were extracted from English Wikipedia and translated into 101 languages by professional translators. The corpus covers a variety of different topics and domains, but it only has the test part. OPUS-100 (Zhang et al., 2020) is an English-centric dataset sampled from the OPUS collection (Tiedemann, 2012) and covers a large variety of topics. CWMT series (Yang et al., 2019; CCM, 2021) corpus provides the corpus of China's ethnic minorities' languages such as Mongolian, Tibetan, and Uyghur aligned with simplified Chinese. However, the language number is not enough to support Chinese-centric CLL research. TED talks corpus (Qi et al.,

2018) has 3540 (60×59) language pairs that rely on volunteers to provide translations for public texts. The content domain of TED corpus includes but not limited to **T**echnology, **E**ntertainment, and **D**esign. Although many language corpora are aligned with Chinese, the content's overlap of different languages corpus can not meet the Chinese-centric multilingual WMT models' training requirements, which are experimentally proved in Section 6.2.

The above benchmarks are designed for multi-task multilingual NMT training or evaluation. Due to limitations on language and content distributions, they can not meet the CLL requirements. In Section3.4, we compare CN-25 with the above benchmarks in detail.

Continual language learning Due to the introduction of Transformer to multilingual NMT, adding new languages to translation models has been increasingly studied in the past few years. Thompson et al. (2019) use the diag Fisher matrix as importance weights to retain the important parameters for old tasks. Although it is a classical constrained optimization method designed for continual domain learning, the idea is still suitable for the CLL scenario. Lyu et al. (2020) introduces a modularized method by adding a new language-specific encoder/decoder modules to the multilingual NMT model, which can satisfy industrial requirements. Similar architecture-based approaches (Escolano et al., 2019, 2020, 2021) have proved valid in the CLL scenario, while the drawback is that the model size will grow concomitantly as the task count increases. Garcia et al. (2021) propose a "vocabulary substitution" approach to augment untranslatable languages of the multilingual NMT model. The core thought is reusing the overlapped embedding parameters between new and old multilingual vocabulary. And the author declares that the success of the "vocabulary substitution" approach is due to the large size vocabulary of origin multilingual NMT model. Berard (2021) proposes to add language-specific adapter modules and freeze the major structure when learning a new language, and experimentally proves no degradation of the existing language pairs. However, learning a new long-distance language may be limited by the frozen major structure. Because these models are executed in different continual learning settings and evaluated by inconsistent methods, the performance comparison is difficult to perform.

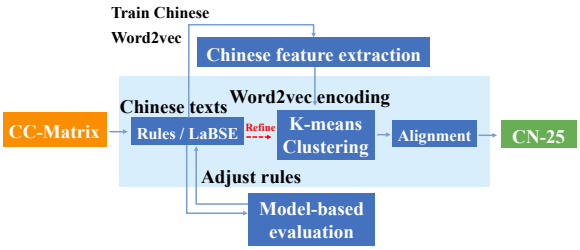

Figure 1: The whole process of Chinese-centric parallel corpus processing.

3 The CN-25 corpus

In this section, we introduce the domain-aligned corpus CN-25 and compare it with existing multilingual corpora. The processing pipeline of CN-25 is shown in Figure 1. In the first step, we utilize the multilingual encoding model LaBSE and text-based filter rules to refine the CC-matrix corpus. Then we use a model-based method to evaluate the refined corpus and the refinement process is regulated according to the evaluation results. To align the topic distribution of different languages, we utilize the central language as the agent and cluster the whole corpus into 100 topics through K-means clustering. In each topic, the sentence number is regulated according to the median value and LaBSE score.

3.1 CC-Matirx data source

CC-Matrix (Schwenk et al., 2021) is a parallel corpus containing a wide range of languages, which is obtained from a large number of web snapshots through parallel corpus mining technologies. Through exploiting the highly optimized FAISS (Johnson et al., 2021) vector retrieval library and language-agnostic BiLSTM (Artetxe and Schwenk, 2019) to encode and retrieve monolingual data, the sentence pairs with a higher probability of translation relationship are preliminarily found. The quality of the sentence pairs are further judged by LASER [2] margin score with a threshold around 1.06.

3.2 Corpus refinement

LASER supports the encoding of more than 100 languages, but the language-agnostic BiLSTM is not trained for translation ranking (Guo et al., 2018). LaBSE (Feng et al., 2022) utilizes the combination of pre-training and dual transformer encoder finetuning to boost the performance on the

[2]https://github.com/facebookresearch/LASER

Figure 2: Correlation matrix of K-means topic distribution. Dark colors indicate low relevancy. Left: original correlation matrix. Right: adjusted correlation matrix.

translation ranking task. We utilize it to re-score the CC-Matrix dataset and refine the parallel corpus with the threshold 0.8 which is selected manually by inspecting the refined samples. Through manual analysis, we find that parts of the corpus are adulterated with other languages. We discard the sentences containing too many foreign-language texts according to the proportion of other language characters.

3.3 Corpus domain alignment

The CLL task focuses more on processing language expansion rather than domain shift which is the main challenge of continual domain learning. Hence, the domain distribution alignment is an important restriction of the CLL benchmark to distinguish it from the benchmark of continual domain learning. We use the central language as the agent to get the domain distribution of the refined corpus. A word2vec (Mikolov et al., 2013) model is trained on the monolingual sentences of the central language to encode the sentences (averaging the word embeddings in a sentence). After that, the entire corpus is clustered into 100 topics (according to the agent language) through the encoding and the K-means clustering method. For each language, we rank the sentences in each topic according to the LaBSE score and only retain the sentences whose scores are above the median value for all languages. The correlation matrix of topic distribution before and after adjustment is shown in Figure 2. The valid and test set are sampled from the top 100 and the top 100-150 sentences in each topic.

3.4 Compared with existing multilingual corpus

We compare CN-25 with existing multilingual benchmarks from multiple aspects including language diversity, corpus amount, domain alignment, and content overlap across languages. Table 1 shows the comparison of CN-25 with WMT-2022, WAT-2022, CWMT-2022, TED talks, CC-matrix, OPUS-100, and FLORES-101 benchmarks.

Language diversity and the count of sentences. CN-25 includes 25 languages from 17 language families. In each family, the typical and commonly used languages are selected. In addition, the number of sentences in each language is controlled in balance. Most languages (except for Tamil and Swahili) have more than 650k sentences aligned with Chinese. For the multilingual parallel corpus including Chinese, as shown in Table 1, the corpus size of CN-25 is greater than the widely used corpora in CLL researches (Berard, 2021; Garcia et al., 2021) such as TED and OPUS-100. And the CN-25 corpus is the largest corpus satisfying the domain alignment restriction.

Domain distribution alignment and content overlap. The domain difference has a critical impact on continual learning, which has been studied in the continual domain learning scenario (Thompson et al., 2019). In the CLL scenario, the domain distribution should be aligned to eliminate the influence of domain differences. The content overlap across languages refers to the central language repetition across different language sentence

Benchmarks	Domain alignment	Content non-overlap	Language diversity	Sentence count (avg)
CCMT-2022	X	✓	5	2.9m
WAT-2022*	X	✓	9	-
WMT-2022*	X	✓	10	-
CC-Matrix	X	✓	38	17.6m
TED†	✓	X	60	253k (en)
OPUS-100‡	X	X	100	546k (en)
FLORES-101	✓	X	101	3k
CN-25	✓	✓	26	730k

Table 1: Compared with existed datasets from four aspects. Sentence count means the average count of all translation directions. *We select the corpus of the general machine translation task in WMT-2022 and the corpus of the document-level translation task in WAT-2022. †TED has average of 90k sentence pairs of Chinese to/from other languages, and the average sentence pairs of English to/from 20 commonly used languages is about 253k. ‡OPUS-100 consists of 55M English-centric sentence pairs covering 100 languages, and has 1000k English-Chinese pairs .

pairs. When trained on the content-overlapped corpus, the central language sentences are repeatedly learned in every translation direction, which increases the overfitting risk of the central language. None of the existing benchmarks considers both constraints simultaneously. For instance, the TED and FLORES-101 corpus translate a sentence from one language to multiple languages. Although the domain distribution of different languages is coincident, the content overlap problem appears. We experimentally analyze it in Section 6.2.

4 The continual language learning tasks

We then introduce two challenging tasks for CLL in our benchmark with different language families and CLL stage settings. Both the language families and the number of learning stages have an impact on the catastrophic forgetting, as the former is empirically analyzed in Appendix A and the latter is shown by Hsu et al. (2018). CLL models are evaluated on the newly and historically learned languages by the averaged BLEU score after each learning stage.

CLCL task: close-distance language continual learning. In this scenario, the whole CLL task consists of two stages: training a multilingual NMT model at the first stage and then continuing learning new languages at the second stage. The learning sequence in this scenario is relatively short and no new language families are added to the subsequent stage. The languages learned at the second stage belong to the language families of the previous stage. For instance, participants are required to train a multilingual NMT model on three language families including Germanic (Dutch, German, Swedish), Romance (Portuguese, Spanish), and Slavic (Russian, Czech) at the first stage, and then continuously train on three languages including English (Germanic), French (Romance), and Polish (Slavic).

LFCL Task: language family continual learning. In this more challenging scenario, we set more learning stages and introduce new language families to add difficulty to CLL approaches. For example, the Germanic family (Dutch, German, English, and Swedish), the Romance family (French Portuguese, and Spanish), and the Slavic family (Polish, Russian and Czech) are learned in sequence. In the long sequence of learning, the catastrophic forgetting problem could be serious. Existing CCL models which try to retain knowledge from previous stages to alleviate catastrophic forgetting may suffer more time and memory consumption. Take EWC, a classical continual learning approach, for example, as the learning stage increases, the time of traversing old training sets and computing will significantly increase when calculating the fisher matrix for the parameters.

Performance metric. We use the average BLEU from and to Chinese (the centrical language) to evaluate the performance of methods on CLL tasks, which is the commonly used metrics in CLL task (Berard, 2021). Let L_i represent the set of historical learned languages up to stage-i, $j \in L_i$ represent the learned language, $\overrightarrow{b_{i,j}}$ and $\overleftarrow{b_{i,j}}$ be the test BLEUs of zh->j and j->zh after the model has finished learning stage-i, the average BLEUs from and to Chinese are respectively defined as

$$B_i = \frac{1}{2|L_i|} \sum_{j \in L_i} \overleftarrow{b_{i,j}} + \overrightarrow{b_{i,j}} \quad (1)$$

The average BLEU B_{-1} of the last stage represents the performance on the CLL task.

5 COMETA framework for continual language learning tasks

We propose the COMETA framework which is based on constrained optimization and meta-learning. The core idea is to utilize a new network to evaluate the importance of the NMT model's parameters for old languages, where we call the new network as meta-model. The meta-model is trained to learn the training process of the NMT

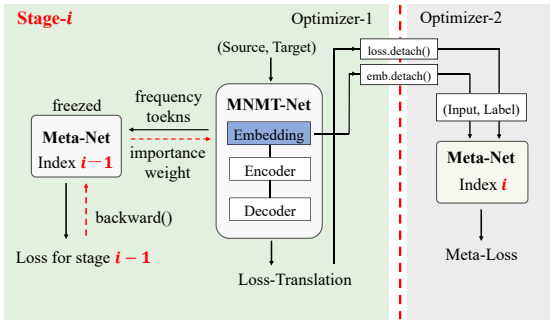

Figure 3: Framework of COMETA at stage-i.

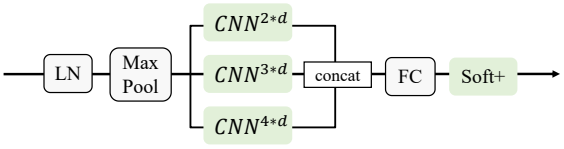

Figure 4: The structure of meta-model.

model according to the NMT model's parameters and loss. After that, the meta-model calculates importance weights to constrain the change to the embeddings of old languages when learning new languages. The framework of COMETA at stage-i is shown in Figure 3. At stage-i, we employ two independent computation graphs to train the multilingual NMT model and meta-model respectively.

5.1 Meta model and importance weight

As shown in Figure 4, the meta-model, a CNN-based network, predicts the translation loss according to the language-specific embeddings which correspond to the language-specific frequent tokens. The LayerNorm and max-pooling operators are used to rescale and resize the embedding, and then multiple CNN kernels of different sizes are utilized to extract the features. The features are concatenated and fed to the fully connected layer. To predict the loss, we use the SoftPlus (Dugas et al., 2000) activation function which only returns non-negative values. The input of the meta-model is the language-specific embedding, and the objective is to fit the translation loss.

We use the meta-model to predict a loss value, and compute the gradient of the language-specific embedding respect to the predicted loss. Then we use the gradient (absolute value) as the importance weights to constrain the change to embeddings when learning new languages. We only penalize the change to embeddings because the embeddings are more language-specific than the parameters of encoder/decoder layers, and the bias of the embedding (Cao et al., 2021) is observed when learning new languages.

5.2 Training progress

As shown in Figure 3, at each stage-i we train the multilingual NMT model $f_\theta(\cdot)$ and train meta-model $g^i_\phi(\cdot)$ apart in two calculation graphs. The embedding and the translation loss (θ^i_{emb}, $L^i_{translate}$) in the left calculation graph are used to train the meta-model in the right calculation graph.

Meta model training. The embeddings θ^i_{emb} are fed to the meta-model $g^i_\phi(\cdot)$ to predict a translation loss value, and the meta loss (loss of training meta-model) is computed by the mean square error loss function $M_{SE}(\cdot)$:

$$L^i_{meta} = M_{SE}(g^i_\phi(\theta^i_{emb}), L^i_{translate}) \quad (2)$$

Then the parameter ϕ is updated by Adam (Kingma and Ba, 2015) optimizer:

$$\phi \leftarrow Adam^1(\partial L^i_{meta}/\partial\phi, \phi)$$

Multilingual NMT model training. At the learning stage-i, the total loss L^i of NMT model combines the translation loss $L^i_{translate}$ and the knowledge retain loss L^i_{retain}, namely

$$L^i = L^i_{translate} + \gamma \cdot L^i_{retain} \quad (3)$$

Given a batch of source sentences src^i, corresponding target sentences tgt^i and cross entropy loss function $C_E(\cdot)$, the translation loss of Multilingual NMT model $f_\theta(\cdot)$ is

$$L^i_{translate} = C_E(f_\theta(src^i), tgt^i) \quad (4)$$

The knowledge retain loss is

$$L^i_{retain} = W^i \cdot |\theta^i_{emb} - \theta^{i-1}_{emb}|^2 \quad (5)$$

where the importance weights W_i is computed by the meta-model $g^{i-1}_\phi(\cdot)$ trained at stage-(i-1)

$$W^i = \alpha \cdot |\partial g^{i-1}_\phi(\theta^i_{emb})/\partial\theta^i_{emb}|^2 + (1-\alpha)\cdot W^{i-1} \quad (6)$$

The parameters θ is updated by another Adam optimizer:

$$\theta \leftarrow Adam^2(\partial L^i_{oss}/\partial\theta, \theta)$$

5.3 Training tricks

When continually learning new languages, the parameters of encoder and decoder layers are frozen at early steps, and only the embeddings of the NMT model are updated, which makes the new language embedding adapt to the well-trained encoder and decoder layers. Otherwise, the parameters of the encoder and decoder layers will be updated sharply, which leads to quick forgetting.For the language-specific embeddings, we design a dynamic frozen strategy to gradually unfreeze the embeddings that corresponded tokens are not commonly used. As observed in Appendix A Figure 5-6, at the early stage of learning a new language, the gradient of the embedding is large and the update of the optimizer is drastic. Hence, freezing part of word embeddings in the early steps can avoid updating the parameters drastically and erasing the learned knowledge stored in the embedding.

6 Experiments

In this section, we evaluate the CN-25 corpus and the COMETA method. All the experiments are evaluated under SacreBLEU (Post, 2018) metric.

6.1 Corpus analysis

We list the language information and parallel sentence amount in Table 2. To get the specific domain distribution, we train a FastText model to classify the Chinese sentences aligned with each language into several categories. Same with the K-means clustering results (Figure 2 right part), the Tamil corpus domain distribution also has a large difference from other languages. Because the original Tamil corpus in CC-Matrix is too scarce, it is hard to adjust the count of samples in each cluster to subject to the similar distribution. The categories distribution by FastText model is shown in Appendix B Table 9.

6.2 Corpus Qualitative evaluation

We assess the corpus quality of the CN-25 through manual and model-based methods. For intuitive comparison, we present top 30 Chinese-English sentence pairs with the highest LASER score of CC-Matrix and CN-25.

To quantitatively verify the refined data quality, we use the sentence pairs in CN-25 and the sentence pairs that are filtered out (regarded as low-quality) to finetune the M2M-418M (Fan et al., 2021) model respectively, then evaluate the model under WMT2020 and TED benchmarks. We select the seven commonly used languages (200k sentences per language) to finetune the M2M-418m model for 90k step (128 sentences per batch) , as shown in Table 3, in each translation direction, the corpus in CN-25 brings more performance improvement. And the corpus filtered out even damages the original M2M model performance in several directions. It proves that the refining process can filter out low-quality parallel sentence pairs. Analogously, we evaluate the model under the WMT-2020 benchmark (only English corpus aligned with Chinese) through finetuning and training from scratch, as shown in Table 4, the corpus with a higher LaBSE score still keeps the advantage.

To verify the performance of models trained on CN-25, we train two Chinese-centric multilingual NMT models respectively on TED and CN-25. Both models are trained on 10 languages for 450k steps (128 sentences per batch) and evaluated on the test dataset of TED. As shown in Table 5, we find two models have competitive performance in "Chinese->xx" directions, which proves that the CN-25 corpus has high data quality. While the model trained on TED severely overfits Chinese due to the content overlap problem.

6.3 Baselines evaluation

We reproduce EWC (Thompson et al., 2019) and MAS (Aljundi et al., 2018) baselines and compare them with COMETA under six experiment settings (2 model sizes * 3 replay settings). Table 6 and 7 present the average BLEU at each learning stage on CLCL task and LFCL task. Compared with EWC and MAS, COMETA does not use the source and target sentences of the historical training corpus, while the average performance of COMETA still has an advantage, which proves the meta-model can identify the important parameters for old languages. However, in each replay setting, the catastrophic forgetting is remarkable, especially in the zero-replay scenario. It proves that the CLCL and LFCL tasks are challenging and the CLL methods still have a large room to improve.

7 Conclusion

We propose the first CLL benchmark — CLLE with the CN-25 corpus and two CLL tasks — CLCL and LFCL. Compared with existing multilingual benchmarks, CLLE considers several restrictions

ISO	Language	Family	Script	Bitext Number (k)					
				Train	Valid	Test	Aligned	Refined	Original
nl	Dutch	Germanic	Latin	831.0	4.8	9.7	845.5	4.23E+03	8.20E+03
de	German	Germanic	Latin	773.7	4.7	9.6	788.0	9.55E+03	1.86E+04
en	English	Germanic	Latin	643.6	4.4	9.2	657.2	4.51E+04	7.14E+04
sv	Swedish	Germanic	Latin	826.2	4.9	9.7	840.8	3.43E+03	7.45E+03
fr	French	Romance	Latin	787.6	4.7	9.5	801.8	1.28E+04	2.14E+04
pt	Portuguese	Romance	Latin	795.5	4.7	9.5	809.7	7.15E+03	1.22E+04
es	Spanish	Romance	Latin	741.1	4.6	9.5	755.2	1.51E+04	2.41E+04
ja	Japanese	Japonic	Kanji;Kana	700.7	4.9	9.7	715.3	7.18E+03	1.24E+04
ko	Korean	Koreanic	Hangul	668.3	4.8	9.7	682.8	3.00E+03	5.10E+03
vi	Vietnamese	Vietic	Latin	780.7	4.8	9.8	795.2	4.31E+03	8.05E+03
ar	Arabic	Arabic	Arabic	743.9	4.8	9.7	758.4	4.10E+03	6.58E+03
fa	Farsi	Iranian	Arabic	706.7	4.9	9.8	721.4	1.67E+03	4.92E+03
he	Hebrew	Semitic	Hebrew	639.1	4.9	9.8	653.7	1.78E+03	5.24E+03
fi	Finnish	Uralic	Latin	838.5	4.8	9.7	853.1	2.35E+03	4.61E+03
hu	Hungarian	Uralic	Latin	834.8	4.8	9.7	849.3	2.46E+03	4.79E+03
lt	Lithuanian	Baltic	Latin	835.5	4.8	9.8	850.2	1.54E+03	3.34E+03
pl	Polish	Slavic	Latin	858.6	4.8	9.7	873.1	3.42E+03	7.45E+03
ru	Russian	Slavic	Cyrillic	768.8	4.7	9.7	783.3	6.96E+03	1.31E+04
cs	Czech	Slavic	Latin	808.0	4.8	9.7	822.5	3.40E+03	6.56E+03
hi	Hindi	Indo-Atyan	Devanagari	693.1	4.8	9.7	707.7	8.00E+02	2.27E+03
ta	Tamil	Dravidian	Tamil	107.1	4.0	9.1	120.2	1.20E+02	1.10E+03
id	Indonesian	Malayo-Polyn	Latin	811.2	4.7	9.7	825.6	3.15E+03	6.24E+03
sw	Swahili	Niger-Congo	Latin	199.3	4.7	9.6	213.6	2.10E+02	1.07E+03
tr	Turkish	Turkic	Latin	784.8	4.9	9.8	799.5	3.32E+03	7.22E+03
el	Greek	Hellenic	Greek	807.3	4.8	9.7	821.8	2.45E+03	4.97E+03

Table 2: The statistics of 25 languages aligned with Chinese. Groups are divided according to M2M-100. "Original": the amount of CC-matrix corpus. "Refined": the amount of corpus after rules filtering, and LaBSE refining. "Aligned": amount of corpus of domain alignment.

for CLL, including domain distribution alignment, content overlap, language diversity, and the balance of corpus. For the CLL tasks, we introduce a novel method COMETA based on constrained optimization and meta-learning to retain the important parameters for old languages through the meta-model. The experiments prove that CN-25 is a high-quality corpus, that the CLL tasks are challenging and that our proposed method outperforms other strong baselines.

Limitations

We discuss the limitations of the CN-25 corpus and the COMETA method. For the CN-25 corpus, the data quality of the refined corpus relies on the LaBSE model, which prefers better on the high-resource languages. Hence, we can't guarantee that each language of CN-25 has the same high-quality corpus. Furthermore, the topic alignment is a resource-consuming process. We need to cluster nearly 1 billion sentences into 100 topics, if a new corpus arrives, then the clustering process needs to execute again. For the COMETA method, the limitation is that the meta-model size increases with the translation model size. And the meta-model is hard to process the parameters with a complex structure such as self-attention layers.

Ethics Statement

Training a multilingual NMT model from scratch usually costs expensive computing resources, researching CLL can effectively reduce resource consumption and carbon emissions. And using the meta-model to alleviate the catastrophic forgetting provides a new perspective for studying continual learning.

Acknowledgements

This research was supported in part by the National Key Research and Development Program of China

ISO	Dataset	M2M-418m		Finetune	
		zh->xx	xx->zh	zh->xx	xx->zh
en	**CN-25**	14.90	22.03	**16.30**	**25.90**
	filtered out			15.10	22.41
de	**CN-25**	10.60	19.11	**11.80**	**21.93**
	filtered out			11.30	19.42
fr	**CN-25**	11.50	18.29	**12.70**	**22.22**
	filtered out			11.80	19.63
fi	**CN-25**	7.40	16.06	**8.00**	**17.49**
	filtered out			7.30	15.08
cs	**CN-25**	8.80	18.86	**9.70**	**20.75**
	filtered out			9.10	17.84
ru	**CN-25**	10.10	17.65	**11.20**	**19.54**
	filtered out			10.20	17.39
tr	**CN-25**	6.00	17.27	**6.70**	**19.39**
	filtered out			6.20	16.74

Table 3: M2M-418m is finetuned on CN-25 and filtered out sentences, then evaluated on TED. The red font means performance decline.

ISO	Dataset	M2M-418m	Finetune	From scratch
en	**CN-25**	17.8	**18.8**	**10.0**
	filtered out		17.2	4.9

Table 4: M2M-418m is finetuned on CN-25 and filtered out sentences, then evaluated on wmt20. The red font means performance decline.

ISO	xx->zh		zh->xx	
	TED	CN25	TED	CN25
en	4.07	**9.48**	**16.52**	14.71
de	3.50	**7.65**	8.75	**9.73**
nl	3.45	**7.76**	**9.94**	9.46
sv	3.69	**7.64**	7.88	**10.30**
fr	3.94	**8.12**	**16.69**	9.63
pt	3.03	**8.76**	8.90	**11.10**
es	3.26	**8.43**	10.66	**11.51**
ja	1.96	**4.69**	**1.44**	0.65
ko	3.49	**6.54**	**4.00**	2.48
vi	3.66	**7.49**	12.80	**14.50**

Table 5: The performance of the transformer-base trained on TED and CN-25 corpus respectively.

2021ZD0112905, in part by the Major Key Project of PCL PCL2021A06, PCL2022D01, in part by the National Natural Science Foundation of China under Grants 62106115, 62006062, and 62176076, in part by the Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies 2022B1212010005.

RS	Method	Transformer-small		Transformer-base	
		Stage-0	Stage-1	Stage-0	Stage-1
0	**EWC**	13.74	7.94	23.60	15.48
	MAS	13.45	8.38	23.69	15.40
	Ours	13.56	**9.47**	23.56	**15.92**
1k	**EWC**	13.55	8.84	23.39	18.46
	MAS	13.58	8.75	23.69	18.62
	Ours	13.75	**10.05**	23.12	**19.45**
2k	**EWC**	13.56	9.67	23.52	19.79
	MAS	13.54	9.52	23.67	19.55
	Ours	13.53	**10.59**	23.45	**20.13**

Table 6: The performance of baselines in CLCL task. RS means the number of replay samples in each old direction.

RS	Method	Transformer-small			Transformer-base		
		Stage-0	Stage-1	Stage-2	Stage-0	Stage-1	Stage-2
0	**EWC**	15.16	9.89	5.89	25.33	16.52	10.38
	MAS	15.57	9.46	5.92	26.03	16.86	10.85
	Ours	15.09	9.89	**6.42**	24.95	17.10	**11.37**
1k	**EWC**	15.39	10.01	8.43	25.90	19.87	17.10
	MAS	15.55	9.86	8.25	25.49	20.00	17.21
	Ours	15.41	10.61	**8.93**	25.72	21.01	**17.57**
2k	**EWC**	15.37	10.82	9.19	25.25	21.07	18.31
	MAS	15.07	10.58	9.08	25.63	20.98	18.17
	Ours	15.40	11.51	**9.56**	25.35	21.75	**18.42**

Table 7: The performance of baselines in LFCL task with different replay settings.

References

2021. *Machine Translation : 17th China Conference, CCMT 2021, Xining, China, October 8-10, 2021, Revised Selected Papers*, 1st ed 2021. edition. Communications in computer and information science 1464. Springer Singapore : Imprint: Springer, Singapore.

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina Espaa-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (wmt21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. 2018. Memory aware synapses: Learning what (not) to forget. In *Computer Vision – ECCV 2018*, pages 144–161, Cham. Springer International Publishing.

Mikel Artetxe and Holger Schwenk. 2019. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of*

the Association for Computational Linguistics, 7:597–610.

Loïc Barrault, Ondřej Bojar, Marta R Costa-Jussa, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, et al. 2019. Findings of the 2019 conference on machine translation (wmt19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61.

Alexandre Berard. 2021. Continual learning in multilingual NMT via language-specific embeddings. In *Proceedings of the Sixth Conference on Machine Translation*, pages 542–565, Online. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, et al. 2017. Findings of the 2017 conference on machine translation (wmt17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, et al. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198.

Yue Cao, Hao-Ran Wei, Boxing Chen, and Xiaojun Wan. 2021. Continual learning for neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3964–3974.

Cyprien de Masson d'Autume, Sebastian Ruder, Lingpeng Kong, and Dani Yogatama. 2019. Episodic memory in lifelong language learning. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Charles Dugas, Yoshua Bengio, François Bélisle, Claude Nadeau, and René Garcia. 2000. Incorporating second-order functional knowledge for better option pricing. *Advances in neural information processing systems*, 13.

Carlos Escolano, Marta R. Costa-jussà, and José A. R. Fonollosa. 2019. From bilingual to multilingual neural machine translation by incremental training. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 236–242, Florence, Italy. Association for Computational Linguistics.

Carlos Escolano, Marta R. Costa-jussà, José A. R. Fonollosa, and Mikel Artetxe. 2020. Training multilingual machine translation by alternately freezing language-specific encoders-decoders. *CoRR*, abs/2006.01594.

Carlos Escolano, Marta R. Costa-jussà, José A. R. Fonollosa, and Mikel Artetxe. 2021. Multilingual machine translation: Closing the gap between shared and language-specific encoder-decoders. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 944–948, Online. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Xavier Garcia, Noah Constant, Ankur Parikh, and Orhan Firat. 2021. Towards continual learning for multilingual machine translation via vocabulary substitution. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1184–1192, Online. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Mandy Guo, Qinlan Shen, Yinfei Yang, Heming Ge, Daniel Cer, Gustavo Hernandez Abrego, Keith Stevens, Noah Constant, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Effective parallel corpus mining using bilingual sentence embeddings. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 165–176, Brussels, Belgium. Association for Computational Linguistics.

Yen-Chang Hsu, Yen-Cheng Liu, Anita Ramasamy, and Zsolt Kira. 2018. Re-evaluating continual learning scenarios: A categorization and case for strong baselines. In *NeurIPS Continual learning Workshop*.

Aman Hussain, Nithin Holla, Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2021. Towards a robust experimental framework and benchmark for lifelong language learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. 3rd International Conference on Learning Representations, ICLR 2015.

Bin Liang, Xiang Li, Lin Gui, Yonghao Fu, Yulan He, Min Yang, and Ruifeng Xu. 2022. Few-shot aspect category sentiment analysis via meta-learning. *ACM Transactions on Information Systems (TOIS)*.

Bin Liang, Rongdi Yin, Jiachen Du, Lin Gui, Yulan He, Min Yang, and Ruifeng Xu. 2021. Embedding refinement framework for targeted aspect-based sentiment analysis. *IEEE Transactions on Affective Computing*.

Sungwon Lyu, Bokyung Son, Kichang Yang, and Jaekyoung Bae. 2020. Revisiting Modularized Multilingual NMT to Meet Industrial Demands. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5905–5918, Online. Association for Computational Linguistics.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2019. Putting evaluation in context: Contextual embeddings improve machine translation evaluation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2799–2808, Florence, Italy. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Toshiaki Nakazawa, Chenchen Ding, Hideya Mino, Isao Goto, Graham Neubig, and Sadao Kurohashi. 2016. Overview of the 3rd workshop on Asian translation. In *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, pages 1–46, Osaka, Japan. The COLING 2016 Organizing Committee.

Toshiaki Nakazawa, Nobushige Doi, Shohei Higashiyama, Chenchen Ding, Raj Dabre, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Yusuke Oda, Shantipriya Parida, Ondřej Bojar, and Sadao Kurohashi. 2019. Overview of the 6th workshop on Asian translation. In *Proceedings of the 6th Workshop on Asian Translation*, pages 1–35, Hong Kong, China. Association for Computational Linguistics.

Toshiaki Nakazawa, Shohei Higashiyama, Chenchen Ding, Hideya Mino, Isao Goto, Hideto Kazawa, Yusuke Oda, Graham Neubig, and Sadao Kurohashi. 2017. Overview of the 4th workshop on Asian translation. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 1–54, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Toshiaki Nakazawa, Hideya Mino, Isao Goto, Sadao Kurohashi, and Eiichiro Sumita. 2014. Overview of the 1st workshop on Asian translation. In *Proceedings of the 1st Workshop on Asian Translation (WAT2014)*, pages 1–19, Tokyo, Japan. Workshop on Asian Translation.

Toshiaki Nakazawa, Hideya Mino, Isao Goto, Graham Neubig, Sadao Kurohashi, and Eiichiro Sumita. 2015. Overview of the 2nd workshop on Asian translation. In *Proceedings of the 2nd Workshop on Asian Translation (WAT2015)*, pages 1–28, Kyoto, Japan. Workshop on Asian Translation.

Toshiaki Nakazawa, Hideki Nakayama, Chenchen Ding, Raj Dabre, Shohei Higashiyama, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, Chenhui Chu, Akiko Eriguchi, Kaori Abe, Yusuke Oda, and Sadao Kurohashi. 2021. Overview of the 8th workshop on Asian translation. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 1–45, Online. Association for Computational Linguistics.

Toshiaki Nakazawa, Hideki Nakayama, Chenchen Ding, Raj Dabre, Shohei Higashiyama, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, and Sadao Kurohashi. 2020. Overview of the 7th workshop on Asian translation. In *Proceedings of the 7th Workshop on Asian Translation*, pages 1–44, Suzhou, China. Association for Computational Linguistics.

Toshiaki Nakazawa, Katsuhito Sudoh, Shohei Higashiyama, Chenchen Ding, Raj Dabre, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, and Sadao Kurohashi. 2018. Overview of the 5th workshop on Asian translation. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation: 5th Workshop on Asian Translation: 5th Workshop on Asian Translation*, Hong Kong. Association for Computational Linguistics.

Mariana Neves, Antonio Jimeno Yepes, Aurélie Névéol, Cristian Grozea, Amy Siu, Madeleine Kittner, and Karin Verspoor. 2018. Findings of the WMT 2018 biomedical translation shared task: Evaluation on Medline test sets. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 324–339, Belgium, Brussels. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana. Association for Computational Linguistics.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan.

2021. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Lucia Specia, Zhenhao Li, Juan Pino, Vishrav Chaudhary, Francisco Guzmán, Graham Neubig, Nadir Durrani, Yonatan Belinkov, Philipp Koehn, Hassan Sajjad, et al. 2020. Findings of the wmt 2020 shared task on machine translation robustness. In *Proceedings of the Fifth Conference on Machine Translation*, pages 76–91.

Brian Thompson, Jeremy Gwinnup, Huda Khayrallah, Kevin Duh, and Philipp Koehn. 2019. Overcoming catastrophic forgetting during domain adaptation of neural machine translation. 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 2062–2068.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Zirui Wang, Sanket Vaibhav Mehta, Barnabas Poczos, and Jaime Carbonell. 2020. Efficient meta lifelong-learning with limited memory. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 535–548, Online. Association for Computational Linguistics.

Muyun Yang, Xixin Hu, Hao Xiong, Jiayi Wang, Yiliyaer Jiaermuhamaiti, Zhongjun He, Weihua Luo, and Shujian Huang. 2019. Ccmt 2019 machine translation evaluation report. In *Machine Translation*, pages 105–128, Singapore. Springer Singapore.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

A Analyze the continual language learning task

In this section, we empirically analyze the challenge in CLL. We train a multilingual NMT model based on multilingual vocabulary released by M2M, then add new languages from different families. To be specific, at the first stage, we train a multilingual NMT model on seven languages from three families including Germanic (Dutch, German, Swedish), Romance (Portuguese, Spanish), and Slavic (Russian, Czech) language families. At the second stage, we directly finetune the model on a new language and then replay 2000 samples of each old direction. We execute the experiments on three new languages including English (Germanic), French (Romance), and Polish (Slavic). As shown in Table 8, the old languages from the same family with the new language have the minimum degree of forgetting. However, the forgetting of all families is still remarkable.

A.1 Gradient visualization of the embedding

To investigate the reason for the remarkable forgetting, we visualize the gradients of language-specific embeddings. We select 1000 frequent tokens of each language by traversing the entire corpus. Then, we record the gradient of each embedding in the finetuning process and average the gradient along the feature dimension. The visualization of gradients is shown in Figure 5, the vertical axis shows the frequent tokens of different languages, the horizontal axis represents the training step of the new language.

Compare the magnitude of the gradient visualized in Figure 5, we can observe several phenomenons:

- If the new and old languages use different scripts, the gradient on the old languages' embedding is minimal. For example, Russian (ru, Slavic) uses Cyrillic script, and other languages use Latin script, the magnitude of the gradient of Russian-specific embedding is very small when learning any new languages.
- If the new and old languages use the same script, learning a new language will create a larger gradient of the languages from the same family. For example, when learning English (en, Germanic), the gradient of the Germanic languages (German, Dutch, Swedish) is large.
- At the beginning of learning a new language, the gradient magnitude is larger (in a darker color) than the subsequent process.

Task	zh-de	zh-nl	zh-sv	AVG	zh-pt	zh-es	AVG	zh-ru	zh-cs	AVG	zh-en	zh-fr	zh-pl
multi-tasks	11.79	13.61	13.63	13.01	18.26	18.77	18.52	11.66	9.21	10.43			
+ en	6.13	6.96	6.97	**6.69**	9.34	9.33	9.33	6.96	4.14	5.55	22.79		
+ fr	6.08	6.50	6.29	6.29	8.82	10.61	**9.72**	6.83	4.02	5.43		16.58	
+ pl	5.99	5.87	6.50	6.12	9.27	9.38	9.32	7.75	4.65	**6.20**			9.63

Table 8: The influence of adding a new language on old languages. Languages in same color come from the same family.

A.2 The semantic shift phenomenon

We guess that the above phenomenons are due to the BPE-based (Sennrich et al., 2016) multilingual vocabulary. Under the multi-tasks scenario, employing a subword-shared vocabulary across multi-languages can promote the semantic knowledge learning, multilingual semantic knowledge can be learned from shared tokens embedding. While in the CLL scenario, shared token embeddings are trained across multiple stages, and the new knowledge may wash out old knowledge. It is worth noting that the semantics of embeddings does not change in the continual domain learning scenario. In the CLL scenario, the shared token may represent different semantics in different languages. The language shared embedding will be updated in new language learning stages, which brings more challenges for the CLL method. We call this phenomenon as *semantic shift*.

To verify the existence of the semantic shift phenomenon, in Figure 6, we present the L2-Norm and L2-Distance of embedding when continually learning the new languages. In each sub-figure, 1000 frequent

tokens of new and old languages and 100 shared tokens (or less than 100 if not exist) are selected to plot the curves. From the L2-Norm and L2-Distance curves, we observe that the shared token embedding varies faster than the old languages' embedding except for Russian (not Latin script), which proves the existence of semantic shift. It is noticed that the semantic shift exists in the languages which use the same script due to the BPE-based multilingual vocabulary. Inspired by the semantic shift phenomenon, to reduce the catastrophic forgetting, we can 1) control the optimizer's updating of the shared embedding (the strategy of COMETA), or 2) utilize a new method to generate the multilingual vocabulary with fewer shared tokens.

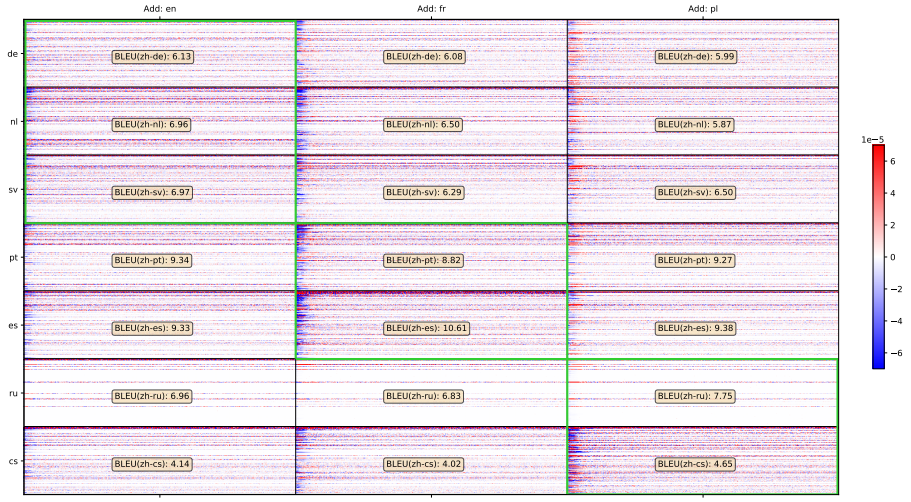

Figure 5: Gradients of the language-specific embedding. The vertical axis shows the tokens of the old language, the horizontal axis represents the finetuning step of new languages. The intensity of the color indicates the magnitude of the gradient, red indicates a positive value while blue indicates a negative value.

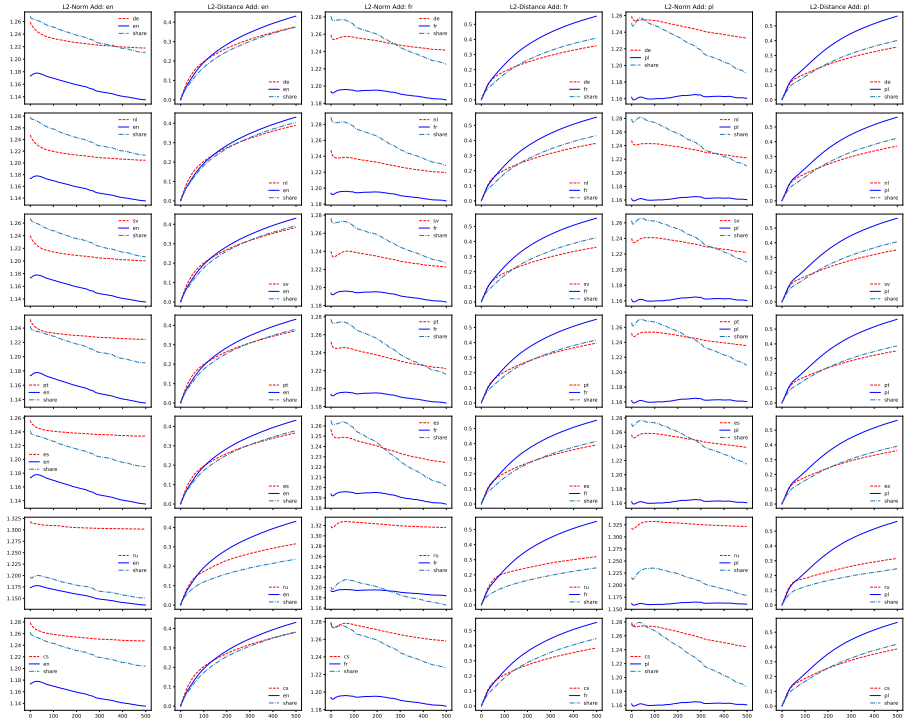

Figure 6: The L2-Norm and L2-Distance when finetuning new languages. The horizontal axis represents the finetuning step. The L2-Distance is calculated according to step 0 and the current step.

B The detail results and hyper-parameters of experiments

B.1 Text classification results by FastText

ISO	Politics	Education	Fashion	Sports	Entertainme	Technology	Healthy	Fiction	Game	Social	Asset	Stocks	Lottery	Economics	Other
ar	26.91%	11.97%	8.15%	7.42%	5.49%	6.58%	6.38%	5.08%	4.04%	1.98%	2.29%	1.58%	0.58%	0.81%	10.73%
cs	20.56%	11.91%	8.57%	8.23%	6.68%	6.93%	6.96%	5.84%	4.74%	2.25%	2.68%	1.47%	0.58%	0.68%	11.89%
de	21.69%	10.74%	8.61%	7.97%	6.28%	7.78%	7.34%	5.48%	4.92%	1.86%	2.92%	1.54%	0.49%	0.65%	11.73%
el	21.29%	11.82%	8.39%	8.49%	6.57%	7.42%	6.04%	5.85%	4.51%	2.43%	2.11%	1.68%	0.58%	0.70%	12.13%
en	26.07%	8.62%	6.37%	10.40%	6.16%	7.09%	6.01%	6.65%	3.48%	1.79%	2.78%	1.69%	0.31%	0.69%	11.89%
es	24.99%	9.90%	7.69%	8.51%	5.98%	7.98%	6.70%	5.04%	4.37%	1.88%	2.66%	1.58%	0.43%	0.65%	11.63%
fa	15.75%	13.56%	11.42%	7.06%	9.11%	6.08%	5.09%	6.28%	6.18%	2.33%	1.66%	1.10%	0.66%	0.51%	13.20%
fi	16.52%	13.42%	9.88%	7.33%	7.72%	6.93%	6.69%	6.09%	5.57%	2.17%	2.52%	1.26%	0.62%	0.51%	12.78%
fr	22.50%	10.77%	8.89%	8.12%	6.43%	7.75%	6.68%	5.03%	5.44%	1.75%	2.74%	1.54%	0.52%	0.57%	11.29%
he	12.16%	13.04%	10.97%	8.84%	10.29%	5.21%	4.91%	6.52%	5.18%	3.07%	1.89%	0.95%	0.79%	0.32%	15.84%
hi	16.22%	12.64%	10.37%	8.55%	8.00%	5.51%	5.17%	7.43%	4.24%	3.01%	1.96%	1.16%	0.60%	0.55%	14.60%
hu	17.18%	12.98%	9.57%	7.63%	7.73%	7.45%	6.57%	5.28%	6.14%	2.15%	2.26%	1.42%	0.67%	0.54%	12.42%
id	18.61%	11.65%	9.47%	9.98%	7.06%	7.96%	5.56%	6.25%	5.09%	2.37%	1.93%	1.40%	0.54%	0.63%	11.49%
ja	18.58%	12.83%	9.84%	6.16%	9.09%	7.20%	5.69%	3.95%	6.99%	2.50%	2.00%	1.75%	0.74%	0.58%	12.11%
ko	11.45%	13.38%	12.72%	6.96%	9.88%	7.59%	5.58%	5.40%	8.02%	2.13%	1.95%	1.31%	0.75%	0.34%	12.53%
lt	18.45%	13.83%	9.43%	7.46%	7.45%	6.33%	6.32%	5.28%	5.05%	2.42%	2.13%	1.36%	0.65%	0.62%	13.22%
nl	18.41%	11.98%	9.86%	7.36%	6.86%	7.99%	7.24%	5.63%	5.54%	1.85%	2.91%	1.48%	0.52%	0.61%	11.76%
pl	17.62%	12.67%	9.64%	7.28%	7.61%	7.78%	6.95%	5.22%	5.92%	2.01%	2.37%	1.59%	0.63%	0.66%	12.04%
pt	22.08%	11.62%	8.60%	7.53%	6.22%	8.36%	6.59%	5.12%	5.35%	1.87%	2.44%	1.72%	0.49%	0.66%	11.35%
ru	26.95%	10.95%	7.96%	7.57%	6.82%	7.22%	5.72%	4.70%	4.50%	1.90%	2.20%	1.61%	0.49%	0.72%	10.70%
sv	15.14%	13.05%	10.64%	7.70%	7.52%	7.72%	7.15%	6.09%	5.94%	1.91%	2.61%	1.24%	0.58%	0.46%	12.25%
sw	20.56%	10.90%	7.36%	9.27%	8.05%	5.08%	4.18%	6.73%	5.41%	2.97%	1.80%	0.85%	0.77%	0.32%	15.75%
ta	5.08%	12.79%	10.01%	8.95%	14.10%	2.99%	2.56%	9.90%	6.71%	4.42%	1.18%	0.87%	1.22%	0.12%	19.10%
tr	13.94%	12.43%	11.16%	7.81%	8.79%	7.94%	5.93%	6.03%	6.90%	2.29%	1.87%	1.22%	0.60%	0.37%	12.73%
vi	16.97%	11.43%	10.15%	8.52%	9.01%	7.46%	5.05%	6.21%	6.38%	2.84%	1.69%	1.34%	0.61%	0.55%	11.79%
AVG	18.63%	12.04%	9.43%	8.04%	7.80%	6.97%	5.96%	5.88%	5.47%	2.33%	2.22%	1.39%	0.62%	0.55%	12.68%

Table 9: Text classification by FastText model which is trained by the THUCNews dataset. Percentages are calculated by text classification based on aligned Chinese sentences. The distribution of Tamil is different from other languages, which is consistent with the result (Figure 2 right par) of the topic distribution obtained by K-means clustering.

B.2 The hyper-parameters of experiments for CLL tasks

Model Size	Small	Base
Multilingual NMT model		
architecture	transformer	transformer
vocabulary	M2M-100 vocab	M2M-100 vocab
vocabulary size	128k	128k
train epoches	3	3
replay epoches	3	3
metric tool	SacreBLEU	SacreBLEU
encoder layers	3	6
decoder layers	3	6
dimension	128	512
decoder ffn dim	512	2048
heads	8	16
optimizer	Adam	Adam
high-frequency tokens	5000	5000
retain loss weight (γ)	5.0	5.0
use fp16	True	True
learning rate	0.0005	0.0005
max source positions	256	256
share all embeddings	True	True
criterion	label-smoothed-cross-entropy	
label smoothing	0.1	0.1
dropout	0.1	0.1
accumulation steps	8	16
batch size	128	64
replay samples	0/1000/2000	0/1000/2000
Meta model (only for COMETA)		
dimension	128	512
meta learning rate	0.001	0.001
parameters number	296705	1182209
criterion	MSE	MSE
dropout	0.5	0.5
kernal sizes	(2,3,4)	(2,3,4)
kernal number	256	256
pool operator	avg-pool-1d	avg-pool-1d
activation function	Softplus	Softplus

Table 10: The hyper-parameters of experiments

C EN-25 corpus

We construct the EN-25 corpus with the same process as CN-25. Due to the large amount of English-centric corpus, we only sample part of the corpus for domain alignment.

ISO	Language	Family	Script	Bitext Number (k)						
				Train	Valid	Test	Aligned	Sampled	Refined	Original
nl	Dutch	Germanic	Latin	1780.2	5.0	10.0	1795.2	2000.0	4.28E+04	5.00E+04
de	German	Germanic	Latin	1759.1	5.0	10.0	1774.1	2000.0	4.72E+04	5.00E+04
sv	Swedish	Germanic	Latin	1744.1	5.0	10.0	1759.1	2000.0	3.56E+04	5.00E+04
fr	French	Romance	Latin	1720.4	5.0	10.0	1735.4	2000.0	4.90E+04	5.00E+04
pt	Portuguese	Romance	Latin	1700.9	5.0	10.0	1715.9	2000.0	4.80E+04	5.00E+04
es	Spanish	Romance	Latin	1671.0	5.0	10.0	1686.0	2000.0	4.94E+04	5.00E+04
ja	Japanese	Japonic	Kanji;Kana	1582.0	5.0	10.0	1597.0	2000.0	1.80E+04	4.09E+04
ko	Korean	Koreanic	Hangul	1558.6	5.0	10.0	1573.6	2000.0	8.94E+03	1.94E+04
vi	Vietnamese	Vietic	Latin	1625.2	5.0	10.0	1640.2	2000.0	2.42E+04	3.43E+04
ar	Arabic	Arabic	Arabic	1597.4	5.0	10.0	1612.4	2000.0	3.66E+04	4.97E+04
fa	Farsi	Iranian	Arabic	1720.0	5.0	10.0	1735.0	2000.0	1.44E+04	2.46E+04
he	Hebrew	Semitic	Hebrew	1639.7	5.0	10.0	1654.7	2000.0	1.48E+04	2.52E+04
fi	Finnish	Uralic	Latin	1731.5	5.0	10.0	1746.5	2000.0	2.26E+04	3.60E+04
hu	Hungarian	Uralic	Latin	1713.0	5.0	10.0	1728.0	2000.0	2.33E+04	3.64E+04
lt	Lithuanian	Baltic	Latin	1585.3	5.0	10.0	1600.3	2000.0	1.41E+04	2.33E+04
pl	Polish	Slavic	Latin	1726.0	5.0	10.0	1741.0	2000.0	3.66E+04	5.00E+04
ru	Russian	Slavic	Cyrillic	1672.5	5.0	10.0	1687.5	2000.0	4.37E+04	5.00E+04
cs	Czech	Slavic	Latin	1723.9	5.0	10.0	1738.9	2000.0	3.51E+04	5.00E+04
hi	Hindi	Indo-Atyan	Devanagari	1637.7	5.0	10.0	1652.7	2000.0	7.66E+03	1.51E+04
bn	Bengali	Indo-Atyan	Eastern-Nagari	1511.9	5.0	10.0	1526.9	2000.0	3.83E+03	1.01E+04
ta	Tamil	Dravidian	Tamil	576.6	5.0	10.0	591.6	722.9	7.23E+02	7.29E+03
id	Indonesian	Malayo-Polyn	Latin	1528.5	5.0	10.0	1543.5	2000.0	4.02E+04	5.00E+04
sw	Swahili	Niger-Congo	Latin	1320.2	5.0	10.0	1335.2	1896.7	1.90E+03	5.76E+03
tr	Turkish	Turkic	Latin	1683.2	5.0	10.0	1698.2	2000.0	2.74E+04	4.71E+04
el	Greek	Hellenic	Greek	1665.2	5.0	10.0	1680.2	2000.0	3.22E+04	4.93E+04

Table 11: The statistics of 25 languages aligned with English (EN-25). The domain alignment is processed on the sampled corpus. "Sampled": the number of sampled stentences from the refined corpus.

Figure 7: Correlation matrix of K-means topic distribution on EN-25. Left: original correlation matrix. Right: adjusted correlation matrix.