The TALP-UPC Participation in WMT21 News Translation Task: an mBART-based NMT Approach

Carlos Escolano[1], **Ioannis Tsiamas**[1], **Christine Basta**[1 2], **Javier Ferrando**[1], **Marta R. Costa-jussà**[1], **José A. R. Fonollosa**[1]
[1] TALP Research Center, Universitat Politècnica de Catalunya, Barcelona
[2] Institute of Graduate Studies and Research, Alexandria University, Egypt
{carlos.escolano,ioannis.tsiamas,christine.basta
javier.ferrando.monsonis,marta.ruiz,jose.fonollosa}@upc.edu

Abstract

This paper describes the submission to the WMT 2021 news translation shared task by the UPC Machine Translation group. The goal of the task is to translate German to French (De-Fr) and French to German (Fr-De). Our submission focuses on fine-tuning a pre-trained model to take advantage of monolingual data. We fine-tune mBART50 using the filtered data, and additionally, we train a Transformer model on the same data from scratch. In the experiments, we show that fine-tuning mBART50 results in 31.69 BLEU for De-Fr and 23.63 BLEU for Fr-De, which increases 2.71 and 1.90 BLEU accordingly, as compared to the model we train from scratch. Our final submission is an ensemble of these two models, further increasing 0.3 BLEU for Fr-De.

1 Introduction

Monolingual data is usually more abundant than parallel data as it does not need any human processing. Neural Machine Translation (NMT) has focused traditionally on parallel data between languages and monolingual data as back-translation (Sennrich et al., 2016). This method consists of translating a monolingual corpus with an NMT system and training the model using the synthetic data. Alternatively, in recent years, pre-trained models have been proposed using monolingual data as a pre-training step with self-supervised learning, before performing task-specific fine-tuning. An example of this approach is BERT (Devlin et al., 2019), which is a Transformer model (Vaswani et al., 2017), pre-trained on masked language modeling and next sentences prediction on a large unlabeled corpus. While BERT is used primarily for classification tasks, BART (Lewis et al., 2020) has been proposed for sequence-to-sequence tasks. BART is a Transformer encoder-decoder, pre-trained as a Denoising Autoencoder (DAE) on monolingual unlabeled text. Since BART is pre-trained on monolingual data, an additional encoder should be introduced during fine-tuning to obtain a bilingual NMT system. mBART overcomes this restriction by being pre-trained on multilingual denoising. mBART (Liu et al., 2020; Tang et al., 2020) is liable to fine-tuning on several translation directions in order to obtain a multilingual NMT system.

Our participation to the news translation task at WMT focuses on translating between German (De) and French (Fr) in both directions, De-Fr and Fr-De. To accomplish this, we employ a pre-trained mBART model, and more specifically mBART50 (Tang et al., 2020), which is pre-trained with 50 languages. We fine-tune the mBART50 on both translation directions to obtain a single multilingual model for the task. To measure the importance of the pre-training step, we additionally train a Transformer with the same architecture and hyperparameters but randomly initialized. Our experiments show that the fine-tuned mBART50 can achieve better translation quality in both directions, with improvements of 2.71 for De-Fr and 1.9 for Fr-De. Apart from fine-tuning a pre-trained model, our approach also includes extensive filtering of a large bilingual corpus to ensure high-quality training data. Finally, we have considered ensembling the fine-tuned mBART50 and the trained-from-scratch Transformer for our submission. This ensembling has resulted in BLEU scores of 31.69 and 23.93 for De-Fr and Fr-De accordingly.

The rest of this paper is organized as follows: In Section 2 we describe the background techniques this work builds upon, multilingual NMT and mBART. In Section 3 we present the datasets we used for training and the techniques applied for filtering them. In Section 4 we provide the system description along with the implementation details and Section 5 involves the results of our experiments. Finally, in Section 6 we discuss the conclusions of this work and present possible directions for further research.

Proceedings of the Sixth Conference on Machine Translation (WMT), pages 117–122
November 10–11, 2021.

2 Background

2.1 Neural Machine Translation

Neural Machine Translation (NMT) uses sequence-to-sequence models, with an encoder-decoder architecture, built upon deep neural networks (Sutskever et al., 2014; Bahdanau et al., 2014; Gehring et al., 2017; Vaswani et al., 2017). In a sequence-to-sequence model, the source sentence is mapped to its contextualized representation and fed to the decoder to generate the translation output in an auto-regressive way. Traditionally, recurrent neural networks (Hochreiter and Schmidhuber, 1997) have been used for the encoder and decoder, with an attention mechanism (Bahdanau et al., 2014) that enables each target token to concentrate on certain tokens in the source sentence. Recently, the Transformer (Vaswani et al., 2017) led to large improvements on sequence-to-sequence tasks and NMT, relying exclusively on attention mechanisms. The systems trained in this work, are also based on Transformer models.

2.2 Multilingual NMT

Multilingual NMT aims to provide a single model that can translate several language directions (Firat et al., 2016; Johnson et al., 2017). These can be one-to-many, many-to-one, or many-to-many, with the "one" being usually English due to the broadly available corpora. Previous studies have explored different design approaches, focusing on sharing parts of the model between the different languages, with shared encoder-decoder attention between languages (Firat et al., 2016), a shared encoder (Sen et al., 2019), a task-specific attention (Blackwood et al., 2018), shared parameters (Zhu et al., 2020) and full model sharing (Johnson et al., 2017). Recently, the paradigm of full model sharing has been extended to accommodate for many more languages and directions by training huge models for massively multilingual NMT (Arivazhagan et al., 2019; Fan et al., 2020). Our submission is also based on multilingual models that are fully shared between the two language directions German-French and French-German.

2.3 mBART

BART (Lewis et al., 2020) is a Transformer encoder-decoder, which is pre-trained with self-supervised learning on reconstructing the text corrupted by a noise function. Its multilingual version, mBART (Liu et al., 2020), uses the same self-supervised approach, but reconstructs corrupted text from multiple monolingual corpora. The nature of this pre-training makes mBART a good initialization for a multilingual NMT system. mBART can be fine-tuned on multiple bitext corpora providing gains in all directions of 25 languages, except for the very highest resource ones (Liu et al., 2020). Our system is initialized with mBART50 (Tang et al., 2020) (an extension of mBART from 25 to 50 languages). This initialization is followed by multilingual fine-tuning on both directions of a large German-French bitext.

3 Data

In this section we introduce the datasets used for training our systems and we go through the data filtering process that was applied in each one of them.

3.1 Datasets

In order to train Transformer models for Machine Translation, commonly, a large parallel corpus is needed. For the purpose of this research, we focus on creating a French-German parallel corpus from several publicly available datasets. More specifically, these are the Europarl (Koehn, 2005), Paracrawl (Bañón et al., 2020), Common Crawl[1], News Commentary (Tiedemann, 2012), Wiki Titles[2], Tilde Rapid and EESC (Rozis and Skadiņš, 2017), WikiMatrix[3] and TED Talks (Cettolo et al., 2012). If the dataset contains more languages, we only keep examples that have non-empty sentences for French-German. We use the development and test data of the news test datasets of 2019 and 2020, as provided by WMT. The size of each dataset can be found in the first column of Table 1.

3.2 Data Filtering

We employ two stages of data filtering to ensure that our system is trained on high-quality data. In the first stage, we process each example, either from the French or German side, separately by altering its content. This process includes the following steps in the listed order:

1. Removal of non-utf8 characters
2. De-escaping html characters

[1] http://data.statmt.org/wmt19/translation-task/fr-de/bitexts

[2] http://data.statmt.org/wikititles/v3/

[3] https://github.com/facebookresearch/LASER/tree/master/tasks/WikiMatrix

3. Normalization of different types of punctuation

4. Normalization of spacing

5. Removal of redundant apostrophes

The normalization of spacing and punctuation is applied using the SacreMoses[4] package. During the second stage of filtering, we completely remove whole examples, when either the French or the German sentence contain noise or inconsistent information.

1. **Basic Filtering.** We remove examples where either side is empty, or the two sides contain the same lower-cased text.

2. **Language Filtering.** Here we intend to identify sentences that are written in languages other than French and German. We remove examples where the language of either sentence does not match the expected one. To predict the language of a sentence, we use a pre-trained language detection model from FastText (Joulin et al., 2016).

3. **Length Filtering.** Here we aim to identify sentences or examples with unnatural length characteristics that potentially result from noise. We remove examples where either side is found to have a large number of words (greater than 200), an extreme character-to-word ratio (lower than 1.5 or greater than 12), at least one word with a high number of characters (greater than 25), or the example has an extreme source-to-target word ratio (lower than 0.4 or greater than 2.5). For setting the boundaries of what is an acceptable length or ratio, we follow (Shi et al., 2020).

4. **Alignment Filtering.** At this point, we want to identify noisy pairs by computing their alignment scores. We use the fast-align library (Dyer et al., 2013) and remove examples where the alignment score is 2.5 times above the average alignment score of the corpus. The alignment score of an example is calculated as the normalized log-probability of the German-French alignment, and the alignment score of the corpus is the sum of the alignment scores of its examples. Specifically for the WikiMatrix pairs, which are not human-generated and possibly contain more noise, we follow a more aggressive approach and remove a pair if its alignment score is 15 absolute points above the average.

[4] https://github.com/alvations/sacremoses

Dataset	Size (thousands)	
	Original	Filtered
Europarl	1,803	1,480
Paracrawl	7,223	5,893
Common Crawl	622	523
News Commentary	304	236
Wiki Titles	1,007	134
Title Rapid	983	849
EESC	2,844	2,392
WikiMatrix	2,807	1,936
TED Talks	292	279
Total	**17,885**	**13,722**

Table 1: Training sets before and after filtering.

The size of the clean corpus can be found in the second column of Table 1.

4 System Description

In this section we are going to describe the two main steps of our submission, fine-tuning of a pre-trained with the provided data and ensemble of pre-trained and not pre-trained models.

4.1 Multilingual fine-tuning

Traditionally, models are trained from random initialization. We initialize our model with mBART50 (Tang et al., 2020) pre-trained weights. These weights act as a more informed initialization that already contains useful features for language representation. Given the support of the public model to the French and German languages, no modifications of the embedding model were needed. Following mBART50 strategy, we fine-tune all the layers of the multilingual model on all the filtered French-German and German-French data. To condition the language generation (Johnson et al., 2017), a source language token was added as the first token of the source sentence, and a target language token was added as the first decoder token to the decoder.

Implementation Both randomly initialized baseline and mBART50 models were trained using *fairseq*'s (Ott et al., 2019)[5] mBART large imple-

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

mentation for multilingual fine-tuning. The architecture consists of 12 layers with 1024 embedding size, 4096 feed-forward size, and 16 attention heads, both for encoder and decoder, with a total number of parameters of 610, 878, 464. Models were trained for approximately 400k updates or seven epochs using validation loss as an early stopping criterion, with a learning rate of $1 * 10^{-4}$ and $3 * 10^{-5}$ for the baseline and fine-tuned model, respectively. Both models are trained using the original vocabulary of 250k tokens, at subword level using the *sentencepiece*[6] model available with the mBART50 model. Both models were trained on two Nvidia GTX 3090 with eight batches of gradient accumulation. At generation time, beam size was set to eight.

4.2 Model ensemble

Model ensembling is a popular technique to leverage the features learned by several models. This is especially important in our case as the pre-trained model has been trained on data not constrained to the domain. As the pre-training step is performed on data that does not belong to the task domain, it could generate structures or patterns that are not commonly used, even when keeping the sentence's intended meaning. In order to balance the provided data, we ensemble the best checkpoint from the baseline and the fine-tuned mBART. Thus, next token prediction during inference is done according to the combined probability from both models.

Implementation Ensembling was performed using the standard *fairseq* generation script. The two models ensembled were the same checkpoints at 400k updates reported for the individual systems. Beam size was set to eight as in the previous experiments.

5 Experiments and Results

Multilingual fine-tuning. Our first hypothesis is that the use of pre-trained models could improve translation performance. Therefore, we compare our system with a baseline system with the same vocabulary and parameter configuration with random initialization to measure the impact of the translation step. Tables 2 and 3 show the translation results for German-French and French-German translation directions, respectively. Results show that fine-tuning of pre-trained models improves translation quality in both directions by 2.3 BLEU points on average, showing that a more informed model initialization significantly impacts the final model performance. It is worth noticing that we expected the fine-tuning approach to converge faster than the randomly initialized baseline, but they both show similar behaviors and required approximately 400k training updates. Figure 1 show that the fine-tuned model's validation loss is lower over the entire training but both converge to the best value at the seventh epoch.

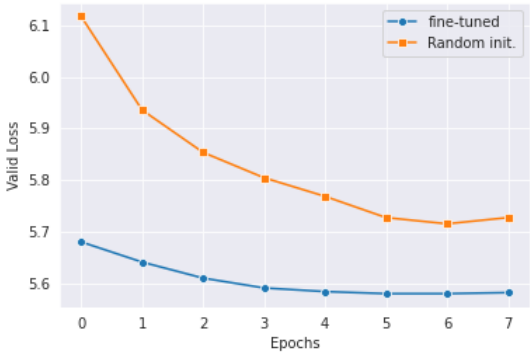

Figure 1: Validation loss during training for the fine-tuned mBART (fine-tuned) and randomly initialized (Random init.) models.

	BLEU	**ΔBLEU**
Baseline	28.98	-
mBART50	**31.69**	2.71
+Ensemble	**31.69**	0.00

Table 2: Results measured in BLEU for the German to French translation direction

Model Ensemble. Our second hypothesis is that the pre-training step on the out-of-domain data may affect the model's phrasing at inference time, and ensembling with the baseline trained only on the provided constrained data could improve its performance. Results show that, although a minor improvement of 0.3 BLEU points has been reported for the French to German translation direction, it is not consistent on German to French, where no performance difference has been observed. These results may indicate that the pre-training step has a limited impact on the final domain performance and that the fine-tuning step on the provided constraint data is the most crucial factor in the final model's domain adaptation.

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

	BLEU	ΔBLEU
Baseline	21.73	-
mBART50	23.63	1.90
+Ensemble	**23.93**	0.30

Table 3: Results measured in BLEU for the French to German translation direction

6 Conclusions

This work describes the TALP-UPC system for the WMT 2021 shared news translation task for French-German and German-French. Experimental results show that pre-trained models help improve translation performance in this kind of scenario, even for high-resource language pairs with millions of parallel sentences available, with 2.71 points for German-French translation direction and 1.90 points for French-German. Results also show that ensembling of pre-trained and randomly initialized models can lead to minor performance improvements (up to 0.3 BLEU) but not consistently on both tested languages.

In future work, better results may be obtained by combining fine-tuned pre-trained models with traditional back translation. Both techniques would benefit from the additional monolingual in two different aspects of the NMT model, initialization and additional training on the monolingual data provided.

Acknowledgments

This work is supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 947657).

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*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Graeme Blackwood, Miguel Ballesteros, and Todd Ward. 2018. Multilingual neural machine translation with task-specific attention. *ArXiv*, abs/1806.03280.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Annual conference of the European Association for Machine Translation*, pages 261–268, Trento, Italy. European Association for Machine Translation.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

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

Orhan Firat, Baskaran Sankaran, Yaser Al-onaizan, Fatos T. Yarman Vural, and Kyunghyun Cho. 2016. Zero-resource translation with multi-lingual neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 268–277, Austin, Texas. Association for Computational Linguistics.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *International Conference on Machine Learning*, pages 1243–1252. PMLR.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Trans. Assoc. Comput. Linguistics*, 5:339–351.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

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

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

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 NAACL-HLT 2019: Demonstrations*.

Roberts Rozis and Raivis Skadiņš. 2017. Tilde MODEL - multilingual open data for EU languages. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 263–265, Gothenburg, Sweden. Association for Computational Linguistics.

Sukanta Sen, Kamal Kumar Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Multilingual unsupervised NMT using shared encoder and language-specific decoders. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3083–3089, Florence, Italy. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Tingxun Shi, Shiyu Zhao, Xiaopu Li, Xiaoxue Wang, Qian Zhang, Di Ai, Dawei Dang, Xue Zhengshan, and Jie Hao. 2020. OPPO's machine translation systems for WMT20. In *Proceedings of the Fifth Conference on Machine Translation*, pages 282–292, Online. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

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.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Changfeng Zhu, Heng Yu, Shanbo Cheng, and Weihua Luo. 2020. Language-aware interlingua for multilingual neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1650–1655, Online. Association for Computational Linguistics.