EXPLOITING GENERAL PURPOSE SEQUENCE REPRESENTATIONS FOR LOW RESOURCE NEURAL MACHINE TRANSLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper, we propose a Neural Machine Translation with General Purpose Sequence Representations (NMTwGSR) system for low-resource machine translation. The proposed system does not have explicit encoder as is the case for general encoder-decoder based NMT systems, instead, it exploits readily avail- able pre-trained sequence representations. The decoder as well takes the input from the masked BERT model and train further to learn encoder-decoder attentions between source and target language. The proposed system with the general-purpose sequence representations is more practical in low resource settings. Especially, if there are no auxiliary high-resource language pairs or monolingual data available to pre-train the NMT system. We evaluate the proposed NMT system on four low-resource language pairs (Ro-En, Fi-En, Tr-En, Lv-En) and empirical results show that our approach is efficient in handling the low resource translation task. For instance, on the Ro-En parallel corpus, our system attains 5.68 BLEU points improvement compared to the competitive NMT system which does not exploit these sequence representations in low a resource environment (one percent of the full corpus).

1 INTRODUCTION

Neural Machine Translation (NMT) has achieved near human-performance on several language pairs with the help of abundant parallel training data (Wu et al., 2016; Hassan et al., 2018). However, it is difficult to gather such large-scale parallel data for all the language pairs.

More recently, several approaches have been proposed to handle low resource translation. These approaches mainly fall into two categories: (1) utilizing monolingual data, and (2) using the knowledge obtained from related high resource language pairs. Many research efforts have been spent on incorporating monolingual data into machine translation, for example, Gülçehre et al. (2015); Zhang & Zong (2016) uses multi-task learning; Sennrich et al. (2016) uses back-translation, and Artetxe et al. (2017); Lample et al. (2018); Chen et al. (2018) proposes unsupervised machine translation.

In the second approach, several works such as Firat et al. (2016); Lee et al. (2017); Ha et al. (2016) exploit the knowledge of auxiliary translations or even auxiliary tasks. In **?**Gu et al. (2018a) leverage multilingualism into NMT. For instance, Gu et al. (2018a) simultaneously train multiple translation tasks using universal lexical representation which facilitate the sharing of embedding information across different languages. Meta-learning based NMT model (Gu et al., 2018b; Li et al., 2019) has shown improvements for low resource translation by pretraining on several auxiliary high resource translation tasks or by leveraging domain data from multiple sources. However, obtaining such universal lexical representations and related auxiliary tasks is not always easy. Moreover, most of these models assume target language is same for all language pairs (for examples, English as the target language in all pairs) making them difficult to apply for reverse translation tasks (English to other languages).

Recently, general purpose sequence representations (Peters et al., 2017; Alec Radford & Sutskever, 2018; Devlin et al., 2018) have led to strong improvements in several Natural Language Processing (NLP) tasks. In this context, a Transformer encoder/decoder is trained on a large unsupervised text corpus, and then fine-tuned on NLP tasks such as question-answering (Rajpurkar et al., 2016),

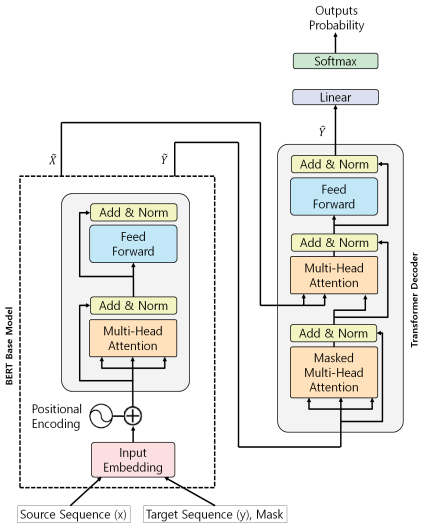


Figure 1: Overview of the proposed system based.

and named entity recognition (Tjong Kim Sang & De Meulder, 2003). There have been many attempts to utilize such representations into NMT systems as well. For example, Lample & Conneau (2019); Song et al. (2019) show that initialization with pretrained MaskedLM or Masked Sequence-to-Sequence is beneficial for machine translation, and Yang et al. (2019); Chen et al. (2019) distill sequence representation knowledge from BERT into NMT model. Also, Clinchant et al. (2019); Imamura & Sumita (2019) proposed incorporating BERT model into the NMT task. However, their incorporation is limited to the encoder, while our proposed method utilizes representations in the target side also.

Inspired by these latest approaches based on pretrained sequence representations (Devlin et al., 2018) in NLP, in this work, we propose a Neural Machine Translation with General Purpose Sequence Representations (NMTwGSR) approach for low resource translation. The proposed system falls into the first category of utilizing monolingual data. However, we do not explicitly pretrain the NMT model on monolingual data compared to the previous works such as Ramachandran et al. (2017), instead, we use pretrained sequence representations. There are several advantages to the proposed NMT system:

- It doesn't assume the availability of several high resource language pairs for pretraining as required by the previous approaches such as transfer or meta learning.
- It achieves significant BLEU point improvement on several low-resource translation tasks compared to the competitive NMT system (Vaswani et al., 2017).
- The previously proposed techniques such as back-translation, multilingual-NMT, transfer/meta-learning training strategies are straight forward to integrated into to the proposed system whenever there is an availability of related high resource language pair(s) for the corresponding low-resource language pair training.
- We extensively evaluate the proposed approach on four low-resource translation tasks. The proposed model consistently outperforms the strong NMT baseline which does not exploits sequence representations and the gap widens as the number of training examples decreases.
- The size of the above four languages are very small compared to the well-known language pairs such as En-De and En-Fr. The performance improvements on these language pairs achieved by our model reveals that it can be easily adopted to a new language pair coming from low resource languages such as Asian and African, whenever the sequence representations are available for these in language models like BERT.

2 NMT SYSTEM FOR LOW RESOURCE TRANSLATION

In Section 2.1 we briefly describe the BERT model used to obtain the pretrained token representations for source and target sequences. The brief details about the Transformer decoder for generating the target sequence are provided in Section 2.2. Finally, in Section 2.3 we introduce our proposed system.

2.1 PRETRAINED LANGUAGE REPRESENTATIONS

The architecture of BERT is based on multi-layer Transformer encoder. Figure 1 contains one layer of BERT model. The language representations from the BERT model drastically improved the performance of several classification tasks from GLUE benchmark (Wang et al., 2018). Unlike traditional language models, BERT computes token representations using left-to-right and right-to-left contextual information. To facilitate the training of this bi-directional language model it uses masked language model and next sentence prediction as training objectives. Due to the bidirectional nature and multi-language model training, the BERT model can efficiently represent the source and target sequence of the translation task.

2.2 TRANSFORMER DECODER

The second component in the proposed system is based on the Transformer decoder block. Each block in the Transformer decoder contains three sub-layers. The first two sub-layers are a position-wise fully connected feed-forward network, a multi-head self-attention mechanism, and used to compute the representations for the input sequence. The third sub-layer is used to compute the attention (context) vector of the source-target sequence based on soft-attention approaches (Bahdanau et al., 2015). One layer of the Transformer decoder is shown in Figure 1.

2.3 PROPOSED SYSTEM

The overview of the proposed system is shown in Figure 1. The tokens in the source and target languages are represented by word-piece ids. We use word-piece vocabulary available in BERT Multilingual Case model (section 2.1) to tokenize both the sequences.

BERT for source language The source language sequence ($\boldsymbol{x}$) is fed into BERT model to get the token representations ($\tilde{\boldsymbol{X}}$),

$$\tilde{\boldsymbol{X}} = \mathrm{BERT}(\boldsymbol{x}) \in \mathbb{R}^{m \times d}. \tag{1}$$

At this stage each token in the source language contains contextual information from all other tokens falling on left-side as well as right-side to it.

BERT for target language By nature of the design, each token representation in the BERT model contains contextual information from the right-side tokens. However, predicting the target sequence token should only depend on the previously predicted tokens. To address this issue we explicitly mask the right side tokens. The target language sequence representation ($\tilde{\boldsymbol{Y}}$) is obtained as follows:

$$\tilde{\boldsymbol{Y}} = \mathrm{BERT}_{\mathrm{Masked}}(\boldsymbol{y}, \mathrm{mask}) \in \mathbb{R}^{n \times d}. \tag{2}$$

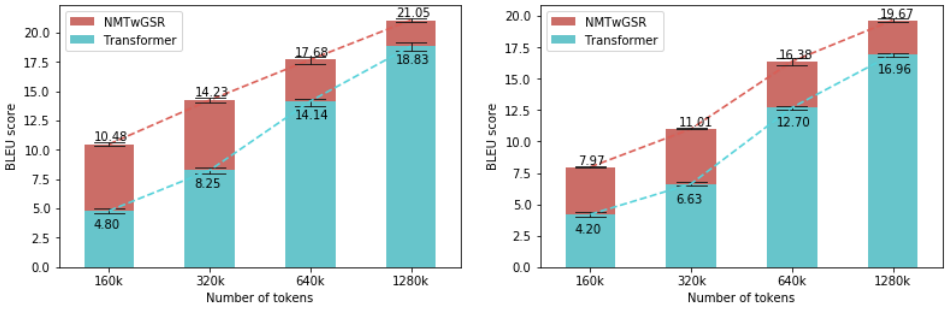
Here, mask prevents model to attend right-side tokens.

Decoder The decoder is based on the 6 identical layers of Transformer decoder (section 2.2). It takes both the source language representations ($\tilde{X}$) and the target language representations ($\tilde{Y}$) as input and computes $\hat{Y}$ using the three sub-layers presented in each layer of the Transformer decoder.

$$\hat{\boldsymbol{Y}} = \mathrm{Transformer}_{\mathrm{decoder}}(\tilde{X}, \tilde{Y}). \tag{3}$$

Finally, we apply the learned transformation and softmax function to convert the output from the decoder to predict the target token probabilities.

Training The parameters from the pre-trained BERT model are kept constant. We only learn parameters from the decoder and the output layer.



(a) Ro-En (b) En-Ro

Figure 2: The models are trained on 160k, 320k, 640k, and 1280k sampled subsets.

#Tok	Ro-En		En-Ro		Fi-En		En-Fi		Tr-En		En-Tr		Lv-En		En-Lv	
	Base	Ours	Base	Ours	Base	Ours	Base	Ours	Base	Ours	Base	Ours	Base	Ours	Base	Ours
Full	31.76	32.68	-	-	20.20	22.39	-	-	13.74	15.19	-	-	15.15	17.39	-	-
1280K	18.83	21.05	16.96	19.67	7.70	9.26	7.20	8.02	9.40	10.24	9.87	10.12	6.17	7.19	5.32	6.39
640K	14.14	17.68	12.70	16.38	5.53	7.62	3.67	4.58	5.08	7.76	4.58	6.43	3.80	5.22	3.01	4.30
320K	8.25	14.23	6.63	11.01	3.24	5.47	1.98	2.92	2.59	5.44	1.74	3.91	2.29	3.52	1.71	2.76
160K	4.80	10.48	4.20	7.97	1.04	3.12	0.53	1.91	1.44	3.49	1.02	2.15	0.85	2.15	0.96	1.72

Table 1: Test BLEU results of models trained on 160k, 320k, 640k, and 1280k sampled subsets of four langauge paris both for forward and backward. The models trained on Full dataset is only presented with forward direction.

3 Experimental Settings

In this section we discuss the datasets used for conducting the experiments and implementation details of the proposed model.

3.1 Datasets

We evaluate our model on four different language pairs: Romanian(Ro) / Finnish(Fi) / Turkish(Tr) / Latvian(Lv) - English(En). The Ro-En dataset is taken from WMT'16 and the remaining 3 languages; Fi-En, Tr-En and Lv-En are taken from WMT'17. We use the standard train, validation and test splits provided in the WMT'16 and WMT'17 tasks.

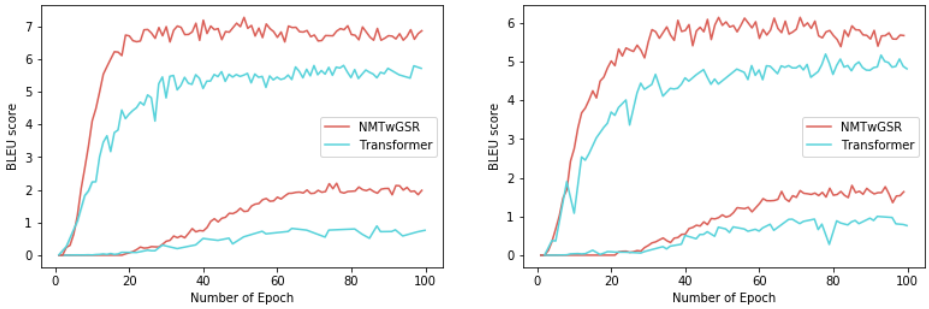
Low resource environment is simulated by randomly sampling the training set from the corpus based on 160k, 320k, 640k, and 1280k English tokens. For all the language pairs, we sample training set for five times for each subset and report the average BLEU score and its standard deviation. The total number of tokens in the full corpus of Lv/Fi/Ro/Tr-En are 67.24M, 64.5M, 16.66M and 5.58M respectively. For example, subset 160k and 320K in Ro-En language pair only contain approximately one and five percent of the full corpus respectively.

3.2 Implementation details

The token representation for source and target sequence are computed using `BERT-Base, Multilingual Cased model` which is release by Devlin et al. (2018). The proposed model is implemented based on Tensorflow(Abadi et al., 2015) and OpenNMT(Klein et al., 2017) frameworks. The weights of the BERT are fixed during training our model. The Transformer model(Vaswani et al., 2018) is used as a baseline with *Tranformer base* settings. The decoder in our models also uses the same set of hyperparameters similar to the Transformer model ($n_{layer} = 6, n_{head} = 8, warmup_{steps} = 16k$).

4 Results

To see the effect of general purpose sequence representations on low resource translation tasks, we compare our proposed system which utilizes sequence represenations against Trannsformer model Vaswani et al. (2017) without these representations.



(a) Lv-En (b) En-Lv

Figure 3: The BLEU scores for Lv-En and En-Lv pairs at different stages of training. In each plot, the upper curves are obtained using 1280k subset and lower curves are obtained using 160k subset.

4.1 Training in Low Resource Environment

We train the proposed, NMTwGSR and Transformer methods on 160k, 320k, 640k, and 1280k sampled training subsets of Ro/Fi/Tr/Lv-En translation tasks. To see the effectiveness of our model in generating low resource language as the target language, we also conduct experiments on the reverse translation task (En-Ro/Fi/Tr/Lv). The results of all the language pairs obtained on the test sets are shown in Table 1. We aslo plot the bar graph in 2 for Ro-En and En-Ro language pairs to show the improvements visually, and other three languages also have similar visual improvements.

From Table 1, we can see that our model clearly outperforms the Transformer model in low resource environments for both forward and reverse translation tasks. This empirical results from various languages show that our model tackles the problem of low resource translation in an simple way compared to collecting huge parallel corpora or monoligula corpora and training from scratch. Moreover the proposed model is eaiser train to due to the less number of trainable parameters compared to the baseline.

4.1.1 Learning curves

In Figure 3, we show the learning curves of both the models obtained on 160K and 1280K subsets of Lv-En and En-Lv pairs. The curves are plotted based on the average of five runs for each subset. The advantage of general purpose sequence representations for low resource translation is clearly observed in Figure 3. Even though in the initial phase of training both the models achieve the same level of BLEU points, the NMTwGSR surpasses the Transformer model as the training progress. Similar trends are observed across different language pairs and training subsets.

4.2 Training on Full Corpus

Along with low resource environment, we also test the general applicability of our model by training it on the full corpus. The number of the parallel sentence in each language pair ranges from 4.46M to 0.21M. The results for all the language pairs are also provided in Table 1. The BLEU scores of the Transformer model are taken from Gu et al. (2018b). It can be noted that the proposed method achieved significant BLEU score improvement in all the translation tasks.

5 Conclusion

In this work, we propose an NMT system based on the general purpose pretranied sequence representations for low resource translation. The proposed system is based on the recently released BERT model and adapted it to work for the decoding step. With the help of multilingual representations available from the BERT model, it can be applied to many new language pairs without depending auxiliary high resource language pairs. The propopsed system can be easily integreted with new better language models whenever they are available. Our experimental results on four low resource language pairs show the effectiveness of the proposed approach for both the forward and reverse translation tasks.

REFERENCES

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL `https://www.tensorflow.org/`. Software available from tensorflow.org.

Tim Salimans Alec Radford, Karthik Narasimhan and Ilya Sutskever. Improving language understanding with unsupervised learning. *Technical report, Open AI*, 2018.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 451–462, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1042. URL `https://www.aclweb.org/anthology/P17-1042`.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. 2015.

Yen-Chun Chen, Zhe Gan, Yu Cheng, Jingzhou Liu, and Jingjing Liu. Distilling the knowledge of bert for text generation. *arXiv preprint arXiv:1911.03829*, 2019.

Yun Chen, Yang Liu, and Victor O. K. Li. Zero-resource neural machine translation with multi-agent communication game. *AAAI*, 2018.

Stephane Clinchant, Kweon Woo Jung, and Vassilina Nikoulina. On the use of bert for neural machine translation. *Proceedings of the 3rd Workshop on Neural Generation and Translation*, 2019. doi: 10.18653/v1/d19-5611. URL `http://dx.doi.org/10.18653/v1/d19-5611`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL `http://arxiv.org/abs/1810.04805`.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 866–875, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1101. URL `https://www.aclweb.org/anthology/N16-1101`.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O.K. Li. 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)*, pp. 344–354, New Orleans, Louisiana, June 2018a. Association for Computational Linguistics. doi: 10.18653/v1/N18-1032. URL `https://www.aclweb.org/anthology/N18-1032`.

Jiatao Gu, Yong Wang, Yun Chen, Victor O. K. Li, and Kyunghyun Cho. Meta-learning for low-resource neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3622–3631, Brussels, Belgium, October-November 2018b. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/D18-1398`.

Çaglar Gülçehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loïc Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. On using monolingual corpora in neural machine translation. *CoRR*, abs/1503.03535, 2015. URL `http://arxiv.org/abs/1503.03535`.

Thanh-Le Ha, Jan Niehues, and Alexander H. Waibel. Toward multilingual neural machine translation with universal encoder and decoder. *CoRR*, abs/1611.04798, 2016. URL http://arxiv.org/abs/1611.04798.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William 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. Achieving human parity on automatic chinese to english news translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4791–4796, 2018.

Kenji Imamura and Eiichiro Sumita. Recycling a pre-trained BERT encoder for neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pp. 23–31, Hong Kong, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5603. URL https://www.aclweb.org/anthology/D19-5603.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pp. 67–72, Vancouver, Canada, July 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P17-4012.

Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*, 2019.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 5039–5049, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D18-1549.

Jason Lee, Kyunghyun Cho, and Thomas Hofmann. Fully character-level neural machine translation without explicit segmentation. *Transactions of the Association for Computational Linguistics*, 5:365–378, 2017. doi: 10.1162/tacl_a_00067. URL https://www.aclweb.org/anthology/Q17-1026.

Rumeng Li, Xun Wang, and Hong Yu. Metamt, a metalearning method leveraging multiple domain data for low resource machine translation. *arXiv preprint arXiv:1912.05467*, 2019.

Matthew E. Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. Semi-supervised sequence tagging with bidirectional language models. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 2017.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL https://www.aclweb.org/anthology/D16-1264.

Prajit Ramachandran, Peter Liu, and Quoc Le. Unsupervised pretraining for sequence to sequence learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 383–391, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1039. URL https://www.aclweb.org/anthology/D17-1039.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 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)*, pp. 86–96, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1009. URL https://www.aclweb.org/anthology/P16-1009.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mass: Masked sequence to sequence pre-training for language generation. *arXiv preprint arXiv:1905.02450*, 2019.

Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pp. 142–147, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics. doi: 10.3115/1119176.1119195. URL https://doi.org/10.3115/1119176.1119195.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan N. Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. Tensor2tensor for neural machine translation. *CoRR*, abs/1803.07416, 2018. URL http://arxiv.org/abs/1803.07416.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W18-5446.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. In *Transactions of the Association for Computational Linguistics*, pp. 339–351, 2016.

Jiacheng Yang, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Yong Yu, Weinan Zhang, and Lei Li. Towards making the most of bert in neural machine translation. *arXiv preprint arXiv:1908.05672*, 2019.

Jiajun Zhang and Chengqing Zong. Exploiting source-side monolingual data in neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1535–1545, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1160. URL https://www.aclweb.org/anthology/D16-1160.