Sub-Word Alignment is Still Useful: A Vest-Pocket Method for Enhancing Low-Resource Machine Translation

Minhan Xu, Yu Hong[*]

School of Computer Science and Technology, Soochow University, China
cosmosbreak5712@gmail.com, tianxianer@gmail.com

Abstract

We leverage embedding duplication between aligned sub-words to extend the Parent-Child transfer learning method, so as to improve low-resource machine translation. We conduct experiments on benchmark datasets of My→En, Id→En and Tr→En translation scenarios. The test results show that our method produces substantial improvements, achieving the BLEU scores of 22.5, 28.0 and 18.1 respectively. In addition, the method is computationally efficient which reduces the consumption of training time by 63.8%, reaching the duration of 1.6 hours when training on a Tesla 16GB P100 GPU. All the models and source codes in the experiments will be made publicly available to support reproducible research.

1 Introduction

Low-resource machine translation (MT) is challenging due to the scarcity of parallel data and, in some cases, the absence of bilingual dictionaries (Zoph et al., 2016; Miceli Barone, 2016; Koehn and Knowles, 2017; Zhang et al., 2017). Unsupervised, multilingual and transfer learning have been proven effective in the low-resource MT tasks, grounded on different advantages (section 2).

In this paper, we follow Aji et al. (2020)'s work to utilize cross-language transfer learning, of which the "parent-child" transfer framework is first proposed by Zoph et al. (2016). In the parent-child scenario, a parent MT model and a child MT model are formed successively, using the same neural network structure. In order to achieve the sufficient warm-up effect from scratch, the **parent** is trained on **high**-resource language pairs. Further, the **child** inherits the parent's properties (e.g., inner parameters and embedding layers), and it is boosted by the fine-tuning over **low**-resource language pairs. One of the distinctive contributions in Aji et al. (2020)'s

study is to demonstrate the significant effect of embedding duplication for transference, when it is conducted between the morphologically-identical sub-words in different languages.

We attempt to extend Aji et al. (2020)'s work by additionally duplicating embedding information among the aligned multilingual sub-words. It is motivated by the assumption that if the duplication between morphologically-identical sub-words contributes to cross-language transference, the duplication among any other type of equivalents is beneficial in the same way, such as that of the aligned sub-words, most of which are likely to be morphologically-dissimilar but semantically-similar (or even exactly the same).

In our experiments, both the parent and child MT models are built with the transformer-based (Vaswani et al., 2017) encoder-decoder architecture (Section 3.1). We use the unigram model from SentencePiece (Kudo and Richardson, 2018) for tokenizing, and carry out sub-word alignment using eflomal (Section 3.2). On the basis, we develop a normalized element-wise embedding aggregation method to tackle the many-to-one embedding duplication for aligned sub-words (Section 3.3). The experiments show that our method achieves substantial improvements without using data augmentation.

2 Related Work

The majority of previous studies can be sorted into 3 aspects in terms of the exploited learning strategies, including unsupervised, multilingual and transfer learning.

- **Unsupervised** MT conducts translation merely conditioned on monolingual language models (Lample et al., 2018a; Artetxe et al., 2017). The ingenious method that has been explored successfully is to bridge the source and target languages using a shareable

---

[*] Corresponding author.

representation space (Lample et al., 2018b), which is also known as interlingual (Cheng et al., 2017) or cross-language embedding space (Kim et al., 2018). To systematize unsupervised MT, most (although not all) of the arts leverage bilingual dictionary induction (Conneau et al., 2018; Søgaard et al., 2018), iterative back-translation (Sennrich et al., 2016a; Lample et al., 2018b) and denoised auto-encoding (Vincent et al., 2008; Kim et al., 2018).

- **Multilingual** MT conducts translation merely using a single neural model whose parameters are thoroughly shared by multiple language pairs (Firat et al., 2016; Lee et al., 2017; Johnson et al., 2017; Gu et al., 2018a,b), including a variety of high-resource language pairs as well as a kind of low-resource (the target language is fixed and definite). Training on a mix of high-resource and low-resource (even zero-resource) language pairs enables the shareable model to generalize across language boundaries (Johnson et al., 2017). The benefits result from the assimilation of relatively extensive translation experience and sophisticated modes from high-resource language pairs.

- **Transferable** MT is fundamentally similar to multilingual MT, whereas it tends to play the aforementioned Parent-Child game (Zoph et al., 2016). A variety of optimization methods have been proposed, including the transfer learning over the embeddings of WordPieces tokens (Johnson et al., 2017), BPE sub-words (Nguyen and Chiang, 2017) and the shared multilingual vocabularies (Kocmi and Bojar, 2018; Gheini and May, 2019), as well as the transference that is based on the artificial or automatic selection of congeneric parent language pairs (Dabre et al., 2017; Lin et al., 2019). In addition, Aji et al. (2020) verify the different effects of various transferring strategies of sub-word embeddings, such as that among morphologically-identical sub-words.

In this paper, we extend Aji et al. (2020)'s work, transferring embedding information not only among the morphologically-identical sub-words but the elaborately-aligned sub-words.

3 Approach

3.1 Preliminary: Basic Transferable NMT

We follow Kim et al. (2019) and Aji et al. (2020) to build neural MT (NMT) models with 12-layer transformers (Vaswani et al., 2017), in which the first 6 layers are used as the encoder while the subsequent 6 layers the decoder.

Embedding Layer As usual, the encoder is coupled with a trainable embedding layer, which maintains a fixed bilingual vocabulary and trainable subword embeddings. Each embedding is specified as a 512-dimensional real-valued vector.

Parent-Child Transfer We follow Zoph et al. (2016) to conduct Parent-Child transfer learning. Specifically, we adopt an off-the-shelf transformer-based NMT[1] which was adequately trained on high-resource De→En (German→English) language pairs. The publicly-available data of OPUS (Tiedemann, 2012) is used for training, which comprises about 351.7M De→En parallel sentence pairs[2]. We regard this NMT model as the Parent. Further, we transfer all inner parameters of the 12-layer transformers from Parent to Child.

By contrast, the embedding layer of Parent is partially transferred to Child, which has been proven effective in Aji et al. (2020)'s study. Assume V_h denotes the high-resource (e.g., the aforementioned De-En) vocabulary while V_l the low-resource, the morphologically-identical sub-words V_o are then specified as the ones occurring in both V_h and V_l (i.e., $V_o = V_h \cap V_l$). Thus, we duplicate the embeddings of morphologically-identical sub-words V_o from the embedding layer of Parent to that of Child. Further, we randomly initialize the embeddings of the rest sub-words V_r in the Child's embedding layer ($V_r = V_l - V_o$), where random sampling from a Gaussian distribution is used.

Both the transferred inner parameters and the duplicated embeddings constitutes the initial state of the Child NMT model. On the basis, we fine-tune Child on the low-resource language pairs, such as the considered 18K My→En (Burmese→English) parallel data in our experiments.

3.2 Tokenizer and Alignment

We strengthen Parent-Child transfer learning by additionally duplicating embeddings for aligned sub-words (between low and high-resource languages).

	Doc.	Sent.	Token
My	113K	1.1M	17.4M
Id	1.1M	8.3M	156.2M
Tr	705K	5.8M	128.2M

Table 1: Statistics of monolingual Wikipedia data.

	Train.	Val.	Test
My-En (ALT)	18K	1K	1K
Id-En (BPPT)	22K	1K	1K
Tr-En (WMT17)	207K	3K	3K

Table 2: Statistics for low-resource parallel datasets.

The precondition is to produce the word-level alignment and equivalently assign it to sub-words.

Word Alignment We use Eflomal[3] to achieve the word alignment. It is developed based on EF-MARAL (Östling et al., 2016), where Gibbs sampling is run for inference on Bayesian HMM models. Eflomal is not only computationally efficient but able to perform n-to-1 alignment. We separately train Eflomal on the low-resource My→En, Id (Indonesian)→En and Tr (Turkish)→En parallel data (Section 4).

Sub-word Tokenizer We train a sub-word tokenizer using the unigram model of SentencePiece for each low-resource language, including My, Id and Tr. The tokenizers are trained on monolingual plain texts which are collected from Wikipedia's dumps[4]. The toolkit wikiextractor[5] is utilized to extract plain texts from the semi-structured data. The statistics of training data is shown in Table 1.

We uniformly set the size of sub-word vocabulary to 50K when training the tokenizers. The obtained vocabulary of each low-resource language is utilized for sub-word alignment, towards the mixed De-En sub-word vocabulary in the Parent NMT model. The size of De-En vocabulary is 58K.

Sub-word Alignment Given a pair of aligned bilingual words, we construct the same correspondence for their sub-words by many-to-many mappings. See the De→Tr example in (1).

(1) Word Alignment: | *produktion↔üretme*
| *Harnstoff↔üre*
Sub-word Alignment: | *produck↔{**üre**, tme}*
| *tion↔{**üre**, tme}*
| *Harn↔{**üre**}*
| *stoff↔{**üre**}*

It is unavoidable that some of the aligned sub-words are non-canonical. Though, the positive effect on transfer learning may be more substantial than negative. It motivated by the findings that the use of sub-words ensures a sufficient overlap

---

[3]https://github.com/robertostling/eflomal
[4]https://dumps.wikimedia.org
[5]https://github.com/attardi/wikiextractor

---

between vocabularies (Nguyen and Chiang, 2017), and thus enables the transfer of a larger number of concrete embeddings rather than random ones.

3.3 N-to-1 Embedding Duplication

Assume that V_l^a denotes the sub-words in low-resource vocabulary that have aligned sub-words in high-resource vocabulary, the mapping is $D(x)$, note that $\forall x \in V_l^a$, $D(x)$ is a set of sub-words. Thus, in the embedding layer of Child, we extend the range of sub-words for embedding transfer, including both the identical sub-words V_o and the aligned V_l^a. To enable the transfer, we tackle n-to-1 embedding duplication. It is because that, in a large number of cases, there is more than one high-resource sub-word corresponding to a single low-resource sub-word (see "*üre*" in (1)).

Given a sub-word x in V_l^a and the aligned sub-words v_x in $D(x)$, we rank v_x in terms of the frequency with which they were found to be aligned with x in the parallel data. On the basis, we carry out two duplication methods as below.

- **Top-1** We take the top-1 sub-word $\check{x}$ from v_x, and perform element-wise embedding duplication from $\check{x}$ to x: $\forall i, E_i(\check{x}) = E_i(x)$ (i is the i-th dimension of embedding $E(*)$).

- **Mean** We adopt all the sub-words in v_x, and duplicate their embedding information by the normalized element-wise aggregation (where, n denotes the number of sub-words in v_x):

$$\forall i, E_i(\check{x}) = \sum_{x \in v_x} E_i(x)/n$$

4 Experimentation

4.1 Datasets and Evaluation Metric

We evaluate the transferable NMT models for three source languages (My, Id and Tr). English is invariably specified as the target language. There are three low-resource parallel datasets used for training the Child NMT model, including Asian Language Treebank (ALT) (Ding et al., 2018), PAN Localization BPPT[6] and the corpus of WMT17 news

---

[6]http://www.pan10n.net/english/OutputsIndonesia2.htm

Model	My-En	Id-En	Tr-En
Baseline	20.5	26.0	17.0
MI-PC	21.0	27.5	17.6
Top-1-PC	21.9	27.6	18.0
Mean-PC	**22.5**	**28.0**	**18.1**

Table 3: Results using **SentencePiece** tokenizer.

Model	My-En	Id-En	Tr-En
Baseline	20.2	24.5	16.5
MI-PC	20.4	24.2	16.8
Top-1-PC	21.2	26.9	16.9
Mean-PC	**21.9**	**27.1**	**16.9**

Table 4: Results using **BPE** tokenizer.

translation task (Bojar et al., 2017). The statistics in the training, validation and test sets is shown in Table 2. We evaluate all the considered NMT models with SacreBLEU (Post, 2018).

4.2 Hyperparameters

We use an off-the-shelf NMT model as Parent (Section 3.1), whose state variables (i.e., hyperparameters and transformer parameters) and embedding layer are all set. On the contrary, the Child NMT model needs to be regulated from scratch.

When training and developing Child, we adopt the following hyperparameters. Each source language was tokenized using SentencePiece (Kudo and Richardson, 2018) with 50k vocabulary size. Training was carried out with HuggingFace Transformers library (Wolf et al., 2020) using the Adam optimizer with 0.1 weight decay rate. The maximum sentence length was set to 128 and the batch size to 64 sentences. The learning rate was set to 5e-5 and checkpoint frequency to 500 updates. For each model, we selected the checkpoint with the lowest perplexity on the validation set for testing.

5 Results and Analysis

Table 3 shows the test results, where all the considered Parent-Child transfer models are marked with "PC", and the baseline is the transformer-based NMT (Section 3.1) which is trained merely using low-resource parallel data (without transfer learning). MI-PC is the reproduced transfer model in terms of Aji et al. (2020)'s study, in which only the embedding transference of morphologically-identical sub-words is used. We report NMT performance when MI-PC is used to enhance the baseline, as well as that when our auxiliary transfer
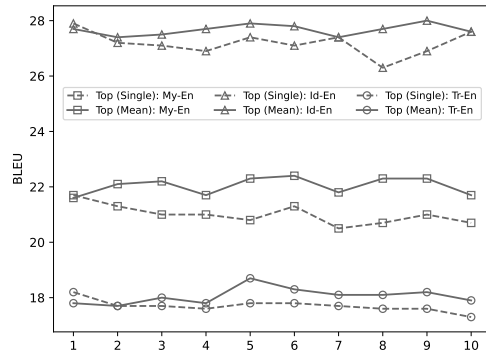

Figure 1: Comparison between embedding duplication of a single aligned sub-word (denoted with **Single**) and that of multiple sub-words (**Mean**).

Model	My-En	Id-En	Tr-En
Baseline	1.30	1.27	4.49
MI-PC	1.30	1.35	3.53
Top-1-PC	1.11	1.00	3.07
Mean-PC	**0.96**	**0.94**	**2.14**

Table 5: The time (in hour) that different MT models consumed during training in all experiments (0.9 hour is equivalent to 54 minutes).

models (i.e., Top-1 and Mean in Section 3.3) are additionally adopted, separately.

It can be observed that, compared to MI-PC, both Top-1-PC and Mean-PC yield improvements for all the three low-resource MT scenarios. The most significant improvement occurs for My→En MT, reaching up to 1.5 BLEU. Both the models generalize well across changes in the input sub-words. It can be illustrated in a separate experiment where the BPE (Sennrich et al., 2016b) tokenizer is used (instead of SentencePiece (Kudo and Richardson, 2018)), and all the transfer models are run over the newly-aligned sub-words. As shown in Table 4, both Top-1-PC and Mean-PC still outperform MI-PC, yielding an improvement of 2.9 BLEU at best (for Id→En MT).

Due to unavoidable errors in the sub-word alignment, the utilization of a single aligned sub-word for embedding duplication easily results in performance degradation. Aggregating and normalizing embeddings of all possible aligned sub-words help to overcome the problem. Figure 1 shows the NMT performance obtained when the i-th top-ranked aligned sub-word is exclusively used for transfer, as well as the aggregation of top-i sub-words is used. It can be found that the latter model almost always outperforms the former model.

We compare the training time consumption of all experiments, the result is shown in Table 5. We use mixed precision for training the child MT model. All experiments are conducted on a single NVIDIA P100 16GB GPU.

Obviously, the time that Mean-PC consumes during training is less than other models. In the scenario of Tr-En MT, the training duration is even shortened from 4.49 hours (i.e., about 269 minutes) to 2.14, compared to the baseline model. Most probably, it is caused by the transferring of a larger number of sub-word embeddings during training. In other word, Mean-PC actually transfers not only morphologically-identical sub-words but the aligned ones. This contributes more to the avoidance of redundant learning over sub-word embeddings. All in all, Mean-PC is less time-consuming when producing substantial improvements.

6 Conclusion

We enhance transferable Parent-Child NMT by duplicating embeddings of aligned sub-words. The experimental results demonstrate that the proposed method yields substantial improvements for all the considered MT scenarios (including My-En, Id-En and Tr-En). More importantly, we successfully reduce the training duration. The efficiency can be improved with the ratio of about 50% at best.

Additional survey in the experiments reveals that phonetic symbols can be used for transfer learning between the languages belonging to different families. For example, the phonologies of *hamburger* in German and Burmese are similar (Hámburger vs hambhargar). In the future, we will study bilingual embedding transfer of phonologically-similar words, so as to further improve low-resource NMT.

Acknowledgements

References

Alham Fikri Aji, Nikolay Bogoychev, Kenneth Heafield, and Rico Sennrich. 2020. In neural machine translation, what does transfer learning transfer? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7701–7710, Online. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2017. Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041*.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Yong Cheng, Yang Liu, Qian Yang, Maosong Sun, and Wei Xu. 2017. Joint training for pivot-based neural machine translation. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 3974–3980.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *Proceedings of the 6th International Conference on Learning Representations*.

Raj Dabre, Tetsuji Nakagawa, and Hideto Kazawa. 2017. An empirical study of language relatedness for transfer learning in neural machine translation. In *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation*, pages 282–286. The National University (Phillippines).

Chenchen Ding, Masao Utiyama, and Eiichiro Sumita. 2018. Nova: A feasible and flexible annotation system for joint tokenization and part-of-speech tagging. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 18(2):1–18.

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

Mozhdeh Gheini and Jonathan May. 2019. A universal parent model for low-resource neural machine translation transfer. *arXiv preprint arXiv:1909.06516*.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O.K. Li. 2018a. 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.

Jiatao Gu, Yong Wang, Yun Chen, Victor O. K. Li, and Kyunghyun Cho. 2018b. Meta-learning for low-resource neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods*

in Natural Language Processing, pages 3622–3631, Brussels, Belgium. Association for Computational Linguistics.

Melvin Johnson, 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. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Yunsu Kim, Yingbo Gao, and Hermann Ney. 2019. Effective cross-lingual transfer of neural machine translation models without shared vocabularies. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1246–1257, Florence, Italy. Association for Computational Linguistics.

Yunsu Kim, Jiahui Geng, and Hermann Ney. 2018. Improving unsupervised word-by-word translation with language model and denoising autoencoder. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 862–868, Brussels, Belgium. Association for Computational Linguistics.

Tom Kocmi and Ondřej Bojar. 2018. Trivial transfer learning for low-resource neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 244–252, Brussels, Belgium. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. 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.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *Proceedings of the 6th International Conference on Learning Representations*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018b. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium. Association for Computational Linguistics.

Jason Lee, Kyunghyun Cho, and Thomas Hofmann. 2017. Fully character-level neural machine translation without explicit segmentation. *Transactions of the Association for Computational Linguistics*, 5:365–378.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

Antonio Valerio Miceli Barone. 2016. Towards cross-lingual distributed representations without parallel text trained with adversarial autoencoders. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 121–126, Berlin, Germany. Association for Computational Linguistics.

Toan Q. Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 296–301, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Robert Östling, Jörg Tiedemann, et al. 2016. Efficient word alignment with markov chain monte carlo. *The Prague Bulletin of Mathematical Linguistics*.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

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 (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788, Melbourne, Australia. Association for Computational Linguistics.

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, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1959–1970, Vancouver, Canada. Association for Computational Linguistics.

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*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.