Improving NMT from a Low-Resource Source Language: A Use Case from Catalan to Chinese via Spanish

Yongjian Chen[1], **Antonio Toral**[1], **Zhijian Li**[2], **Mireia Farrús**[3,4]
Center for Language and Cognition, University of Groningen, Netherlands[1]
School of Foregin Languages, Guangzhou City University of Technology, China[2]
Centre de Llenguatge i Computació, Universitat de Barcelona, Spain[3]
Institut de Recerca en Sistemes Complexos, Universitat de Barcelona, Spain[4]
{yongjian.chen, a.toral.ruiz}@rug.nl[1]
lizhijian@geu.edu.cn[2]
mfarrus@ub.edu[34]

Abstract

The effectiveness of neural machine translation is markedly constrained in low-resource scenarios, where the scarcity of parallel data hampers the development of robust models. This paper focuses on the scenario where the source language is low-resource and there exists a related high-resource language, for which we introduce a novel approach that combines pivot translation and multilingual training. As a use case we tackle the automatic translation from Catalan to Chinese, using Spanish as an additional language. Our evaluation, conducted on the FLORES-200 benchmark, compares our new approach against a vanilla baseline alongside other models representing various low-resource techniques in the Catalan-to-Chinese context. Experimental results highlight the efficacy of our proposed method, which outperforms existing models, notably demonstrating significant improvements both in translation quality and in lexical diversity.

1 Introduction

The development of neural machine translation (NMT) has considerably benefited translation between language pairs abundant in parallel data, enhancing translation accuracy and fluency across diverse linguistic landscapes (Hassan Awadalla et al., 2018; Popel et al., 2020). However, its effect is challenged by the fact that building an effective NMT system requires a large amount of parallel data. This challenge is particularly pronounced in the case of low-resource languages, that is, language pairs with limited parallel language resources, remaining a significant hurdle in achieving universal communication.

An exemplary case highlighting such hurdle involves the translation dynamics between Catalan and Chinese, (CA–ZH) two languages characterized by limited parallel corpora. The year 2022 marked a significant increase in Chinese investments in Catalonia[1], and the Chinese population emerged as the fourth largest foreign community in Catalonia[2]. These together highlight the growing economic and social interactions between these regions and thus the pressing need for effective communication tools between Catalan and Chinese speakers. Despite the potential benefits, the development of a robust NMT system for the CA–ZH language pair faces notable challenges, primarily due to the scarcity of direct parallel data.

Addressing this gap, previous works have sought to navigate the low-resource landscape of the CA-ZH language pair. The research by Costa-Jussà et al. (2019) was the first work to specifically focus on addressing the low-resource CA–ZH language pair, where they broke new ground by generating non-human-written parallel sentences, i.e. pseudo-parallel corpus via pivot translation and then used them to train ZH→CA NMT models. Another work (Zhou, 2022) concerned building CA–ZH parallel data, where CA–ZH bitexts was first mined from Wikipedia with the help of the open-source *LASER* tookit[3] and then passed to san-

[1]Data taken from https://catalonia.com.
[2]Data taken from https://www.idescat.cat/novetats/?id=4489\&lang=en.
[3]https://github.com/facebookresearch/LASER

ity check according to Kreutzer's (2022) methodology. Therefore, unlike the training datasets created by Costa-Jussà et al. (2019), Zhou's (2022) parallel corpus consists of human-selected bitexts. Subsequently, Liu (2022) used Zhou's (2022) dataset to fine-tune a massively pre-trained multilingual NMT model, i.e. M2M-100 with parameters of 418M by Fan et al. (2020) for CA↔ZH, presenting better translation performance in both directions as compared to the original M2M-100.

Furthermore, other work (Schwenk et al., 2019; El-Kishky et al., 2020; Schwenk et al., 2020), related to building parallel data for many language pairs, included CA–ZH. The multilingual bitexts therein were massively mined from web-based resources and subsequently utilized to train multilingual NMT models like M2M-100, which is also workable with the CA↔ZH translation.

The previous studies on the CA–ZH language pair contributed to enhancing the automated translation between these two languages. However, each focused primarily on employing a singular, specific low-resource NMT technique, e.g. pivot translation in Costa-Jussà et al.(2019), fine-tuning in Liu (2022), multilingual training in Fan el al. (2020), etc. Unlike these approaches, our work aims to propose a novel integration, pivot translation-aided multilingual training (PTAMT), and compare it against existing methods (multilingual training, fine-tuning, pivot translation, fine-tuning coupled with pivot translation). We focus on the CA→ZH translation direction, as a use case in which the source language is low-resource and there is a related higher-resourced language, Spanish (ES). The technique we introduce, PTAMT, uses additional data from ES both as pivot and as multilingual training.

The contributions of our work can be summarized as follows:

1. Our work introduces a novel approach that effectively leverages pseudo-parallel and authentic data to enhance translation quality and mitigate the effects of source-side *machine translationese*, setting a new standard for NMT from low-resource languages.

2. Our work, to the best of our knowledge, is the first one to provide systematic empirical evidence highlighting the effectiveness of different low-resource NMT techniques for the CA–ZH language pair.

3. Our work underscores the important role of a modest amount of authentic parallel data in the target language pair(s) in the training and fine-tuning processes.

2 Related Work

2.1 Low-resource Techniques

Multilingual training refers to training for different language pairs in a single NMT model (Wang et al., 2021) via various methods of sharing parameters, e.g. full parameters sharing (Ha et al., 2016; Johnson et al., 2017; Tan et al., 2019), attention mechanism sharing (Firat et al., 2016; Lu et al., 2018), etc. Through multilingual training, low-resource language pairs can be trained together with high-resource language pairs, and thus desired low-resource languages can benefit from high-resource auxiliary languages when the model learns linguistic knowledge, contextual information, and commonalities, etc. from different languages. Furthermore, if auxiliary languages are related to low-resource languages of interest, they can effectively benefit translation quality in a low-resource scenario (Gu et al., 2018; Neubig and Hu, 2018).

Fine-tuning is performed when a parent NMT model is first trained on high-resource language pairs, and the trained model is used to initialize a child model's parameters, which are subsequently fine-tuned on a low-resource language pair (Zoph et al., 2016). In this way, whereas knowledge learnt from high-resource auxiliary languages can be transferred to low resource languages, the pre-trained NMT model can also be forced to primarily focus on the desired low-resource language pair only. By contrast, since a model has constrained capacity, multilingual training can potentially favor high-resource language pairs due to imbalanced data ratio (Arivazhagan et al., 2019; Wang et al., 2020). Fine-tuning can be combined with multilingual training if a model is first trained on multiple high-resource languages as well as the desired low-resource language pair and then is fine-tuned on the latter only, which has been proved as an effective way to improve low-resource translation (Thillainathan et al., 2021; Adelani et al., 2022).

Pivot translation is applicable for a low-resource translation condition if an auxiliary language has parallel data with both languages of the

low-resource language pair, and this auxiliary language is called a pivot language (Costa-Jussà et al., 2019).

There are mainly two approaches in pivot translation. The first one is the cascade approach, aiming to train two separate translation systems from source to pivot and from pivot to target, and then combine them together for source→target translation, which is common in early statistical machine translation (Cohn and Lapata, 2007; Wu and Wang, 2007; El Kholy et al., 2013).

Another approach is more widely used in state-of-the-art NMT, which is used to synthesize pseudo-parallel data for a low-resource language pair, with data either from the source side synthesized through pivot→source translation (Zheng et al., 2017) or from the target side synthesized through pivot→target translation (Chen et al., 2017). In this case, to ensure the effectiveness of synthetic data via pivot translation, it is important to obtain qualified pivot→source translation or pivot→target translation. For instance, Costa-Jussà et al. (2019) compared two pseudo-parallel CA–ZH parallel corpora. One was built by translating the Spanish sentences from the ES–ZH parallel corpus *United Nations Parallel Corpus v1.0* (Ziemski et al., 2016) into Catalan, whereas the other was created by translating the Spanish sentences from the ES-CA parallel corpus *El Periódico* (Costa-jussà et al., 2014) into Chinese. They used them to train two separate NMT models with same neural network architecture for ZH→CA translation, and discovered that the NMT model trained on the former yielded a higher BLEU score, as the ES→CA translation was of higher quality than the ES→ZH translation.

In contrast to this direct synthesis approach, studies by Lakew et al. (2018) and Currey and Heafield (2019) leveraged pivot resources differently. Rather than solely relying on pivot→source or pivot→target translations to generate pseudo-parallel data, these studies initiated their process with multilingual NMT model training using both source–pivot and target–pivot parallel data. Afterwards, Lakew et al. (2018) used their multilingual NMT model to back-translate source and target language data into the corresponding target and source languages, respectively. This generated pseudo-parallel source–target data was then used alongside the original parallel data to iteratively re-train the multilingual NMT model. Differently, Currey and Heafield (2019) leveraged their multilingual model to back-translate monolingual data from the pivot language into both the source and target languages, thereby obtaining pseudo-parallel source–target data used to further train or fine-tune the model to enhance the direct translation capabilities between the source and target languages.

2.2 Machine Translationese

Machine translationese refers to the artificially impoverished language in MT outputs, marked by reduced fluency, lexical diversity, and distinct syntactic structures compared to original or human-translated texts (Vanmassenhove et al., 2021; Chae and Nenkova, 2009; Ilisei et al., 2010). Such characteristics can make synthetic machine translated data ill-suited for capturing the nuances of human language, potentially leading to deviations in real-world language usage (Dutta Chowdhury et al., 2022). When synthetic data is utilised as training data (as can be the case in pivot translation, see Section 2.1), models may inadvertently learn from the machine translationese present in the synthetic data, leading to the generation of translations or language constructs that are inconsistent with the target language.

Efforts to mitigate translationese have included techniques such as data tagging, where training datasets are annotated to distinguish between original and translated texts. This tagging helps models recognize and avoid translationese during training, as in Riley et al. (2020) and Freitag et al. (2022b). Another approach involves transforming machine-translated texts into more original-like content using style transfer or by re-generating text from abstracted representations like AMR (Jalota et al., 2023; Wein and Schneider, 2024).

These studies mainly focus on improving the quality of machine-translated output by reducing translationese. However, less attention has been given to the effects of source-side artefacts in synthetic data on NMT training. We contemplate this case in this work, comparing different models that deal with synthetic source-side training data in terms of machine translationese.

3 Proposed Method

Our proposed method, PTAMT, couples pivot translation with multilingual training to leverage the advantages of both techniques. Distinct from

previous work (Lakew et al., 2018; Currey and Heafield, 2019), which uses a less related pivot language to initially train a multilingual NMT system for back-translating and synthesizing pseudo-parallel data, PTAMT employs a pivot language (ES) that is linguistically closer to the source language. This choice is informed by the synthetic pivot translation approach demonstrated by Costa-Jussà et al. (2019), which is favored due to the greater linguistic affinity between the source (CA) and pivot (ES) languages as compared to the pivot (ES) and target (ZH) languages (Rapp, 2021).

In our implementation, we used an existing ES–ZH corpus to synthesize pseudo-parallel CA–ZH data by translating ES sentences into CA. This strategy aligns the synthetic side with the source language (CA) and uses authentic data for the target language (ZH), enhancing the model's ability to produce natural output.

Nevertheless, as pivot-translated sentences are inherently machine-translated texts, they are prone to containing machine translationese. To address this, PTAMT strategically leverages the ES–ZH bitexts in a multilingual training setup to facilitate effective pivot-based knowledge transfer. This approach helps to mitigate the potential impact of noise introduced by the synthetic CA input. PTAMT incorporates both source languages (CA and ES) in the encoder, and applies encoder parameter sharing throughout training, which is applicable for both from-scratch training and fine-tuning. The same set of weights and biases is shareable in a single encoder for inputs from both source languages. Multilingual training empowers the model to capture and integrate contextual information from both source languages. Given that the ES data is authentic (human-produced), we hypothesize that PTAMT will aid in reducing the influence of noise from the pivot-translated CA training data on the target ZH output.

4 Experimental Setup

4.1 Data Description

Original Data We used the aforementioned CA–ZH parallel corpus, *CA–ZH Wikipedia* (Zhou, 2022), as the foundation, since it contains human-selected parallel sentences with quality control. We refer to this dataset as *CA–ZH-WIKI*.

Pivot-translated Data We made use of ES–ZH bitexts from the public release *United Nations Parallel Corpus v1.0* (Ziemski et al., 2016), to which we refer as *ES-ZH-UN*. Instead of training an ES→CA automatic translation system from scratch as in (Costa-Jussà et al., 2019), we directly used the open-source NMT model provided by Softcatalà[4] to translate ES sentences from *ES-ZH-UN* into CA and then obtain the pseudo-parallel (synthetic source) CA–ZH United Nations parallel dataset, *CA-ZH-PVT*. Addtionally, by translating CA sentences from *CA-ZH-WIKI* into ES through a CA→ES NMT model[5], we generated a pseudo ES–ZH Wikipedia parallel dataset, *ES-ZH-PVT*, for later data augmentation.

Mixed Data As for the CA–ZH language pair, we concatenated the pivot-translated CA–ZH dataset *CA-ZH-PVT* with *CA-ZH-WIKI*, resulting in the mixed parallel dataset *CA-ZH-MIX*. As regards the ES–ZH language pair, we concatenated the pivot-translated ES–ZH Wikipedia dataset *ES-ZH-PVT* with *ES-ZH-UN* to obtain a mixed parallel dataset *ES-ZH-MIX* for the ES–ZH language pair.

4.2 Models

We implemented one vanilla baseline model, four models based on a pre-trained multilingual model (M2M-100-418), two Transformer-based models trained from scratch, and four PTAMT models to compare different low-resource NMT techniques for CA→ZH translation.

Vanilla Baseline Our vanilla baseline was a Transformer-based model trained on the original parallel corpus *CA-ZH-WIKI*. Due to the small size of this training dataset, rather than using the default Transformer-base configuration (Vaswani et al., 2017), we adopted the architecture setting optimized on 40k training sentence pairs (Araabi and Monz, 2020), which consists of 2 attention heads, 5 encoder and decoder layers, and a 512 embedding dimension.

M2M-100-418M Models As a second baseline, we selected the pretrained model M2M-100 (Fan et al., 2020), which is representative of a model that has taken advantage of multilingual training. M2M-100 is a state-of-the-art massively multilingual translation model, which supports translation between Catalan and Chinese. We opted for the

[4] https://github.com/Softcatala/nmt-models
[5] https://github.com/Softcatala/nmt-models

Language pair	Corpus	# of sentence pairs	
		Training	Validation
CA–ZH	CA-ZH-WIKI	58,328	10,293
	CA-ZH-PVT	17,575,795	2,638
	CA-ZH-MIX	17,634,123	12,931
ES–ZH	ES-ZH-UN	17,575,795	2,638
	ES-ZH-PVT	58,328	10,293
	ES-ZH-MIX	17,634,123	12,931

Table 1: Distribution of the datasets in the experiments.

one with the least size of parameters (418M) taking into account available computational resources as well as comparability across the different models in our experiments. M2M-100-418M is a Transformer-based model that contains 12 encoder and decoder layers with a 1024 embedding dimension.

Large-scale multilingual pre-trained NMT models can be further leveraged to improve low-resource machine translation by fine-tuning them on low-resource language pairs. Therefore, we examined three fine-tuned M2M-100-418M models for the CA→ZH translation. The first one was obtained from the aforementioned work by Liu (2022), accessible on a Hugging Face repository[6], which was solely fine-tuned on the *CA-ZH-WIKI* training dataset. We fine-tuned the second one on the pseudo-parallel dataset *CA-ZH-PVT* and the third one on the mixed parallel dataset *CA-ZH-MIX*. These last two models represent those that leverage pivot translation (either without or with original parallel data) paired with fine-tuning.

From-scratch Trained Models We additionally trained two models from scratch using a Transformer architecture, with 6 encoder and decoder layers and a 512 embedding dimension, respectively on the pseudo-parallel dataset *CA-ZH-PVT* and the mixed parallel dataset *CA-ZH-MIX*. These two models represent those that leverage pivot translation (either without or with original parallel data) under from-scratch training conditions.

PTAMT-enhanced Models We implemented PTAMT to enable simultaneous benefits from *CA-ZH-MIX* and *ES-ZH-MIX* in both from-scratch training and fine-tuning scenarios. Under the from-scratch training condition, we trained a single NMT model that has the same network architecture as the previous from-scratch trained models. Whereas the language pair of interest is still CA–ZH, this model supports both CA→ZH and ES→ZH translation, effectively operating as a many-to-one NMT system. The encoder parameters are shared between CA and ES without increasing the model size, where a special token was added to the source side to specify the input language. Likewise, in the fine-tuning condition, both language pairs were included, and thus M2M-100-418M was fine-tuned on both *CA-ZH-MIX* and *ES-ZH-MIX*.

During the training or fine-tuning phase, ES was engaged as an auxiliary language. Despite potential noise introduced by pivot-translated CA sentences, the model could still learn relevant linguistic properties and characteristics related to CA from their ES equivalents, and thereby enhancing the CA→ZH translation. These two models represent PTAMT in from-scratch training and fine-tuning scenarios, respectively. Furthermore, we applied a second-step fine-tuning to both models on *CA-ZH-WIKI*.

4.3 Preprocessing

As for the parallel datasets used in our experiments, we only worked on sentence-level translation and so we removed lengthy sentence pairs by restricting them to maximum length of 100 words, then split them into training set and validation set (see Table 1), and went through different preprocessing pipelines depending on the models to be trained, as detailed next.

M2M-100-418M models For this model and its fine-tuned variants, including two with PTAMT, we employed the pre-trained SentencePiece tok-

[6] https://huggingface.co/projecte-aina/m2m100_418M_ft_ca_zh

enizer designed for M2M-100[7]. This tokenizer was used to tokenize all the parallel sentences.

Other models For the remaining models, we applied pre-tokenization using separate segmenters tailored for each language, following the approach outlined in Costa-Jussà (2019):

- **Chinese**: Since word boundaries in Chinese are not discernible through whitespace, we utilized the Jieba segmenter [8] to segment Chinese sentences into words.
- **Catalan and Spanish**: We relied on the spaCy tokenizer, specifically the models *ca_core_news_sm* [9] and *es_core_news_sm* [10], respectively. These models were used to identify word boundaries and split contractions (e.g., *l'original* into *l'* + *original*).

Following the pre-tokenization step, we trained SentencePiece BPE models using training sentences from the respective datasets and then proceeded with tokenization.

- **Vanilla Baseline**: Following Araabi and Monz (2020), we trained a tokenizer with 12k BPE merge operations for each language.
- **From-scratch Models**: We sampled 5M sentences from the corresponding training set for each language. We then trained a tokenizer with a character coverage of 1.0 for CA and another one for ZH with character coverage of 0.9995. To determine the optimal vocabulary size for training our tokenizers, we initially used the widely-adopted size of 32k. Subsequently, we conducted experiments by both increasing and decreasing the vocabulary size. In evaluating the performance of tokenizers with different sizes, we assessed the occurrences of the *unk* token in the tokenized data. This resulted in a vocabulary of 35K subwords.
- **PTAMT**: For the proposed PTAMT-enhanced model under the from-scratch condition, we sampled 5M sentences from the ES corpus and concatenated it with the CA samples. This combined set was used to train a joint tokenizer for CA and ES. After testing different vocabulary sizes, we finally created a joint vocabulary of 64K. We retained the Chinese tokenizer used in the previous from-scratch trained model for tokenization.

[7] https://dl.fbaipublicfiles.com/m2m_100/spm.128k.model
[8] https://github.com/fxsjy/jieba
[9] https://spacy.io/models/ca
[10] https://spacy.io/models/es

4.4 Training

For maximum comparability across the various models in this study, we conducted all experiments using a single NVIDIA RTX A5000 GPU card. We trained or fine-tuned all models with the Adam Optimizer and label smoothing cross-entropy loss. The configuration of hyper-parameters for all the NMT models is provided in Table 4 (Appendix A), except that hyper-parameters for training the vanilla baseline followed the ones in Araabi and Monz (2020) (see Table 5 in Appendix A). Addtionally, checkpoints were evaluated at an intervals 5k training or fine-tuning steps on the validation set. Throughout this process, we continuously monitored the models' performance by assessing both training and validation losses. To ensure a balance between achieving convergence and avoiding overfitting, we implemented early stop if there was no improvement in the validation loss over 0.02 across three consecutive validation intervals. The epochs are listed in Table 6 in Appendix B.

4.5 Evaluation

4.5.1 Evaluation Benchmark

We benchmarked the models in this work on *FLORES-200* (Team et al., 2022). We used 1012 sentence pairs from its *devtest* set to evaluate the translation quality in the CA→ZH direction in all experiments, where we performed beam search decoding with a beam size of 5.

4.5.2 Evaluation Metrics

We incorporated three distinct sets of automatic evaluation metrics, with the first two aiming to evaluate translation quality and the last one aiming to assess lexical diversity.

SentencePiece BLEU We adopted the SentencePiece BLEU (spBLEU) (Goyal et al., 2021) as one of our quality evaluation metrics, since spBLEU correlated with human ratings slightly better than BLEU (Freitag et al., 2022a). We first detokenized the output from all the NMT models, then imple-

System	Methods	spBLEU	COMET
Vanilla Baseline	-	8.2	0.525
M2M-100-418M Models	multilingual training (pre-trained baseline)	22.0	0.774
	fine-tuning	22.4	0.797
	fine-tuning & pivot (without original data)	22.7	0.779
	fine-tuning & pivot (with original data)	24.6	0.808
	PTAMT	25.2	0.810
	PTAMT & 2nd fine-tuning	**26.7**	**0.828**
From-scratch Trained Models	pivot (without original data)	19.8	0.738
	pivot (with original data)	21.1	0.763
	PTAMT	23.1	0.783
	PTAMT & 2nd fine-tuning	**24.3**	**0.786**

Table 2: Translation quality automatic scores for the baseline, pre-trained models and from-scratch models. The best score per section and metric is shown in bold.

mented the pre-trained SentencePiece tokenizer [11] specific for FLORES-200 to tokenize the MT output and the reference translation, and finally computed spBLEU for each model.

Crosslingual Optimized Metric for Evaluation of Translation COMET (Rei et al., 2020) leverages cross-lingual neural language modelling and is trained to predict human judgement scores for machine-translated texts. COMET caters for a great variety of languages, and takes into account semantic similarities not only between the MT output and the reference translation but also the corresponding source text (Rei et al., 2020). We used the default COMET model[12], feeding it a triplet with detokenized source, MT output, and reference translation.

Measures of Lexical Diversity As discussed in Section 2.2, machine-translated texts exhibit differences in lexical diversity compared to original texts. Therefore, we evaluated lexical diversity in both reference translations and outputs from the NMT models in our experiments to compare the prevalence of machine translationese. Following the approach outlined by Vanmassenhove et al. (2021), lexical diversity was examined using various measures, including lexical frequency profile (LFP), type/token ratio (TTR), Yule's I and the measure of textual lexical diversity (MLTD).

In Vanmassenhove et al. (2021), LFP is used to quantify the richness of a translation by dividing the words of a text into three bands: (i) the percentage of words among the 1000 most common words in that language, (ii) the percentage of words among the next 1000 most common words, and (iii) all other words. These word frequency lists are generated from the training set. TTR assesses a text's repetitiveness by comparing the ratio of unique words (types) to the total number of words (tokens) in the text. MLTD represents the mean length of a text where a given TTR value is maintained. Yule's I, the inverse of Yule's K, measures the constancy of text and the repetitiveness of vocabulary.

Prior to computing the lexical diversity scores for each metric, we tokenized the Chinese references and MT outputs following the same Chinese pre-tokenization steps outlined in Section 4.3. Besides, we utilized the pre-tokenized mixed Chinese training sentences from *CA-ZH-MIX* to obtain the Chinese word frequency list for LFP.

[11] https://github.com/facebookresearch/flores/tree/main/flores200

[12] https://huggingface.co/Unbabel/wmt22-comet-da

System	Methods	B1	B2	B3	TTR	Yule's I	MLTD
Reference	-	0.487	0.090	0.423	0.320	14.679	218.133
Vanilla Baseline	-	0.593	0.074	0.333	0.228	3.873	50.784
M2M-100-418M Models	multilingual training (strong baseline)	0.534	0.089	0.377	0.248	6.091	115.470
	fine-tuning	0.519	0.092	0.389	0.282	8.916	132.883
	fine-tuning & pivot (without original data)	0.517	0.093	0.391	0.279	10.046	121.555
	fine-tuning & pivot (with original data)	0.514	0.094	0.393	0.283	10.257	129.594
	PTAMT	0.517	0.093	0.390	0.288	**10.770**	157.991
	PTAMT & 2nd fine-tuning	**0.515**	0.088	**0.396**	**0.295**	10.011	**167.163**
From-scratch Trained Models	pivot (without original data)	0.584	0.095	0.321	0.246	6.762	137.691
	pivot (with original data)	0.579	0.091	0.331	0.256	6.927	127.027
	PTAMT	0.556	0.093	0.351	0.272	**8.515**	**155.443**
	PTAMT & 2nd fine-tuning	**0.547**	0.090	**0.363**	**0.274**	7.728	149.375

Table 3: LFP scores with 3 bands (B1: 0-1000, B2: 1001-2000, B3: 2001-end), TTR, Yule's I and MTLD scores for the reference and the output of the NMT models in CA→ZH translation. Lower B1 values, indicating fewer matched tokens in frequent cases, along with higher values in B3, TTR, Yule's I, and MTLD, collectively indicate greater lexical richness.

5 Results and Discussion

5.1 Results

Table 2 displays the quality outcomes, while Table 3 shows the lexical diversity outcomes for all the NMT models involved in the CA→ZH translation on the FLORES-200 dataset.

Translation Quality The results indicate a notable advancement in translation quality when examining the from-scratch trained models. Particularly, transitioning from the vanilla baseline to a pivot strategy yields a significant increase in performance metrics, with spBLEU surging by 11.6 points from 8.2 to 19.8, and the COMET score enhancing by approximately 0.213 from 0.525 to 0.738. This trend of improvement extends when integrating the pivot-translated dataset with the original, which further elevates the spBLEU score by 1.3 to 21.1. This enhancement is surpassed by the PTAMT model, marking a spBLEU increase of 2.0 from 21.1 to 23.1. Interestingly, fine-tuning the PTAMT model on the small amount of original dataset led to a further spBLEU boost by 1.2.

In comparison, the M2M-100-418M models begin with a strong foundation, exhibiting a high initial spBLEU score of 22 and a COMET score of 0.774. A slight improvement in spBLEU is noted after fine-tuning on the original training set, increasing modestly to 22.4. The incremental advancement persists when pairing fine-tuning with pivot translation, further elevating the spBLEU to 22.7 when excluding the original parallel data and to 24.6 when combined with the original parallel dataset. Applying PTAMT in the fine-tuning condition boosts spBLEU further to 25.2, with a second-step fine-tuning on the original dataset resulting in a peak spBLEU score of 26.7, accompanied by the highest COMET score of 0.828.

The M2M-100-418M models generally outperform from-scratch models in terms of translation quality. However, the PTAMT-enhanced model in the from-scratch training scenario, whether with second-step fine-tuning or not, still surpasses the M2M-100-418M models reliant solely on multilingual training, fine-tuning, and fine-tuning combined with pivot translation (without original parallel data) in terms of spBLEU scores.

Lexical Diversity Compared to all the NMT models, the reference translation exhibits a lower B1 score and a higher B3 score. This reveals that the 1000 most frequent words represent a smaller proportion of the human-translated sentences, while less frequent words constitute a

larger portion of the original data compared to the outputs of different NMT systems, indicating a preference for less frequent words and a richer vocabulary. This is further evidenced by its superior TTR, Yule's I, and MLTD scores.

However, the results also reveal that incorporating low-resource training approaches into NMT models consistently leads to performance improvements over the vanilla baseline. Specifically, our proposed method, PTAMT, stands out in both from-scratch trained models and atop the M2M-100-418M pre-trained model, by achieving the lowest B1 score, the highest B3 score, and the highest scores of TTR, Yule's I, and MLTD. This suggests that the PTAMT-enhanced models excel in generating linguistically rich and varied outputs across both from-scratch training and fine-tuning scenarios. Furthermore, it was observed that the M2M-100-418M models demonstrate a superior ability to use a wider vocabulary compared to the from-scratch trained models.

5.2 Discussion

Notable improvements in translation quality and lexical diversity have been observed following the implementation of low-resource NMT techniques, underscoring the pivotal role of innovative training strategies in surpassing the limitations traditionally associated with NMT models in low-resource contexts.

While different approaches have exhibited different degrees of enhancement, the overall superiority of the M2M-100-418M models can be attributed to the extensive multilingual pretraining of the initial M2M-100-418M model, which is equipped with a broad variety of linguistic knowledge, enabling itself to benefit substantially from subsequent low-resource training strategies. Among the low-resource methods examined, PTAMT has set a new standard for generating translations, allowing the M2M-100-418M model to capitalize on both the CA–ZH and ES–ZH training datasets, achieving superior translation quality and lexical diversity compared to the other M2M-100-418M models examined. Interestingly, despite the inherent advantages of the M2M-100-418M's large-scale multilingual training base, the from-scratch trained models leveraging the PTAMT method exhibit unique capacity to optimize translation quality beyond the capabilities of the M2M-100-418M models that rely solely on the approaches of multilingual training, fine-tuning, and fine-tuning combined with pivot translation (without original parallel data).

Furthermore, PTAMT is particularly effective in reducing the impact of source-side machine-translationese introduced by the pivot-translated data (i.e. source-side machine-translated Catalan from Spanish) on the target output. PTAMT does not only include as training data a large amount of pseudo-parallel data for the desired source-target language pair (CA–ZH), but also integrates authentic linguistic input from the pivot–target language pair (ES–ZH). This approach enables the models to not only be exposed to a wider range of lexical items and usage contexts but also effectively discern and replicate the subtleties of natural language usage. Empirical evidence from our results of the lexical diversity metrics corroborates PTAMT's positive impact. The improved scores in these metrics for PTAMT-enhanced models reflect diversified word usage and a departure from the simplified and often repetitive language characteristic of synthetic data-driven translations, thereby diminishing the hallmarks of machine translationese.

Besides lexical level, we have also observed a syntax-semantics phenomenon uniquely captured by the PTAMT-enhanced models. A translation sample (see Table 7 in Appendix C) is illustrated where the CA source sentence contains three elements conveying negative meaning, whereas the ZH reference exhibits only one negative marker. This is because CA is a negative concord language, where multiple negative markers do not cancel but affirm one another to intensify the negation, and thus combine into a single negation (Espinal et al., 2016; Tubau et al., 2023). By contrast, ZH is a language without negative concord, meaning that negative markers spell out one another and thus two negatives resolve to a positive (Yang, 2011). Therefore, the triple negatives in the CA source sentence actually resolve to a single negation. When translating the the CA source sentence to ZH, only one negative needs to be retained. In our experiments, only the PTAMT models accurately captured this linguistic phenomenon, while other models erroneously included two negatives in the ZH translation, resulting in a completely opposite meaning. Surprisingly, after fine-tuning the from-scratch trained PTAMT-enhanced model on the original parallel corpus, this understanding was

lost. Conversely, the M2M-100-418M PTAMT-enhanced model gained this understanding after the second-step fine-tuning.

Finally, we have also noted the substantial impact of incorporating a small quantity of authentic parallel data in the desired language pair (dataset *CA–ZH-WIKI*). When training or fine-tuning on pivot-translated data alongside a modest amount of original CA-ZH parallel corpus, there is a marked increase in spBLEU and COMET scores, compared to using pivot-translated data alone. Moreover, fine-tuning successively the from-scratch trained PTAMT-enhanced model and the M2M-100-418M PTAMT-enhanced model on a small portion of original CA-ZH parallel data also resulted in a notable enhancement in spBLEU and COMET scores for both models. Taken together, these findings are likely to imply the significance of authentic parallel data in the target language pair(s) in improving the performance of NMT systems.

6 Conclusions and Future Work

In this work, our comprehensive experimental evaluation of from-scratch trained and M2M-100-418M pre-trained models for the CA→ZH translation task has highlighted the efficacy of low-resource NMT techniques. Significantly, these experiments have confirmed the substantial benefits of these methods on translation quality and lexical diversity, with our novel PTAMT method emerging as a key innovator in addressing the challenges inherent in translating the low-resource language pair.

The PTAMT method, with its ability to effectively utilize pseudo-parallel and authentic parallel data, significantly mitigates the influence of source-side machine translationese and enhances the model's capability to produce translations that are not only accurate but also linguistically rich and varied. This approach not only broadens the lexical range and usage contexts available to the model but also ensures a nuanced understanding and replication of natural language subtleties, as evidenced by the improved lexical diversity metrics and the accurate handling of complex linguistic phenomena such as negative concord.

Moreover, our findings seem to imply the critical role of integrating authentic data in the desired language pair(s) into the training or fine-tuning process, demonstrating that even a small amount of authentic parallel data can substantially elevate the performance of NMT systems. This insight emphasizes the importance of combining pseudo-parallel and authentic inputs to achieve the best possible translation outcomes, particularly in the context of low-resource language pairs.

While our study marks progress in NMT for the low-resource CA–ZH pair, it also unveils areas which deserve further exploration. The potential domain alignment between our test set *FLORES-200* and the original *CA-ZH-WIKI* parallel data raises questions about how the inclusion of a modest amount of authentic parallel data in the target language pair(s) influences translation outcomes. This becomes especially relevant when considering the potential for domain-specific biases to affect the evaluation of NMT systems. Moreover, the observed discrepancies in how from-scratch trained PTAMT-enhanced models and M2M-100-418M PTAMT-enhanced models handle linguistic complexities such as negative concord—both before and after additional fine-tuning—suggest underlying differences in model learning dynamics that deserve closer scrutiny. These divergent model responses highlight the need for a nuanced understanding of how different training approaches impact the NMT models' ability to grasp and accurately render complex linguistic structures.

To navigate these uncertainties and expand upon our findings, we propose several avenues for future research: firstly, building novel and diversified test sets to quantify and generalize the influence of authentic parallel data in the target language pair(s) on model performance; secondly, exploring the models' internal representations and additional fine-tuning processes to pinpoint factors contributing to their distinct responses to linguistic complexities such as negative concord; thirdly, expanding our investigation to include more low-resource language pairs to enable a comprehensive evaluation of the PTAMT method's applicability across diverse linguistic contexts.

7 Acknowledgements

This work was partly funded by the China Scholarship Council (CSC), to whom we express our sincere gratitude. We are also grateful to the anonymous reviewers of EAMT 2024, whose insightful comments significantly enhanced the quality and presentation of this paper.

References

Adelani, David, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen Muhammad, Guyo Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ajibade, Tunde Ajayi, Yvonne Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. 2022. A few thousand translations go a long way! leveraging pre-trained models for African news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070, Seattle, United States. Association for Computational Linguistics.

Araabi, Ali and Christof Monz. 2020. Optimizing transformer for low-resource neural machine translation.

Arivazhagan, Naveen, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges.

Chae, Jieun and Ani Nenkova. 2009. Predicting the fluency of text with shallow structural features: Case studies of machine translation and human-written text. In Lascarides, Alex, Claire Gardent, and Joakim Nivre, editors, *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 139–147, Athens, Greece, March. Association for Computational Linguistics.

Chen, Yun, Yang Liu, Yong Cheng, and Victor O. K. Li. 2017. A teacher-student framework for zero-resource neural machine translation.

Cohn, Trevor and Mirella Lapata. 2007. Machine translation by triangulation: Making effective use of multi-parallel corpora. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 728–735, Prague, Czech Republic. Association for Computational Linguistics.

Costa-jussà, Marta Ruiz, José A. R. Fonollosa, José B. Mariño, Marc Poch, and Mireia Farrús. 2014. A large Spanish-Catalan parallel corpus release for machine translation. *Comput. Informatics*, 33:907–920.

Costa-Jussà, Marta R., Noé Casas, Carlos Escolano, and José A. R. Fonollosa. 2019. Chinese-catalan: A neural machine translation approach based on pivoting and attention mechanisms. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 18(4).

Currey, Anna and Kenneth Heafield. 2019. Zero-resource neural machine translation with monolingual pivot data. In Birch, Alexandra, Andrew Finch, Hiroaki Hayashi, Ioannis Konstas, Thang Luong, Graham Neubig, Yusuke Oda, and Katsuhito Sudoh, editors, *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 99–107, Hong Kong, November. Association for Computational Linguistics.

Dutta Chowdhury, Koel, Rricha Jalota, Cristina España-Bonet, and Josef Genabith. 2022. Towards debiasing translation artifacts. In Carpuat, Marine, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3983–3991, Seattle, United States, July. Association for Computational Linguistics.

El Kholy, Ahmed, Nizar Habash, Gregor Leusch, Evgeny Matusov, and Hassan Sawaf. 2013. Language independent connectivity strength features for phrase pivot statistical machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 412–418, Sofia, Bulgaria. Association for Computational Linguistics.

El-Kishky, Ahmed, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969, Online. Association for Computational Linguistics.

Espinal, M. Teresa, Susagna Tubau, Joan Borràs-Comes, and Pilar Prieto, 2016. *Double Negation in Catalan and Spanish. Interaction Between Syntax and Prosody*, pages 145–176. Springer International Publishing, Cham.

Fan, Angela, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation.

Firat, Orhan, KyungHyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. *CoRR*, abs/1601.01073.

Freitag, Markus, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022a. Results of WMT22 metrics shared

task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid), December. Association for Computational Linguistics.

Freitag, Markus, David Vilar, David Grangier, Colin Cherry, and George Foster. 2022b. A natural diet: Towards improving naturalness of machine translation output. In Muresan, Smaranda, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3340–3353, Dublin, Ireland, May. Association for Computational Linguistics.

Goyal, Naman, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzman, and Angela Fan. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation.

Gu, Jiatao, Hany Hassan, Jacob Devlin, and Victor O.K. 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, New Orleans, Louisiana. Association for Computational Linguistics.

Ha, Thanh-Le, Jan Niehues, and Alexander Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder.

Hassan Awadalla, Hany, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, Will Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic chinese to english news translation.

Ilisei, Iustina, Diana Inkpen, Gloria Corpas Pastor, and Ruslan Mitkov. 2010. Identification of translationese: A machine learning approach. In Gelbukh, Alexander, editor, *Computational Linguistics and Intelligent Text Processing*, pages 503–511, Berlin, Heidelberg. Springer Berlin Heidelberg.

Jalota, Rricha, Koel Chowdhury, Cristina España-Bonet, and Josef van Genabith. 2023. Translating away translationese without parallel data. In Bouamor, Houda, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7086–7100, Singapore, December. Association for Computational Linguistics.

Johnson, Melvin, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation.

Kreutzer, Julia, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoî t Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Lakew, Surafel Melaku, Quintino Francesco Lotito, Matteo Negri, Marco Turchi, and Marcello Federico. 2018. Improving zero-shot translation of low-resource languages. *ArXiv*, abs/1811.01389.

Liu, Zixuan. 2022. Improving chinese-catalan machine translation with wikipedia parallel corpus. Master's thesis, Universitat Pompeu Fabra, Barcelona.

Lu, Yichao, Phillip Keung, Faisal Ladhak, Vikas Bhardwaj, Shaonan Zhang, and Jason Sun. 2018. A neural interlingua for multilingual machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 84–92, Brussels, Belgium. Association for Computational Linguistics.

Neubig, Graham 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, Brussels, Belgium. Association for Computational Linguistics.

Popel, Martin, Markéta Tomková, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdenek Zabokrtsky. 2020. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature Communications*, 11:4381.

Rapp, Reinhard. 2021. Similar language translation for Catalan, Portuguese and Spanish using Marian NMT. In *Proceedings of the Sixth Conference on Machine Translation*, pages 292–298, Online. Association for Computational Linguistics.

Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online, November. Association for Computational Linguistics.

Riley, Parker, Isaac Caswell, Markus Freitag, and David Grangier. 2020. Translationese as a language in "multilingual" nmt.

Schwenk, Holger, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia.

Schwenk, Holger, Guillaume Wenzek, Sergey Edunov, Edouard Grave, and Armand Joulin. 2020. Ccmatrix: Mining billions of high-quality parallel sentences on the web.

Tan, Xu, Yi Ren, Di He, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2019. Multilingual neural machine translation with knowledge distillation.

Team, NLLB, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Thillainathan, Sarubi, Surangika Ranathunga, and Sanath Jayasena. 2021. Fine-tuning self-supervised multilingual sequence-to-sequence models for extremely low-resource nmt. In *2021 Moratuwa Engineering Research Conference (MERCon)*, pages 432–437.

Tubau, Susagna, Urtzo Etxeberria, and M. Teresa Espinal. 2023. A new approach to negative concord: Catalan as a case in point. *Journal of Linguistics*, page 1–33.

Vanmassenhove, Eva, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In Merlo, Paola, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2203–2213, Online, April. Association for Computational Linguistics.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Wang, Xinyi, Yulia Tsvetkov, and Graham Neubig. 2020. Balancing training for multilingual neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8526–8537, Online. Association for Computational Linguistics.

Wang, Rui, Xu Tan, Renqian Luo, Tao Qin, and Tie-Yan Liu. 2021. A survey on low-resource neural machine translation.

Wein, Shira and Nathan Schneider. 2024. Lost in translationese? reducing translation effect using abstract meaning representation.

Wu, Hua and Haifeng Wang. 2007. Pivot language approach for phrase-based statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 856–863, Prague, Czech Republic. Association for Computational Linguistics.

Yang, Huiling. 2011. Is Chinese a negative concord language? In *proceedings of the 23rd North American Conference on Chinese Linguistics (NACCL-23)*, pages 208–223, Eugene, United States), December. University of Oregon Press.

Zheng, Hao, Yong Cheng, and Yang Liu. 2017. Maximum expected likelihood estimation for zero-resource neural machine translation. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, page 4251–4257. AAAI Press.

Zhou, Chenuye. 2022. Building a catalan-chinese parallel corpus from Wikipedia for use in machine translation. Master's thesis, Universitat Pompeu Fabra, Barcelona.

Ziemski, Michał, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia. European Language Resources Association (ELRA).

Zoph, Barret, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation.

A Appendix: Hyper-parameters Configuration

Hyper-parameters	Value
adam betas	0.9, 0.98
learning rate	0.0005
warmup initial learning rate	1.00E-07
label smoothing	0.1
dorpout	0.2
weight decay	0.0001
batch size (in tokens)	4096 ($\times$ 8 steps)
gradients accumulation	8

Table 4: hyper-parameters of the neural models. Note that the same set of hyper-parameters was used for all experiments except that the batch size for the M2M-100-418M models was 2048 tokens ($\times$ 16 steps) due to GPU memory limit.

Hyperparameter	Value
adam betas	0.9, 0.98
learning rate	0.0005
warmup initial learning rate	1.00E-07
label smoothing	0.5
dorpout	0.3
activation dropout	0.3
enc/dec layerDrop	0.0/0.1
weight decay	0.0001
batch size (in tokens)	4096 ($\times$ 8 steps)
gradients accumulation	8

Table 5: Optimal Transformer hyper-parameters settings for 40k datasets.

B Appendix: Training and Fine-tuning Epochs

System	Methods	Epoch
Vanilla Baseline	-	155
M2M-100-418M Models	multilingual training (strong baseline)	-
	fine-tuning	8
	fine-tuning & pivot (without original data)	3
	fine-tuning & pivot (with original data)	4
	PTAMT	4
	PTAMT & 2nd fine-tunig	10
From-scratch Trained Models	pivot (without original data)	13
	pivot (with original data)	11
	PTAMT	13
	PTAMT & 2nd fine-tuning	10

Table 6: Training and fine-tuning epochs for the baseline, pre-trained models and from-scratch models. Note that the epoch of the pre-trained M2M-100-418M model was not reported publicly.

C Appendix: Translation Sample

Model	Sentence
Source(CA)	Adverteix que no hi ha ningú que pugui garantir que cap acció a l'Iraq en aquest moment aconsegueixi aturar la guerra sectària, la violència creixent o una deriva caòtica.
Reference(ZH)	报告警告称,没有人能保证目前在伊拉克采取的任何行动 能够阻止宗派战争、不断增长的暴力或走向混乱。
Meaning in English	It warns no one can guarantee that any action in Iraq at this point will stop sectarian warfare, growing violence, or a slide toward chaos.
Vanilla baseline	亚当斯表示，没有任何人确保伊拉克的行动，并阻止 伊拉克战争、暴力行为、暴力行为或暴力行为。
Meaning in English	Adam shows, no one guarantees the action in Iraq and stops the war in Iraq, violence, violence, or violence.
M2M-100-418M (Strong baseline)	他警告说,没有人能保证伊拉克目前没有任何行动 能阻止种族战争、暴力加剧或混乱的流动。
Meaning in English	He warns no one can guarantee that no action in Iraq at this point will stop ethnic warfare, growing violence, or the flow of chaos.
M2M-100-418M + finetuning	他警告,没有人能保证伊拉克在这一时刻不会采取任何行动 来阻止种族战争、日益暴力或混乱的发生。
Meaning in English	He warns no one can guarantee that Iraq is not taking any action at this point to stop ethnic warfare, growing violence, or the occurrence of chaos.
M2M-100-418M + fine-tuning & pivot (without original data)	他警告说,没有人能够保证目前在伊拉克采取的任何行动 都不会停止派别战争、暴力升级或混乱。
Meaning in English	He warns no one can guarantee that any action in Iraq at this point will not stop sectarian warfare, growing violence, or chaos.
M2M-100-418M + fine-tuning & pivot (with original data)	他警告说,没有人能够保证目前在伊拉克采取的任何行动 都不会阻止派别战争、暴力升级或混乱。
Meaning in English	He warns no one can guarantee that any action in Iraq at this point will not stop sectarian warfare, growing violence, or chaos.
M2M-100-418M + PTAMT	他警告说,没有人能够保证,目前在伊拉克采取的任何行动 都不会成功地制止教派战争、不断升级的暴力或混乱的倾向。
Meaning in English	He warns no one can guarantee that any action in Iraq at this point will not successfully stop sectarian warfare, growing violence, or tendency towards chaos.

Model	Sentence
M2M-100-418M + PTAMT & 2nd-fine-tuning	他警告说,没有人能够保证在伊拉克的任何行动 能阻止教派战争、不断升级的暴力或混乱的倾向。
Meaning in English	He warns no one can guarantee that any action in Iraq will stop sectarian warfare, growing violence, or tendency towards chaos.
From-scratch + pivot translation (without original data)	他指出,没有人能够保证伊拉克目前的任何行动 都不会导致教派战争、日益严重的暴力或混乱。
Meaning in English	He points out no one can guarantee that any action in Iraq at this point will not lead to sectarian warfare, growing violence, or chaos.
From-scratch + pivot translation (with original data)	他警告说,没有人能够保证,伊拉克目前的任何行动 都不会阻止教派战争、不断升级的暴力或混乱的漂流。
Meaning in English	He warns no one can guarantee that any action in Iraq at this point will not stop sectarian warfare, growing violence, or the flow of chaos.
From-scratch + PTAMT	他警告说,没有人能够确保目前在伊拉克的任何行动 能够制止教派战争、不断升级的暴力或混乱。
Meaning in English	He warns no one can guarantee that any action in Iraq at this point will stop sectarian warfare, growing violence, or chaos.
From-scratch + PTAMT & 2nd-fine-tuning	他警告说,没有能保证此时在伊拉克的任何行动 都不会阻止教派战争、不断升级的暴力或混乱的漂流。
Meaning in English	He warns no one can guarantee any action in Iraq at this point will not stop sectarian warfare, growing violence, or the flow of chaos.

Table 7: Translation sample for baseline, fine-tuned, and from-scratch trained models. Note that the underlined elements in the table are words or structural elements that cause negation.