Entropy-Based Vocabulary Substitution for Incremental Learning in Multilingual Neural Machine Translation

Kaiyu Huang[1], **Peng Li**[*1], **Jin Ma**[5,6], **Yang Liu**[* 1,2,3,4,7,8]
[1]Institute for AI Industry Research, Tsinghua University, Beijing, China
[2]Dept. of Comp. Sci. & Tech., Institute for AI, Tsinghua University, Beijing, China
[3]Beijing National Research Center for Information Science and Technology
[4]Beijing Academy of Artificial Intelligence, Beijing, China [5]Tencent
[6]Sch. of Comp. Sci. & Tech., University of Science and Technology of China
[7]International Innovation Center of Tsinghua University, Shanghai, China
[8]Quan Cheng Laboratory
{huangkaiyu,lipeng}@air.tsinghua.edu.cn
majin01@mail.ustc.edu.cn; liuyang2011@tsinghua.edu.cn

Abstract

In a practical real-world scenario, the long-standing goal is that a universal multilingual translation model can be incrementally updated when new language pairs arrive. Specifically, the initial vocabulary only covers some of the words in new languages, which hurts the translation quality for incremental learning. Although existing approaches attempt to address this issue by replacing the original vocabulary with a rebuilt vocabulary or constructing independent language-specific vocabularies, these methods can not meet the following three demands simultaneously: (1) High translation quality for original and incremental languages, (2) low cost for model training, (3) low time overhead for preprocessing. In this work, we propose an entropy-based vocabulary substitution (**EVS**) method that just needs to walk through new language pairs for incremental learning in a large-scale multilingual data updating while remaining the size of the vocabulary. Our method has access to learn new knowledge from updated training samples incrementally while keeping high translation quality for original language pairs, alleviating the issue of catastrophic forgetting. Results of experiments show that EVS can achieve better performance and save excess overhead for incremental learning in the multilingual machine translation task.[1]

1 Introduction

Multilingual neural machine translation (NMT) aims at performing multi-directional translation with a single model. Due to its effectiveness and efficiency, it has attracted intensive attention in recent years (Firat et al., 2016; Johnson et al., 2017; Gu et al., 2018; Wenzek et al., 2021; Goyal et al., 2022). Typically, a multilingual NMT system is trained on a multilingual parallel corpus covering all the interested language pairs. As a result, the work of preparing and optimizing the characteristics of the training corpus makes it cumbersome and time-consuming to extend an existing multilingual NMT system to support new language pairs (Dabre et al., 2020).

The stream of data evolves over time by adding new language pairs in the real-world scenario. Due to the high-cost of GPU resources, an efficient way is to incrementally train the initial translation model, instead of training the model from scratch every time when the new language pairs arrive (Neubig and Hu, 2018; Lakew et al., 2018; Chronopoulou et al., 2020; Garcia et al., 2021). Incremental learning is a potential solution, which allows neural models to learn new knowledge from updated training samples while inheriting the original knowledge (Kirkpatrick et al., 2017; De Lange et al., 2019; Yin et al., 2022). Therefore, the high-cost of incorporating new languages in multilingual NMT models can be alleviated with the incremental training paradigm.

The main challenge in incremental learning is catastrophic forgetting (French, 1993). Moreover, an inevitable part of the incremental learning paradigm for multilingual NMT is how to deal with vocabulary (Dabre et al., 2020; Garcia et al., 2021). A certain amount of the "out-of-vocabulary (OOV)" tokens (<UNK>) will appear in the incremental training samples if we directly utilize the initial vocabulary. This situation will hurt the translation performance naturally (Zhang et al., 2022).

Previous approaches (Chronopoulou et al., 2020;

[*] Corresponding authors: Peng Li (lipeng@air.tsinghua.edu.cn) and Yang Liu (liuyang2011@tsinghua.edu.cn)

[1] https://github.com/koukaiu/evs

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 10537 - 10550
December 7-11, 2022 ©2022 Association for Computational Linguistics

Garcia et al., 2021) attempt to address this problem by replacing the initial vocabulary with a new vocabulary. However, these methods suffer from the following two challenges in a further complicated scenario: (1) Excessive time cost in preprocessing, and (2) More OOV tokens on original language pairs. In particular, the former is due to the rebuilt processing of vocabularies with the standard Byte-Pair Encoding (BPE) (Sennrich et al., 2016) or Sentencepiece Model (SPM) (Kudo and Richardson, 2018) procedures on the sum of the initial and incremental training data. And the latter is due to the diversity between the incremental and original languages (Tan et al., 2019).

Another intuitive and simple scheme for vocabularies is to expand the embedding layers of original NMT models directly (Lakew et al., 2018). However, growing tokens lead to representation sparsity, which may hurt neural translation model learning (Sennrich and Zhang, 2019; Ding et al., 2019). More importantly, as the data continues to be updated rapidly, the embedding size will grow uncontrollably. Both the memory and time overhead increase relentlessly and it is not a sustainable strategy.

In this work, considering the diversity of languages, we construct a further complicated and comprehensive setting for incremental learning in multilingual NMT. Due to the dissimilar scripts and diverse language branches, there is a little token overlap (about a quarter) between the original and rebuilt vocabularies. To alleviate this issue, we propose an entropy-based vocabulary substitution (**EVS**) method to retain the learned knowledge from the original translation model with minimizing the case of <UNK> tokens in texts and is suitable for multilingual translation settings. Moreover, our proposed method maintains the same embedding size with the original vocabulary for incremental learning, and does not expand the size of the initial model to keep the memory and time costs of NMT models.

To sum up, our contributions are as follows:

- We propose an entropy-based vocabulary substitution method to alleviate the issue of low token overlap between the initial vocabulary and the rebuilt vocabulary.

- Our method can retain the previously-learned knowledge from the original translation model and learn new knowledge from updated training samples incrementally.

- Experiments show that our method can retain the translation performance on original language pairs while achieving high translation qualities for new incremental language pairs without the excess overhead.

2 Related Work

Past works develop a universal translation model to provide high-quality translation service between any pair of languages (Firat et al., 2016; Johnson et al., 2017; Gu et al., 2018). They leverage knowledge transfer techniques to train neural translation models on a set of languages. The shared knowledge enables the set of languages to help each other (Dong et al., 2015; Firat et al., 2016; Zoph and Knight, 2016). While these approaches are trained on initially selecting a set of languages. The multilingual models need to be retrained when incorporating new languages or data.

Previous approaches study how to adapt a machine translation model to new languages from an updated stream of data timely (Zoph and Knight, 2016; Lakew et al., 2019) by incremental learning. Neubig and Hu (2018) presents two strategies that can rapidly adapt the translation model to new low-resourced languages. And some approaches attempt to improve the translation qualities of pretrained multilingual machine translation models, which incorporate new data (Bapna and Firat, 2019; Tang et al., 2020). Escolano et al. (2021) leverages the language-specific encoders and decoders to incrementally extend a neural translation model from bilingualism to multilingualism.

However, the situation is further complicated, in which the vocabulary needs to be updated in order to avoid the issue of OOV tokens. Lakew et al. (2018) extends the initial model to adapt new languages by a dynamic vocabulary. Chronopoulou et al. (2020) rebuilds a vocabulary from the sum of initial and updated training data with the standard BPE or SPM procedures.

Garcia et al. (2021) constructs an additional vocabulary to replace the initial vocabulary with the same setting of the initial procedure from a new language. However, the translation quality will decrease dramatically for original languages if the rate of token overlap is low between the initial and additional vocabulary. It is not suitable for the situation where the incremental languages are not related to the original languages. Different from previous works, our proposed method utilizes an entropy-based vocabulary substitution strategy with the minimum-cost reconstruction to alleviate the issue of low token overlap. The method can effectively adapt to large-scale updated training data by incremental learning, without the superfluous time and memory overhead.

3 Background

Multilingual NMT utilizes a single encoder-decoder model to handle different translation directions by jointly training on the multilingual parallel dataset. To achieve better translation performance in multilingual training, we share the embeddings for the encoder and decoder embedding layers. To indicate the target language, a prepending language token is appended to each source sentence (Johnson et al., 2017). Formally, given the source sentence $\mathbf{x}' = (x_1, x_2, ..., x_I)$, the modified source sentence is represented as $\mathbf{x} = (l_i, x_1, x_2, ..., x_I)$. l_i represents the target language. And its target sentence is represented as $\mathbf{y} = (y_1, y_2, ..., y_J)$. A sequence of word embeddings $e(\mathbf{x})$ is fed into the encoder component. The probability of a target sentence is given by:

$$p(\mathbf{y}|\mathbf{x};\theta) = \prod_{j=1}^{J} p(y_i|\mathbf{y_{<j}}, \mathbf{x};\theta) \quad (1)$$

where θ is a set of trainable parameters, $\mathbf{y}_{<j}$ are the generated words before the j-th step. To optimize the trainable parameters θ, the training objective for the multilingual NMT translation model is to maximize the log-likelihood $\mathcal{L}$ with the parallel training corpora $\mathcal{P} = \{\mathcal{D}^{l_i}\}_{i=1}^{L}$:

$$\mathcal{L}_{\mathcal{P}}(\theta) = \sum_{\mathcal{D}^{l_i} \in \mathcal{P}} \sum_{(\mathbf{x},\mathbf{y}) \in \mathcal{D}^{l_i}} \log p(\mathbf{y}|\mathbf{x};\theta) \quad (2)$$

where $\mathcal{D}$ is the parallel training set on only one language pair, L represents the available number of language pairs.

4 Approach

In this work, we aim to leverage vocabulary substitution strategies to incrementally update the initial model when new language pairs arrive. The previously-learned knowledge can be retained with the token overlap between the initial vocabulary and a new vocabulary, alleviating the issue of catastrophic forgetting. As a result, we present a variant scheme for incremental learning in multilingual machine translation, as shown in Figure 1, where the substitution method does not change the size of the initial vocabulary. Moreover, to improve the efficiency and practicability of vocabulary substitution methods in the real-world scenario, our method alleviates the issue of excessive time cost in preprocessing. The subsequent subsections introduce the definition of incremental learning in the multilingual machine translation task and discuss how to alleviate the new challenge in the complicated setting.

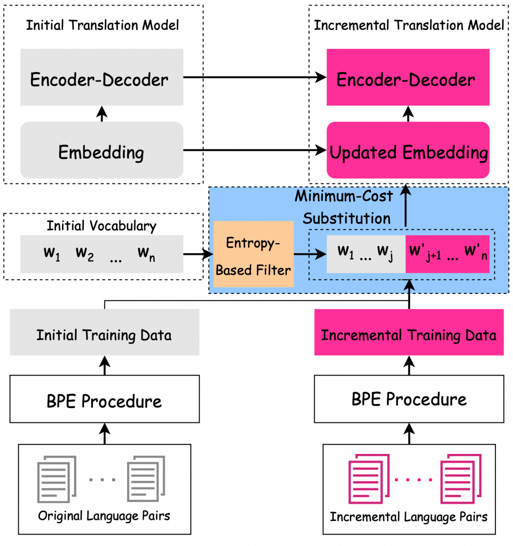

Figure 1: The variant scheme for incremental learning in multilingual NMT. The blue box represents the entropy-based vocabulary substitution method.

4.1 Problem Formulation

As the stream of data is frequently being updated, the number of language pairs L should be updated and changed with time in the real-world scenario. Therefore, it raises a new requirement for the multilingual NMT task, which allows the original NMT model to support new language pairs while retaining the translation quality for original language pairs. Formally, an initial multilingual NMT model $\mathcal{M}_{\mathcal{P}}$, originally trained on first selecting a set of language pairs $L_{\mathcal{P}} = \{1, 2, ..., L\}$. The scope is to extend $\mathcal{M}_{\mathcal{P}}$ to solve the multilingual NMT task on a set of new languages $L_{\mathcal{Q}} \notin L_{\mathcal{P}}$, with $L_{\mathcal{Q}} = \{1, 2, ..., K\}$. And the optimization objective of the multilingual NMT model for incremental learning is given by:

$$\mathcal{L}_{\mathcal{P}\cup\mathcal{Q}}(\theta) = \sum_{\mathcal{D}^{\hat{l}_i} \in \mathcal{P}\cup\mathcal{Q}} \sum_{(\mathbf{x},\mathbf{y}) \in \mathcal{D}^{\hat{l}_i}} \log p(\mathbf{y}|\mathbf{x};\theta) \quad (3)$$

Algorithm 1: Entropy-Based Vocabulary Substitution

```
Input: Training corpora on all language
       pairs $\mathcal{D}$, an initial vocabulary $\mathcal{V}_{\mathcal{N}}$,
       an incremental vocabulary $\mathcal{V}_{\mathcal{M}}$
Merge vocabulary $\mathcal{V} = \mathcal{V}_{\mathcal{N}} \cup \mathcal{V}_{\mathcal{M}}$ ;
Filter $\mathcal{V} \rightarrow \mathcal{V}_{\mathcal{S}}$ using Eq.4 on $\mathcal{D}$ ;
Initialize cost = [][0,1,2,...,$m$] ;
Initialize record = [][0,1,2,...,$m$] ;
//Loop all unduplicated words in the $\mathcal{D}$ ;
while $t$ is OOV do
    for $r \leftarrow 0$ to $m$ do
        min_cost = cost[$t$][$r$] ;
        for $l \leftarrow 0$ to $r$ do
            $s = t[l : r]$ ;
            if $s$ in $\mathcal{V}_{\mathcal{S}}$ then
                $c = 0$ ;
                if $l - 1 \geq 0$ then
                    $c$ =cost[$t$][$l - 1$] ;
                if min_cost $= c$ then
                    min_cost $= c$ ;
                    cost[$t$][$r$]= min_cost ;
                    record[$r$]= $l$ ;
    Replace $t$ with $t^*$ by the record matrix
Rebuild the training data $\mathcal{D} \rightarrow \mathcal{D}_{\mathcal{S}}$
```

where $\mathcal{Q}$ is the updated parallel training corpora and $\mathcal{Q} = \{\mathcal{D}^{K_i}\}_{i=1}^{K}$, K represents the number of the updated language pairs for incremental learning. The initial number of language pairs L increases by K and the combination of the initial and incremental training corpora is represented by $\mathcal{N} \cup \mathcal{M} = \{\mathcal{D}^{\hat{l}_i}\}_{i=1}^{\hat{L}}$.

4.2 Entropy-Based Vocabulary Substitution

Adopting the direct expansion method for vocabularies, one suffers from the risk which increases the difficulty and time-cost of model training. Moreover, a sustainable strategy is to fix the size of vocabulary in the real-world scenario. Due to the limited size of vocabulary, it is beneficial to retain essential words and cover as many languages as possible. To this end, we propose an entropy-based vocabulary substitution method, which consists of two components. In particular, we construct an entropy-based vocabulary filter to keep the fixed size of the merged vocabulary for incremental learning in multilingual machine translation and adopts a substitution strategy with the minimum-cost to address the challenge of OOV tokens.

Entropy-Based Vocabulary Filter. Figure 1 illustrates the process of vocabulary substitution and there will be some more subwords in the merged vocabulary that do not overlap the initial one. As the data continues to be updated rapidly, the merged vocabulary size will grow uncontrollably. Both the memory and time overhead increase relentlessly. Thus we construct a vocabulary filter to fix the size of the merged vocabulary. The filter adopts an entropy-based word importance score to determine whether the subword is retained. Information entropy can be used as a measure of the complexity of a system. We employ this thought to calculate the complexity of a token in the multilingual scenario. Formally, the entropy-based score represents the mean information content of subwords in the parallel corpora of different language pairs. In particular, the entropy-based score is calculated as:

$$\mathcal{H}(w) = -\frac{1}{m_w} \sum_{\hat{l}_i \in \hat{L}} f_{\hat{l}_i}(w) \log f_{\hat{l}_i}(w) \qquad (4)$$

where $f_{\hat{l}_i}(w)$ presents the relative frequency of token w from the training corpus on the current language $\hat{l}_i$, m_w is the length of the current token, $\hat{L}$ represents all the available number of language pairs, i.e., the sum number of initial and incremental language pairs. The entropy-based score represents the complexity of tokens in the multilingual corpus. Discarding the tokens with higher scores increases the chaos of the vocabulary, which makes the translation model difficult to train. While discarding the tokens with lower scores decreases the diversity of the vocabulary, which causes the problem of OOV tokens. Thus, to balance the stability and diversity of the rebuilt vocabulary, we choose the words from both ends of the score list. Specifically, as an extreme case, if tokens only appear in one language, the entropy-based score of this token is zero. It indicates that they are irreplaceable in their corresponding language. And we rerank the score list that considers tokens with a value of zero first.

Minimum-Cost Substitution. The Algorithm 1 also introduces a substitution strategy via a minimum-cost path. As some tokens are discarded by the entropy-based word importance score, these tokens in corpora will be transferred into the <UNK> token. It is detrimental to train neural translation models. Therefore, we propose a token substitution strategy to reconstruct the discarded tokens. A directed graph is constructed to represent all the paths that make up the discarded tokens. The nodes on the graph are reserved tokens in the rebuilt vocabulary. We need to search for the minimum-cost path which can recover the discarded tokens. To reduce the time-cost of searching, we implement the minimum-cost path using the dynamic programming algorithm. The tokens are further fine-grained segmented with a minimum number of subwords. This method not only addresses the problem of OOV tokens but also maintains a relatively short sentence length, which is beneficial to model optimization.

5 Experiments

5.1 Datasets

To examine the translation quality for original and incremental language pairs, we conduct experiments on a popular multilingual machine translation benchmark (WMT-14) (Zhang et al., 2020) as original languages and provide 7 additional languages considered for incremental adaption[2]. We provide the statistics and details of datasets for original and incremental languages in Appendix A.

Language Choice We further make a complicated setting, compared with previous works for adapting translation models to new languages. Past works typically explore the situation of related languages that belong to similar language branches or scripts to original languages. The initial model allows access to data coming from 14 languages. The 7 incremental languages are diverse with respect to scripts and language branches, as shown in Appendix A.2. And both the original and incremental data come WMT training sets for reliable quality.

5.2 Implementation Details

Baselines. For evaluation convincing and future reproducibility, we re-implement a vanilla Transformers (Vaswani et al., 2017) for original languages as the initial model. Then we compare our proposed methods with three intuitive vocabulary substitution baselines for incremental learning. Either original and incremental datasets learn the shared BPE model of 64k tokens using the Sentencepiece library[3].

Unadapted: We build the vocabulary with the standard BPE procedure from the initial parallel training samples. And the multilingual translation model is incrementally trained with the unaltered vocabulary when new language pairs arrive.

Adapted (Garcia et al., 2021): We build a supplementary vocabulary with the standard BPE procedure from only the updated parallel training samples when new language pairs arrive. Then the original vocabulary is replaced with the supplementary vocabulary. The embeddings for subword tokens are reused in the intersection and the original translation model is incrementally trained after vocabulary adaptation.

Frequency-Based: The original vocabulary combines with the supplementary vocabulary to form an entire vocabulary. To keep the embedding size of translation models, the entire vocabulary needs to be truncated. And the frequency of words is an important factor to consider which word should be remained (Sennrich et al., 2016). Therefore, the truncation is based on the frequency of words in all data from high to low.

Training Setup. We implement our experiments using the open-source toolkit fairseq[4] (Ott et al., 2019) which is an advanced neural network library. For a fair comparison, we use Transformers as the basis of multilingual NMT models and follow the configuration of Transformer-Big (Vaswani et al., 2017). We provide more details on the model training in Appendix B.

Evaluation. We evaluate the translation quality of models by the detokenized SacreBLEU score (Post, 2018)[5]. We report the average ΔBLEU on each of three (Low/Med/High) groups for original languages (WMT-14) to indicate the degradation situation of each model. And we also report average BLEU scores on both original and incremental languages to show the overall performance of each method. We utilize beam search decoding with a beam size of 4 and a length penalty of 1.0.

[2] https://www.statmt.org/

[3] https://github.com/google/sentencepiece

[4] https://github.com/pytorch/fairseq

[5] Signature: nrefs:1 | eff:no | tok:13a | smooth:exp | version:2.1.0. English-Chinese: nrefs:1 | eff:no | tok:zh | smooth:exp | version:2.1.0. English-Japanese: nrefs:1 | eff:no | tok:ja-mecab | smooth:exp | version:2.1.0.

Method	Original Lang-Pairs (ΔBLEU)				Incremental Lang-Pairs (BLEU)							
	Low	Med.	High	Avg.	Ja-En	Pl-En	Km-En	Is-En	Ps-En	Ha-En	Ta-En	Avg.
Unadapted	1.04	0.16	0.33	0.56	18.74	31.44	8.36	32.84	15.13	15.89	19.73	20.32
Adapted	-6.87	-11.04	-20.54	-13.09	18.94	31.05	11.45	**33.22**	14.39	15.58	19.93	20.66
Frequency-Based	0.46	0.02	0.09	0.22	12.98	21.78	4.08	23.97	8.59	11.50	12.85	13.68
EVS (Ours)	1.04	-0.08	-0.13	0.34	**19.00**	**31.66**	**11.48**	32.91	**15.38**	**16.17**	**20.14**	**20.96**

Table 1: Results on WMT-14 (original) and incremental languages for xx-to-English. Note that ΔBLEU for original languages (WMT-14) represents the changes in performance of each method compared with the initial model. The highest score for incremental language pairs is highlighted in **bold**.

Method	Original Lang-Pairs (ΔBLEU)				Incremental Lang-Pairs (BLEU)						
	Low	Med.	High	Avg.	En-Ja	En-Pl	En-Is	En-Ps	En-Ha	En-Ta	Avg.
Unadapted	-0.24	-0.40	-0.39	-0.33	12.91	26.64	20.04	4.62	13.30	9.48	14.50
Adapted	-12.68	-19.05	-24.24	-21.00	**13.18**	26.88	**20.73**	3.89	13.01	**10.49**	14.69
Frequency-Based	-0.87	-0.68	-0.41	-0.65	8.38	9.56	8.03	1.15	5.62	1.43	5.70
EVS (Ours)	-0.12	-0.47	-0.45	-0.32	13.14	**26.96**	20.17	**5.20**	**13.54**	9.57	**14.76**

Table 2: Results for the English-to-xx. The highest score for incremental language pairs is highlighted in **bold**.

5.3 Results

Main Results.

As shown in Table 1 and Table 2, our proposed method obtains a better translation quality for both En-xx and xx-En directions in most incremental language pairs, compared with several vocabulary substitution baselines. And considering performance for all language pairs, our method achieves the state-of-the-art performance with respect to the average BLEU scores (27.14 average BLEU on xx-En and 21.32 average BLEU on En-xx). Although the average performance with vocabulary unadapted scheme (Unadapted) is also competitive, this is due to the better performance on original language pairs. The translation qualities of the vocabulary unadapted scheme are worse than our method for both En-xx and xx-En directions in all incremental language pairs, especially in km→en and en→ps translation directions.

We examine the translation quality using different vocabulary substitution methods for both original and incremental language pairs. The result shows the adapted substitution scheme performs poorly for original language pairs, suffering from catastrophic forgetting. And the frequency-based vocabulary substitution method only shows competitive performance for original language pairs. Due to the different scales of training data among translation directions, the subwords with low frequency will be discarded by the frequency-based method. However, these subwords may play an important role in the low-resource scenario. Comparing to the above baselines with vocabulary substitution adaptation, our method achieves better performance for both original and incremental language pairs simultaneously, showing that the entropy-based vocabulary substitution method is effective for incremental learning in multilingual machine translation.

Degradation on Vocabulary Substitution.

To measure the issue of catastrophic forgetting for vocabulary substitution methods, we investigate the translation quality of when the vocabulary is modified. The results of degradation are shown in Figure 2. The baseline is that the initial multilingual translation model is trained incrementally with the original vocabulary. Minor degradation has occurred using vocabulary substitution methods in some of the original languages. In particular, the degradation is more pronounced on English-to-many translation direction. And the results show that the similarity may not be necessarily related to the performance of progress or degradation directly. The performance on Estonian (et) drops slightly, while the performance on Finnish (fi) improves. For another group of similar language, the results on German and French are both positive. Our proposed method significantly alleviates the issue of degradation compared with the other vocabulary substitution methods. Notably, our approach even outperforms the unadapted vocabulary on multiple translation directions and does not incur degradation from the vocabulary substitution.

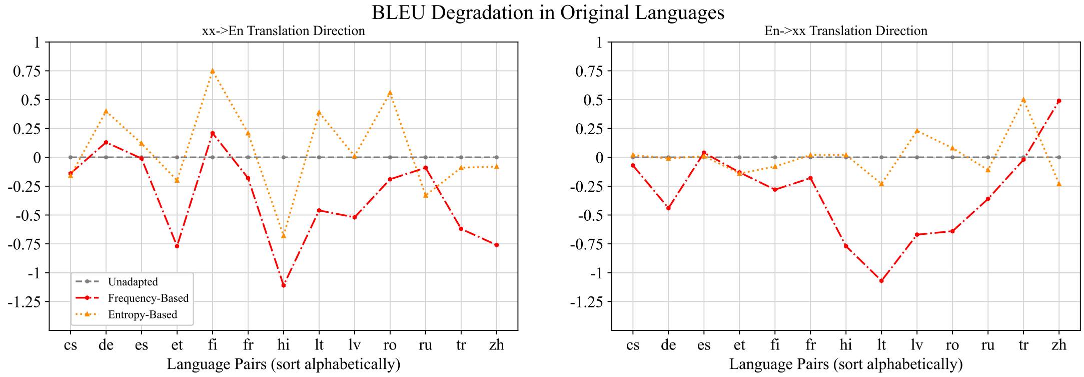

Figure 2: Measuring degradation in BLEU after vocabulary substitution methods. The grey dashed line represents the performance **without** vocabulary substitution. The curves represent the BLEU scores with incremental scheme.

Method	Scheme	Original Lang-Pairs		Incremental Lang-Pairs	
		xx→En	xx←En	xx→En	xx←En
Unadapted	from-scratch	30.29	23.80	20.07	14.50
	incremental	**30.61** (+0.32)	**24.44** (+0.64)	20.38 (**+0.31**)	14.50 (+0.00)
Adapted	from-scratch	16.61	3.62	20.39	14.65
	incremental	16.96 (+0.35)	3.78 (+0.16)	20.66 (+0.27)	14.69 (+0.04)
EVS (Ours)	from-scratch	29.87	23.63	20.65	14.56
	incremental	30.38 (**+0.51**)	24.35 (**+0.78**)	**20.96** (**+0.31**)	**14.76** (**+0.20**)

Table 3: Results on the original and incremental language pairs with different training schemes. The values in parentheses represent the changes in performance between incremental learning and the model trained from scratch. The highest score is highlighted in **bold**.

Incremental Learning and Training from Scratch.

As shown in Table 3, we investigate the benefits of incremental learning when multiple new languages arrive simultaneously, compared with multilingual machine translation models trained from scratch. The results show that the incremental learning scheme has a positive effect on translation qualities for all language pairs. In particular, comparing to the other vocabulary substitution methods, the incremental learning scheme based on our proposed method achieves the greatest progress (up to +0.52/+0.78 for original language pairs on xx-to-En/En-to-xx translation direction; up to +0.31/+0.29 for incremental language pairs on xx-to-En/En-to-xx translation direction). Due to the limitations of tokens coverage for original languages, our method incurs a slight decline on the translation quality for the original languages, compared with the unadapted vocabulary method. The overall performance of our method is competitive. More importantly, our proposed method has access to learn new knowledge from updated training samples incrementally while inheriting the originally learned knowledge, alleviating the issue of catastrophic forgetting. It is more efficient to utilize incremental learning based on our proposed vocabulary substitution strategy than the multilingual machine translation model trained from scratch.

5.4 More Comparisons

We investigate the time and memory overhead of our method, compared with the following stronger incremental strategies based on vocabularies.

Oracle (Chronopoulou et al., 2020): A new vocabulary is rebuilt with the standard BPE procedure from all available training data. The overlap tokens between the new dictionary and the original dictionary inherit previously-learned knowledge.

Expansion (Lakew et al., 2018): we combine the original vocabulary ($\mathcal{V}_{\mathcal{P}}$) and the incremental vocabulary ($\mathcal{V}_{\mathcal{Q}}$) to form an entire vocabulary $\mathcal{V} = \mathcal{V}_{\mathcal{P}} \cup \mathcal{V}_{\mathcal{Q}}$. The embeddings of the initial translation model are expanded to the size of the

Method	Model Size↓	Avg.(BLEU)↑		Time Overhead		
		En←xx	En→xx	Preprocess (hours)↓	Training (hours)↓	Inference (tokens/s)↑
Oracle	243.49M	**27.14**	21.51	14.71	57.31	1780.01
Expansion	279.54M	26.99	**21.68**	**5.02**	82.77	1577.53
EVS (Ours)	**243.27M**	**27.14**	21.32	5.12	**53.05**	**1787.62**

Table 4: Time and memory overhead of vocabulary adaptations. "Inference" indicates the average speed of all languages for English-to-xx directions at inference. The optimal value is highlighted in **bold**.

No.	Filter	Reconstruction	Original Lang-Pairs	Incremental Lang-Pairs
1	Frequency	FMM	29.74	19.64
2	Frequency	Minimum-Cost (ours)	30.06	20.07
3	Entropy (ours)	FMM	29.76	20.57
4	Entropy (ours)	Minimum-Cost (ours)	**30.39**	**20.96**

Table 5: Results on different substitution strategies for xx-to-English.

entire vocabulary ($\mathcal{V}$) and are initialized with the Gaussian distribution.

As shown in Table 4, our proposed method is more efficient and practical than the other strong baselines in the following three aspects: (1) No additional parameter expansion, (2) minimum time overhead on the procedure of preprocessing, training, and inference, (3) negligible performance decrease. As the data continues to be updated rapidly, the embedding size will grow uncontrollably using the Expansion method. It is not sustainable in the oracle setup, because the standard BPE procedure is time-consuming on large-scale data. On the contrary, our method is flexible and sustainable without excess overhead for incremental learning in the multilingual machine translation task.

5.5 Ablation Study

This paper proposes an entropy-based technique for vocabulary adaptation, which consists of the vocabulary filter and the minimum-cost substitution. The technique can alleviate the issues in increment of languages. While there are several potential techniques to serve a similar purpose, e.g., the former max matching (FMM) algorithm (Cheng et al., 1999) and frequency filter (inspired by (Sennrich et al., 2016)). As shown in Table 5, we adopt different combinations of these two techniques and our proposed techniques to show the effectiveness of our method.

The results show that the entropy-based vocabulary filter and the minimum-cost strategy achieve better performance on both original and incremental language pairs. Specifically, the entropy-based filter has a positive effect on the incremental languages according to the comparison between 1 and 3. The minimum-cost has a positive effect on the original languages between 1 and 2.

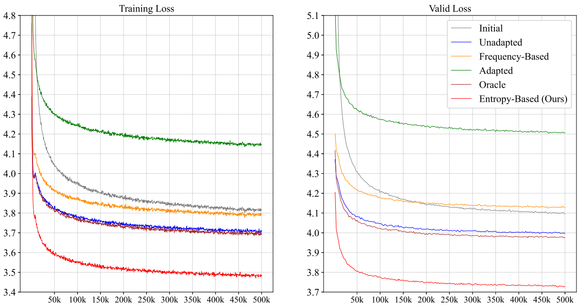

Figure 3: Loss curves of the training and validation process with different vocabulary substitution methods on the xx-to-English translation direction.

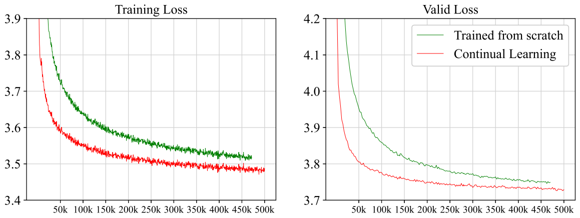

Figure 4: Loss curves of the updated translation model trained from scratch and incremental learning with our method on the xx-to-English translation direction.

6 Analysis

Convergence of Models. We examine the convergence process of the translation model which is trained by different methods. We first depict the loss curves of the training and validation process

Method	Km→En		Hi→En	
	OOV (%)	BLEU	OOV (%)	BLEU
Unadapted	1.7	8.4	**0.0**	26.4
Adapted	**0.0**	**11.5**	40.8	13.5
Frequency-Based	17.0	4.1	0.2	25.4
EVS (Ours)	**0.0**	**11.5**	**0.0**	**26.6**

Table 6: Results on Hindi-to-English and Khmer-to-English, where Hindi is the initial language and Khmer is the incremental language. The OOV rate is counted based on the training corpus.

with different vocabulary substitution methods on the xx-to-English translation direction, as shown in Figure 3. Comparing to the other vocabulary substitution methods, our proposed method achieves the minimum loss and can better incrementally train the initial model. In particular, The loss value falls the most sharply in the first 100K steps with our method. This trend indicates that our method also outperforms the other baselines with limited training time.

In addition, we investigate the loss curves of the incremental translation model with our method to analyze the effect of different learning schemes. Figure 4 plots the results. We find that the incremental learning scheme provides better optimization than the model trained from scratch based on our method because some parameters of the incremental model do not need to be optimized from scratch. It implies that our method can retain the previously-learned knowledge from the original translation model and learn new knowledge from updated training samples incrementally.

Effects of OOV Rate. As shown in Table 6, we investigate the effect of the OOV rate on both one initial language pair and one incremental language pair. The OOV rate may hint at the translation performance before model training. The results show that the method with a lower OOV rate achieves higher translation qualities on all language pairs. In particular, the average OOV rate decreases to 0% by our method.

Effects of Token Overlap. We investigate the situation with a lower token overlap compared with previous methods. And some of the incremental languages have a very low rate of token overlap with the original languages, which is a crucial factor to influence the translation qualities. We collect that the rate of token overlap is less than 20% between the incremental languages and the original languages, as shown in Figure 5. Specifically, there is only 3.7% of tokens overlap between Tamil and the group of old languages (Pashto is 2.4%, Khmer is 0.3%). The results show that our method achieves better performance on these languages (Ta, Km and Ps) in Table 1 and Table 2. And our methods

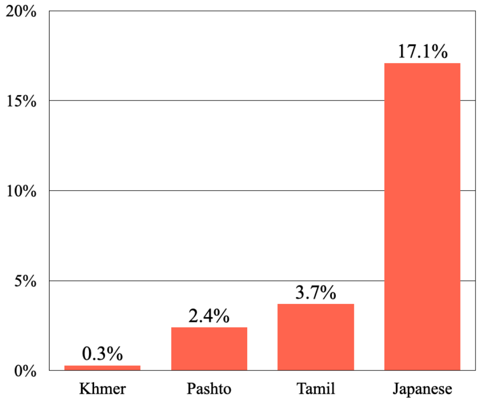

Figure 5: Low rate of token overlap between the original and incremental languages.

7 Conclusion

In this work, we propose an entropy-based vocabulary substitution (EVS) method for incremental learning in multilingual machine translation. And we adopt the incremental learning scheme to learn new knowledge from updated training samples while keeping high translation quality for original language pairs, alleviating the issue of catastrophic forgetting. It is more efficient to utilize incremental learning based on the proposed method than the model trained from scratch. Experimental results demonstrate that the proposed method can also outperform several stronger baselines without the excess time and memory overhead.

Limitations

Our proposed method attempts to extend an existing multilingual NMT system to support a group of new language pairs with an acceptable expense. Besides the advantages, our method has the following limitations:

(1) Diversity of data. We just utilize the parallel data in this work, not monolingual data. The monolingual data is more readily available than high-quality parallel data. It is necessary to investigate the effect of monolingual data for incremental learning in multilingual NMT.

(2) Only the English-Centric translation direction. The translation directions are English-

Centric for both the initial and incremental languages. However, a universal multilingual translation model needs to provide high performance on non-English-centric translation direction in the real-world scenario.

(3) The gap between the practical real-world scenario and our experimental setting. Due to the limited time and the lack of parallel datasets. We only consider 7 incremental language pairs. Moreover, the multilingual NMT model just takes one round of incremental learning, which is different from the situation of constant data updating in the real-world scenario.

The limitations come mainly from the scarcity of data. The in-house data is sensitive, which causes the difference between the real-world scenario and the setting of this work. The incremental learning for the multilingual NMT task is still in its infancy. The definition of this task is vague and further studies will be beneficial. In the future, we will alleviate the above mentioned limitations gradually and further improve the practicability of the NMT system in the real-world scenario.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 61925601), the National Social Science Fund of China (No. 62236011), Beijing Academy of Artificial Intelligence (BAAI), a grant from the Guoqiang Institute, Tsinghua University, and Tencent. We sincerely thank the reviewers for their insightful comments and suggestions to improve the quality of the paper.

References

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, et al. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*.

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1538–1548.

Kwok-Shing Cheng, Gilbert H Young, and Kam-Fai Wong. 1999. A study on word-based and integral-bit chinese text compression algorithms. *Journal of the American Society for Information Science*, 50(3):218–228.

Alexandra Chronopoulou, Dario Stojanovski, and Alexander Fraser. 2020. Reusing a pretrained language model on languages with limited corpora for unsupervised nmt. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2703–2711.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A survey of multilingual neural machine translation. *ACM Computing Surveys (CSUR)*, 53(5):1–38.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. 2019. Continual learning: A comparative study on how to defy forgetting in classification tasks. *arXiv preprint arXiv:1909.08383*, 2(6).

Shuoyang Ding, Adithya Renduchintala, and Kevin Duh. 2019. A call for prudent choice of subword merge operations in neural machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 204–213.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732, Beijing, China. Association for Computational Linguistics.

Carlos Escolano, Marta R Costa-Jussà, and José AR Fonollosa. 2021. From bilingual to multilingual neural-based machine translation by incremental training. *Journal of the Association for Information Science and Technology*, 72(2):190–203.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875.

Robert French. 1993. Catastrophic interference in connectionist networks: Can it be predicted, can it be prevented? *Advances in Neural Information Processing Systems*, 6.

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.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzman, 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.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor OK Li. 2018. Universal neural machine translation for extremely low resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 344–354.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71.

Surafel M Lakew, Aliia Erofeeva, Matteo Negri, Marcello Federico, and Marco Turchi. 2018. Transfer learning in multilingual neural machine translation with dynamic vocabulary. In *International Workshop on Spoken Language Translation*.

Surafel M Lakew, Alina Karakanta, Marcello Federico, Matteo Negri, and Marco Turchi. 2019. Adapting multilingual neural machine translation to unseen languages. In *Proceedings of the 16th International Conference on Spoken Language Translation*.

Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 875–880.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. 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.

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.

Rico Sennrich and Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221.

Xu Tan, Jiale Chen, Di He, Yingce Xia, Tao Qin, and Tie-Yan Liu. 2019. Multilingual neural machine translation with language clustering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 963–973.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010.

Guillaume Wenzek, Vishrav Chaudhary, Angela Fan, Sahir Gomez, Naman Goyal, Somya Jain, Douwe Kiela, Tristan Thrush, and Francisco Guzmán. 2021. Findings of the wmt 2021 shared task on large-scale multilingual machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 89–99.

Wenpeng Yin, Jia Li, and Caiming Xiong. 2022. Contintin: Continual learning from task instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3062–3072.

Biao Zhang, Ankur Bapna, Rico Sennrich, and Orhan Firat. 2020. Share or not? learning to schedule language-specific capacity for multilingual translation. In *International Conference on Learning Representations*.

Shiyue Zhang, Vishrav Chaudhary, Naman Goyal, James Cross, Guillaume Wenzek, Mohit Bansal, and Francisco Guzman. 2022. How robust is neural machine translation to language imbalance in multilingual tokenizer training? *arXiv preprint arXiv:2204.14268*.

Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 30–34, San Diego, California. Association for Computational Linguistics.

A Dataset Details

We conduct experiments on a popular multilingual machine translation benchmark (WMT-14), followed by (Zhang et al., 2020) as original languages. And we gather 7 additional languages for incremental learning from the WMT news translation track. We provide the statistics and of these dataset used in this work. In addition, we introduce the characteristics of languages to analyze the linguistic diversity, as shown in Table 7 and Table 8.

A.1 Data Statistics

As a common setting, we divide the groups of the original and incremental languages into three categories according to the volume of parallel data: low resource (100k~1M), medium resource (1M~10M), and high resource (>10M). For original language pairs, Low resource: Hindi, Lithuanian, Latvian, Romanian, and Turkish; Medium resource: Finnish, German, and Estonian; High resource: Czech, French, Russian, Chinese, and Spanish. For incremental language pairs, Low resource: Hausa and Tamil; Medium resource: Khmer, Icelandic, and Pashto; High resource: Japanese and Polish.

A.2 Language Consideration

"Language family represents a group of languages related through descent from a common ancestor, called the proto-language of that family[6] ." There are various kinds of language families in the real-world. The incremental languages belong to different language families and are large differences in scripts, compared with the original languages. In addition, the grammatical construction of languages and language branch are the consideration factors[7]. The statistics and details of datasets for original and incremental languages are shown in Table 7.

B Model Details

For a fair comparsion, we implement Transformer-Big in all our experiments, which consists of 6 stacked encoder layers, 6 stacked decoder layers, and 16 multi-attention heads. The dimensions of hidden state d_{model} and feed-forward d_{ffn} are 1024 and 4096 respectively. And we use the same learning schedule algorithm and setting with Vaswani et al. (2017). The parameters of multilingual neural models are optimized using Adam optimizer (Kingma and Ba, 2014). Moreover, we reset the learning scheduler and optimizer for incremental learning. To mitigate the imbalance in the multilingual training data, we use the temperature-based sampling scheme with a temperature of $T = 5$ to balance the training data (Arivazhagan et al., 2019). The total training steps are set to 500K with the early stop strategy (patience is 10) and the batch size is 4096 in the training procedure. We evaluate training and inference speed for all models on the same hardware configuration (8 NVIDIA A100 GPUs). We apply half-precision training for speed.

[6] https://en.wikipedia.org/wiki/Languagefamily

[7] https://wit3.fbk.eu/

Code	**Language**	**Genus**	**Family**	**Order**
cs	Czech	Slavic	Indo-European	SVO
de	German	Germanic	Indo-European	SVO
es	Spanish	Romance	Indo-European	SVO
et	Estonian	Finnic	Uralic	SVO
fi	Finnish	Finnic	Uralic	SVO
fr	French	Romance	Indo-European	SVO
hi	Hindi	Indic	Indo-European	SOV
lt	Lithuanian	Baltic	Indo-European	SVO
lv	Latvian	Baltic	Indo-European	SVO
ro	Romanian	Romance	Indo-European	SVO
ru	Russian	Slavic	Indo-European	SVO
tr	Turkish	Turkic	Altaic	SOV
zh	Chinese	Chinese	Sino-Tibetan	SVO
ha	Hausa	West Chadic	Afro-Asiatic	SVO
is	Icelandic	Germanic	Indo-European	SVO
ja	Japanese	Japanese	Japanese	SOV
km	Central Khmer	Khmer	Austro-Asiatic	SVO
pl	Polish	Slavic	Indo-European	SVO
ps	Pashto	Iranian	Indo-European	SOV
ta	Tamil	Southern Dravidian	Dravidian	SOV

Table 7: The characteristics of languages in our setting. The top half part represents the group of the original languages. The second half represents the group of the incremental languages.

Language Pair	Data Sources			# Samples		
	Train	Dev	Test	Train	Dev	Test
Cs-En	WMT19	WMT17	WMT18	64,336,053	3,005	2,983
Fr-En	WMT15	WMT13	WMT14	40,449,146	3,000	3,003
Ru-En	WMT19	WMT18	WMT19	38,492,126	3,000	2,000
Zh-En	WMT19	WMT18	WMT19	25,986,436	3,981	2,000
Es-En	WMT13	WMT13	WMT13	15,182,374	3,004	3,000
Fi-En	WMT19	WMT18	WMT19	6,587,448	3,000	1,996
De-En	WMT14	WMT13	WMT14	4,508,785	3,000	3,003
Et-En	WMT18	WMT18	WMT18	2,175,873	2,000	2,000
Lv-En	WMT17	WMT17	WMT17	637,599	2,003	2,001
Lt-En	WMT19	WMT19	WMT19	635,146	2,000	1,000
Ro-En	WMT16	WMT16	WMT16	610,320	1,999	1,999
Hi-En	WMT14	WMT14	WMT14	313,748	520	2,507
Tr-En	WMT18	WMT17	WMT18	205,756	3,007	3,000
Ja-En	WMT21	WMT20	WMT21	18,001,428	993	1,005
Pl-En	WMT20	WMT20	WMT20	10,206,520	2,000	1,001
Km-En	WMT20	WMT20	WMT20	4,459,608	2,309	2,320
Is-En	WMT21	WMT21	WMT21	4,376,282	2,004	1,000
Ps-En	WMT20	WMT20	WMT20	1,155,942	2,698	2,719
Ha-En	WMT21	WMT21	WMT21	744,856	2,000	997
Ta-En	WMT20	WMT20	WMT20	660,818	1,989	997

Table 8: The Statistics of train, dev, and test data for the original 14 languages (WMT-14) and the incremental 7 languages. The top half part represents the group of the original languages. The second half represents the group of the incremental languages.