TenTrans Multilingual Low-Resource Translation System for WMT21 Indo-European Languages Task

Han Yang[1,2]†, Bojie Hu[2], Wanying Xie[2], Ambyer Han[2], Pan Liu[2], Jinan Xu[1]*, Qi Ju[2]*
[1]Beijing Jiaotong University
[2]TencentMT Oteam
{19120421, jaxu}@bjtu.edu.cn
{bojiehu, wanyingxie, ambyera, galeliu, damonju}@tencent.com

Abstract

This paper describes TenTrans' submission to WMT21 Multilingual Low-Resource Translation shared task for the Romance language pairs. This task focuses on improving translation quality from Catalan to Occitan, Romanian and Italian, with the assistance of related high-resource languages. We mainly utilize back-translation, pivot-based methods, multilingual models, pre-trained model fine-tuning, and in-domain knowledge transfer to improve the translation quality. On the test set, our best-submitted system achieves an average of 43.45 case-sensitive BLEU scores across all low-resource pairs. Our data, code, and pre-trained models used in this work are available in TenTrans evaluation examples[1].

1 Introduction

We participate in the WMT21 Multilingual Low-Resource Translation shared task. This task focuses on the multilinguality in the cultural heritage domain for two Indo-European language families: North-Germanic and Romance. We devote the research into translations among Romance languages, including Catalan→Occitan, Catalan→Romanian, Catalan→Italian. Additionally, this task explicitly encourages the use of data of four related high-resource languages (Spanish, French, Portuguese and English) in the same linguistic family.

For the model architecture, we adopt a universal encoder-decoder architecture that shares parameters across all languages (Johnson et al., 2017). And almost all of the subsequent experiments are based on Transformer *base* model (Vaswani et al., 2017).

To effectively exploit low and high resource data in the multilingual low-resource scenario, we explore several approaches, and each approach shows effectiveness. We employ back-translation (Sennrich et al., 2016a) and pivot-based methods to augment the training corpus. In terms of knowledge transfer, we explore the pre-trained model and the multilingual model that trained with both low and high resource language pairs. Moreover, we extract in-domain corpus by a domain classifier and adapt the model to the target domain by in-domain fine-tuning.

This paper is structured as follows: Section 2 introduces used datasets, data statistic and pre-processing pipeline. Section 3 describes the details of different approaches. In Section 4 we present experimental settings and results. Section 5 draws a brief conclusion of our work in the WMT21.

2 Data

2.1 Datasets

The training datasets are majorly provided by the publicly available OPUS (Tiedemann, 2012) repository. We use almost all available datasets provided in the task, including Europarl, JW300, WikiMatrix, MultiCCAligned, OPUS-100, Bible, ELRC, and 167.2K It-Ro pairs in TED talks as well as 15M/360K sentence pairs of En-It/En-Ro extracted from Wikipedia dumps. For datasets that can be found through the resources search form on the top-level website of OPUS, we use opus-tools[2] to extract low-resource language pairs. As for rest of the data, we download them in the usual way. Statistics of different datasets are showed in Table 1.

2.2 Data Pre-processing

Cleaning datasets is necessary when the datasets are noisy and of low quality. We partially refer to

†This work is done by the author as an intern at TencentMT Oteam.
*Corresponding author.
[1]https://github.com/TenTrans/TenTrans/blob/master/example/WMT21/WMT21-low-resource-MNMT.md
[2]https://pypi.org/project/opustools-pkg/

Proceedings of the Sixth Conference on Machine Translation (WMT), pages 376–382
November 10–11, 2021.

Bilingual	WikiMatrix	MultiCCAligned	Bible/Europarl	JW300	ELRC	OPUS
LRL-LRL	2.7M	11.5M	386.5K	1.2M	0.022K	-
HRL-LRL	8.9M	68.7M	5.9M	12.2M	2.7M	2.0M
Monolingual	**WikiMatrix**	**MultiCCAligned**	**Bible/Europarl**	**JW300**	**ELRC**	**OPUS**
Oc	342.3K	-	-	-	-	35.8K
Ro	3.8M	132.6M	470.1K	54.5M	1.1M	1M
It	9.1M	175.2M	23.3M	66.2M	1.6M	~~1M~~

Table 1: Number of sentences in different datasets. 'LRL-LRL' means the bilingual data between low resource languages, e.g. Ca-Ro. 'HRL-LRL' means the bilingual data between high-low resource languages, e.g. En-Ro. '~~1M~~' means we do not use that data though it is provided. Note that OPUS provides En-Oc/Ro/It bilingual pairs, but we also use the target side Oc/Ro as monolingual data due to lacking data.

	Ca-Oc	Ca-Ro	Ca-It
No filter	138.7K	2.2M	6.3M
Filtered	138.7K	2.1M	5.8M
	It-Ro	**It-Oc**	**Oc-Ro**
No filter	7.2M	122K	81K
Filtered	6.9M	122K	81K

Table 2: Number of sentences in low-resource bilingual data.

M2M-100[3] (Fan et al., 2020) data pre-processing procedures to filter bilingual sentences. We remove sentences with more than 50% punctuation, deduplicate training data and remove all instances of evaluation data from the bilingual training data.

We tokenize all data and normalize punctuation with the Moses tokenizer (Koehn et al., 2007). To enable open-vocabulary and share information among languages, we use joint Byte-Pair-Encoding (BPE) with 32K split operations for subword segmentation (Sennrich et al., 2016b). We also remove sentences longer than 512 as well as sentence pairs with a source/target length ratio exceeding 3.

For monolingual data, we still employ those rules except the length ratio filter. See Table 2 for the statistics of low-resource bilingual data, Table 3 for the statistics of high-low resource bilingual data and Table 4 for the statistics of low-resource monolingual data.

3 System Overview

3.1 Base Systems

In multilingual translation scenarios, one can employ multi-task learning framework using multiple encoders or multiple decoders (Luong et al., 2016; Dong et al., 2015; Firat et al., 2016). Either, one can employ a unified model consisting of a shared encoder and a shared decoder for all the language pairs (Johnson et al., 2017). We experiment with these two models and conduct the conclusion that a universal encoder-decoder model outperforms the model with multiple decoders. The unified architecture is adopt in subsequent experiments in this work. Parameters and vocabulary are shared among all language pairs and this helps the generalization across languages improving the translation for the low-resource language pairs (Aharoni et al., 2019). We also train three separate bilingual models to be regarded as contrastive model with multilingual model. Furthermore, we jointly train on Catalan, Occitan, Romanian, Italian four low-resource languages simultaneously to obtain a many-to-many multilingual model. Detailed results of base systems are shown in Table 6.

We use the Transformer (Vaswani et al., 2017) as our model architecture for all of our systems. We experiment with increasing network capacity but we find that deep and wide model architectures bring training hurdles. So almost all subsequent models are based on the Transformer *base* architecture (Vaswani et al., 2017) as implemented in TenTrans[4], expect for pre-trained model M2M-100 trained using FAIRSEQ[5] (Ott et al., 2019).

3.2 Back-translation

Back-translation (briefly, BT) (Sennrich et al., 2016a) is an effective and commonly used data augmentation technique to incorporate monolingual data into a translation system.

In this work, for translation direction with more than 5 million bilingual data such as Catalan→Italian, we train a dedicated bilingual BT

[3] https://github.com/pytorch/fairseq/tree/master/examples/m2m_100

[4] https://github.com/TenTrans/TenTrans

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

		It	Oc	Ro	Ca
En	No filter	22.4M	73K	14.6M	7.1M
	Filtered	22.3M	59K	14.5M	7.0M
Es	No filter	4.4M	36K	6.4M	12.3M
	Filtered	4.3M	36K	4.2M	6.5M
Fr	No filter	4.8M	124K	1.6M	7.7M
	Filtered	4.7M	124K	1.5M	7.0M
Pt	No filter	24.3M	24K	5.7M	4.9M
	Filtered	15.6M	24K	5.6M	4.6M

Table 3: Number of sentences in high-low resource bilingual data.

	It	Oc	Ro
No filter	275M	378K	193.5M
Filtered	38.3M	225K	13.4M

Table 4: Number of sentences in low-resource monolingual data.

BT System	It-Ca	Oc-Ca	Ro-Ca
Bilingual model	37.74	-	-
Multilingual 4-to-4	31.41	23.72	51.02

Table 5: BLEU scores (%) for reverse models evaluated on the validation data.

model Italian→Catalan to translate Italian monolingual data into Catalan. For other translation directions with less than 5 million bilingual data, we use the jointly pre-trained many-to-many multilingual model with four low-resource languages as its source and target side (see Section 3.1) to back translate Occitan and Romanian monolingual data into Catalan. Beam search with beam size 5 is used when generating the synthetic sentences. Detailed results of reverse models are shown in Table 5.

3.3 Multilingual Model

Arivazhagan et al. (2019) shows that multilingual models can improve the translation performance of medium and low resource languages, as multilingual models are often trained on greater quantities of data compared to individual models. So we utilize high-low resource paired data such as English→Occitan in addition to low-resource bilingual data during training. Training on high-resource and low-resource language pairs together may bring knowledge transfer (Zoph et al., 2016), especially when languages are from the same linguistic family.

In the experiment, we train on four high-resource languages (Spanish, French, Portuguese and English) combined with four target-task low-resource languages together, resulting in an **8-to-4 multilingual** model with Ca, Oc, Ro, It as the target side. We randomly extract 2K sentence pairs from training data as the validation set for each high-low resource languages pairs. BPE codes and multilingual vocabulary are shared among all languages, but a shared multilingual vocabulary runs the risk of favoring high-resource languages over others, due to the imbalance of the dataset size the vocabulary is extracted. To reduce the effect of imbalanced dataset size, we apply a temperature sampling strategy named Vocabulary Sampling to construct a joined vocabulary. Following Arivazhagan et al. (2019), we set sampling temperature $\underline{T} = 5$.

Table 6 shows results on validation set of our baseline systems. Obviously, the universal encoder-decoder model outperforms the model with separate decoders for each target language by 7 BLEU on average. Compared to the bilingual baseline system, our universal multilingual 1-to-3 baseline system performs great improvement on low-resource languages, at the cost of sacrificing performance on relatively rich language Italian. However, the jointly trained multilingual 4-to-4 system shows performance degradation. We ascribe this phenomenon to multilingual model capacity is split for more translation directions, from 3 directions to 12 translation directions in this case.

3.4 Pivot-based Method

Pivot-based approaches are prevalent when addressing the data scarcity problem in machine translation, nonetheless, they suffer from cascaded translation errors: the mistakes made in the source-to-pivot translation will be propagated to the pivot-to-target translation (Dabre et al., 2020). Another pivot-based approach used in zero-resource transla-

Base System	Ca-Oc	Ca-Ro	Ca-It	Average BLEU
Bilingual models	30.02	-	-	28.48
	-	21.51	-	
	-	-	33.91	
Separate Decoders				
Multilingual 1-to-3	25.82	22.03	32.96	26.90
Universal Models				
Multilingual 1-to-3	43.40	24.16	32.92	**33.78**
Multilingual 4-to-4	41.44	22.81	31.01	31.75

Table 6: BLEU scores (%) for baseline systems evaluated in the validation data. And the numbers represent languages used in models, e.g. 1-to-3 means source side of model is Ca but target side consists of Oc, Ro, It, and 4-to-4 means both the source and target side of model consist of four languages Ca, Oc, Ro and It.

tion scenario is that the pivot side of the pivot-target parallel corpus is back-translated to the source language, creating a synthetic source-target parallel corpus (Lakew et al., 2018; Gu et al., 2019). In this work, we adopt the latter pivot-based method.

In practice, we consider four high-resource languages En, Es, Fr, Pt as pivot languages, thus we train a pivot-to-source multilingual model to back translate four pivot languages in pivot-to-target parallel data into source language. Owing to relatively rich data of Catalan-Italian, we only perform experiments on low-resource languages of Occitan and Romanian. To balance distribution between genuine parallel data and synthetic parallel data, we oversample genuine data to be of the same magnitude as synthetic data.

We can combine all synthetic parallel data generated from back-translation and pivot-based method with genuine parallel data to jointly train a multilingual model from scratch, which is named **Combine-All**. Source side of this model is comprised of four rich-resource and four low-resource languages, and target side of this model is comprised of four low-resource languages.

3.5 Pre-trained Model Fine-tuning

Because of the recent popularity of using large scale pre-training models to fine-tune specific languages and tasks, we employ the M2M-100, a true Many-to-Many multilingual translation model (Fan et al., 2020) that can translate between 100 languages which cover four task languages. Our experiments are based on the M2M-100 1.2B model due to its better performance than the 418M model. In the subsequent fine-tuning procedure, we follow the parameters setting in fine-tuning mBART (Liu et al., 2020). In three task directions, we try fine-tuning M2M-100 model with genuine bilingual data (Bilingual FT) and fine-tuning with genuine multilingual data (Multilingual FT). Moreover, we try fine-tuning the M2M-100 1.2B model using **Combine-All** data with four high-resource plus low-resource languages as the source side and four low-resource languages as the target side.

Unfortunately, M2M-100 model trains on SentencePiece(Kudo and Richardson, 2018) rather than Byte-Pair-Encoding so that the fine-tuned model can not be directly combined with the models that listed above for ensembling. We utilize synthetic Catalan-Occitan, Catalan-Romanian data generated through sentence-level knowledge distillation (Kim and Rush, 2016) to train a 'student' model so as to incorporate knowledge of 'teacher' model M2M-100 1.2B into 'student' model. Concretely, in Catalan→Occitan direction, we employ multilingual fine-tuning on M2M-100 1.2B model using Combine-All data for 200K updates (1.1M updates for each epoch), after that, we continue with bilingual fine-tuning using genuine Catalan-Occitan parallel data. As for Catalan→Romanian direction, we directly use the pre-trained model without fine-tuning. We continue to train on **8-to-4 multilingual** model (See Section 3.3) in three task translation directions with data obtained through knowledge distillation and finally get a new model named **M2M-KD**. We do not implement knowledge distillation in Catalan→Italian direction since we find other systems perform equivalently to the pre-trained model. If time permitted, we believe that more improvements will be observed.

3.6 Domain Adaptation

Domains of training data are various, whereas validation and hidden test data belong to the cultural

System	Ca-Oc	Ca-Ro	Ca-It	Average BLEU
Multilingual baseline	43.40	24.16	32.92	**33.78**
+ Back-translation	48.21	22.66	32.9	34.59
+ Pivot	26.98	26.33	34.2	29.17
Combine-All	26.75	29.59	37.49	31.28
M2M-100 418M				
w/o FT	31.04	26.72	34.18	30.65
+ Multilingual FT	40.71	25.26	33.92	33.30
+ Bilingual FT	49.42 - -	- 25.4 -	- - 35.19	36.67
M2M-100 1.2B				
w/o FT	34.70	32.21	38.37	35.09
+ Multilingual FT	42.09	28.11	36.62	35.61
+ Bilingual FT	49.79 - -	- 27.83 -	- - 37.09	38.24
+ Combine-All FT	37.15	30.54	37.28	34.99
+ Bilingual FT	49.86	-	-	-
Multilingual 8-to-4	51.49	29.11	38.26	39.62
+ In-domain-FT	56.60	28.30	38.74	41.21
+ M2M-KD	**65.18**	**32.85**	36.19	44.74
Ensemble				
†In-domain-FT + M2M-KD	64.70	32.85	39.41	45.65
*In-domain-FT + M2M-KD + Combine-All	64.02	32.63	**40.04**	45.56

Table 7: BLEU scores (%) for different systems on the validation data. The number 8 means source side of model consists of both four high-resource languages and four low-resource languages, 4 means target side of model consists of four low-resource languages Ca, Oc, Ro and It. '†' is the submitted primary system. '*' is the submitted contrastive system.

heritage domain. Owing to the domain discrepancy, adapting models to the cultural heritage domain (Luong et al., 2015) is required.

Due to the scarcity of in-domain data, we utilize pre-trained language model multilingual Bert [6] (Devlin et al., 2019) to train a domain classifier for extracting in-domain sentences from genuine bilingual data. To train the domain classifier, we consider validation data of three languages Ca, Ro, It as positive samples, and randomly sample the low-resource side of high-low resource bilingual data as negative samples. Then classifier is exploited to score the source sentences (Ca/Ro/It). We select sentence pairs whose source is predicted to be positive with a probability greater than threshold 0.7 to construct in-domain corpus. In the end, we pick out 60K Catalan-Occitan, 297K Catalan-Romanian and 815K Catalan-Italian data respectively as in-domain corpus. We fine-tune **8-to-4 multilingual** model on the in-domain corpus in three task translation directions and then get the **In-domain-FT** model. For the purpose of preventing overfitting, we set the max-tokens to be 2K with a learning rate of 3e-5 and we force fine-tuning to stop when finishing the first epoch. Note that we do not perform fine-tuning on the validation set.

[6] https://huggingface.co/bert-base-multilingual-cased

4 Experiments

4.1 Settings

Except that the pre-training experiments are trained on 4 NVIDIA V100 GPUs, the rest of our experiments are carried out with 8 NVIDIA P40 GPUs. Except for the pre-training experiments, the rest of our experiments use the following settings. Our models apply Adam (Kingma and Ba, 2015) as optimizer to update the parameters with $\beta_1 = 0.9$ and $\beta_2 = 0.98$. We set the label smoothing and dropout rate to 0.1. The initial learning rate is set to 5e-4

varied under a warm-up strategy with 4000 steps. In the training stage, batch size is 8K tokens per GPU.

We use uncased BLEU scores calculated with Moses multi-bleu.pl[7] toolkit as the evaluation metric. And we choose model checkpoints based on the BLEU score on average of the validation set.

4.2 Main Results

Table 7 shows that the translation quality is largely improved with different systems. Although minority systems encounter the problem of average performance degradation on the validation set, they contribute to at least one translation direction. Back-translation gives a solid improvement by nearly 0.8 BLEU on average. Pivot-based method offers 1∼2 BLEU in Catalan→Romanian, Catalan→Italian directions, however, pivot degrades in Catalan→Occitan direction. When we train an 8-to-4 multilingual model jointly with both the high and low resource languages, the model shows an absolute improvement in three task directions of 6 BLEU on average score. It can be explained by that a larger quantity of genuine data leads to robust encoder/decoder or knowledge can be transferred from high-resource into low-resource languages. As for the pre-trained model, we notice that M2M-100 1.2B model performs very well in Catalan→Romanian, Catalan→Italian directions without fine-tuning. And we find that average bilingual fine-tuning outperforms multilingual fine-tuning by about 2.6 BLEU. We also observe some systems hold a comparable performance with M2M-100 1.2B model in Catalan→Romanian and Catalan→Italian directions when training data is abundant.

Further experiments include the in-domain fine-tuning and M2M-KD based on the **multilingual 8-to-4** system. In-domain fine-tuning is restricted to in-domain data size, but we also obtain a solid improvement of 1.5 BLEU on average, especially in Catalan→Occitan direction. M2M-KD model yields a greater improvement that we get the best BLEU in Catalan→Occitan, Catalan→Romanian directions with 65.18, 32.85 respectively. Ultimately, to take advantages of multiple single models, two or three top performing models are ensembled to be the submitted systems.

[7] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl

5 Conclusions

In this paper, we present the system TenTrans submitted for the WMT21 Multilingual Low-Resource Translation for Indo-European Languages shared task. We focus on Romance languages, translating from Catalan to Occitan, Romanian and Italian. Back-translation, pivot-based method, multilingual model, knowledge distillation using pre-trained model, domain adaptation and ensembles are employed and proven effective in the experiments. Our best submitted system achieves an average of 43.45 case-sensitive BLEU score across all low-resource languages pairs.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 3874–3884. Association for Computational Linguistics.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George F. Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *CoRR*, abs/1907.05019.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A comprehensive survey of multilingual neural machine translation. *CoRR*, abs/2001.01115.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

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.

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. *CoRR*, abs/2010.11125.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor O. K. Li. 2019. Improved zero-shot neural machine translation via ignoring spurious correlations. *CoRR*, abs/1906.01181.

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.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 1317–1327. The Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

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, Brussels, Belgium. Association for Computational Linguistics.

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

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. *CoRR*, abs/2001.08210.

Minh-Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task sequence to sequence learning. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Minh-Thang Luong, Christopher D Manning, et al. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the international workshop on spoken language translation*, IWSLT. Da Nang, Vietnam.

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: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Demonstrations*, pages 48–53. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. 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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, pages 2214–2218. European Language Resources Association (ELRA).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 1568–1575. The Association for Computational Linguistics.