PICT-NLP@WMT22-EMNLP2022: Unsupervised and Very-Low Resource Supervised Translation on German and Sorbian Variant Languages

Aditya Vyawahare * aditya.vyawahare07@gmail.com
Rahul Tangsali * rahuul2001@gmail.com
Aditya Mandke † amandke@ucsd.edu
Onkar Litake † olitake@ucsd.edu
Dipali Kadam ‡ ddkadam@pict.edu
Pune Institute of Computer Technology, India

Abstract

This paper presents the work of team PICT-NLP for the shared task on unsupervised and very low-resource supervised machine translation, organized by the Workshop on Machine Translation, a workshop in collocation with the Conference on Empirical Methods in Natural Language Processing (EMNLP 2022). The paper delineates the approaches we implemented for supervised and unsupervised translation between the following 6 language pairs: German-Lower Sorbian (de-dsb), Lower Sorbian-German (dsb-de), Lower Sorbian-Upper Sorbian (dsb-hsb), Upper Sorbian-Lower Sorbian (hsb-dsb), German-Upper Sorbian (de-hsb), and Upper Sorbian-German (hsb-de). For supervised learning, we implemented the transformer architecture from scratch using the Fairseq library. Whereas for unsupervised learning, we implemented Facebook's XLM masked language modeling approach. We discuss the training details for the models we used, and the results obtained from our approaches. We used the BLEU and chrF metrics for evaluating the accuracies of the generated translations on our systems.

1 Introduction

Neural machine translation has witnessed significant progress in the case of highly spoken languages such as English (Bahdanau et al., 2015), Mandarin (Li et al., 2022), French (Emezue and Dossou, 2020), etc. However, in many cases, it becomes challenging to develop a robust bilingual machine translation system, especially with limited resources (Dong et al., 2015). There are big-tech companies such as Google[1] and Bing[2], which have taken initiatives to build translation systems for multiple languages. Still, languages that are low-resource in nature, such as the Sorbian family of languages (Howson, 2017), have gotten lesser attention in terms of research. The paper focuses on the development of machine translation systems between pairs of languages from German, Lower Sorbian, and Upper Sorbian, using both supervised and unsupervised approaches.

In the Indo-European language family, German (Deutsch) belongs to the western Germanic branch (Durrell, 2006). Approximately 95 million people speak it natively; 28 million speak it as a second language in more than 40 countries. Due to a phonetic mutation called High German Consonant Shift (Vennemann, 2008), German moved away from other Germanic languages. This shift in German consonants occurred between the 3rd and 5th centuries, and probably ended in the 9th century AD.

Lower Sorbian (dolnoserbska rěc) and Upper Sorbian (hornjoserbska rěč) are western Slavonic languages spoken in the region of Lower and Upper Lusatia in the southeast of Germany respectively. They are closely related to other West Slavonic languages, including Polish, Czech, Slovak, and Kashubian. There are seven recognized autochthonous minorities and regional languages in Germany, including Danish, Saterfrisian, North Frisian, Romanes, and Lower German.

We aimed to carry out research in neural machine translation between German, which is high-resource in nature, and Lower and Upper Sorbian, which are low-resource languages. We implement supervised and unsupervised methods for the same. For the supervised approach, we trained transformer (Vaswani et al., 2017) models from scratch using the bilingual data provided by WMT in the 2022 workshop edition. We used the Fairseq[3] (Ott et al., 2019) library for the same, which is a sequence-to-sequence learning toolkit for neural

* equal contribution
† equal contribution
‡ equal contribution
[1]https://translate.google.com/
[2]https://www.bing.com/translator
[3]https://fairseq.readthedocs.io/en/latest/

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 1104 - 1110

Supervised	de-dsb	dsb-de	dsb-hsb	hsb-dsb	de-hsb	hsb-de
Parallel	40,194	40,194	62,565	62,565	70,000	70,000

Table 1: Statistics of the dataset used for supervised training

Unsupervised	de	dsb	hsb
Monolingual	53,309	1,45,198	2,22,027

Table 2: Statistics of the dataset used for unsupervised training

machine translation. Transformer(Vaswani et al., 2017) is a Seq2Seq (Sutskever et al., 2014a) model that uses self-attention to train on input data. The encoder part of the transformer model consists of a self-attention layer and a feed forward neural network (Bebis and Georgiopoulos, 1994). The encoder of the transformer reads the input sequence, one word at a time to produce a hidden vector. The decoder produces the output sequence from the vector received from the encoder. Being part of recent NMT research, transformers perform well compared to baseline models such as CNNs (Albawi et al., 2017) and LSTMs (Hochreiter and Schmidhuber, 1997).

For the unsupervised approach, implemented Facebook XLM's[4] masked language model (cross-lingual language model) for unsupervised learning (Chronopoulou et al., 2021). Training data used for the same was monolingual data provided by the organizers and the OPUS project[5]. We preprocessed the data, and also applied byte-pair encoding (BPE) (Sennrich et al., 2016) to the input and target data. We made use of the fastBPE[6] library for the same. Finally, we applied XLM preprocessing on the data before training.

We experimented our approaches on six language pairs between German, Lower Sorbian and Upper Sorbian. We have used the BLEU (Papineni et al., 2002) and chrF (Popović, 2015) evaluation metrics for computing accuracy, which have been discussed in this paper.

2 Dataset Description

We used the data provided by the WMT22 organizers, and from the OPUS project, recommended by the organizers. The statistics of the training data for both supervised and unsupervised approaches is given in Table 2. For the supervised training, we used the parallel data provided by the organizers, for each language pair. We used the 2022 version of the data itself, as training for larger corpora was proving computationally expensive at our end. For translations between German and Lower Sorbian, validation data size was 1353, whereas for Upper Sorbian-Lower Sorbian and German-Upper Sorbian, validation data sizes were 709 and 2000 for each language respectively.

For unsupervised learning, we used the monolingual data for Lower Sorbian provided by the organizers. For Upper Sorbian, we used the monolingual data provided by the Witaj Sprachzentrum[7]. Whereas for German, we used the monolingual data provided by the OPUS project. The quantitative statistics of these datasets is given in Table 2.

The blind test data provided by the organizers contained 1000 sentences each for translations between Lower Sorbian and German, 1000 sentences each for translations between Lower Sorbian and Upper Sorbian, and 1621 sentences each for translations between Upper Sorbian and German. We submitted the inferences on the blind test data to the shared task leaderboard.

3 Data Preparation

The data preprocessing step was crucial in determining the accuracies of our translations. The goal was not to waste resources (compute power, time) in processing things that don't add much value to extracting the semantics and understanding the text. (Tabassum and Patil, 2020)

For supervised learning, we preprocessed the source and target text using fairseq-preprocess[8], an inbuilt preprocessing script provided by Fairseq. We set the number of parallel workers for preprocessing the text as 20, so as to achieve faster preprocessing. Normalization (Mansfield et al., 2019) and pre-tokenization of the text is done before passing the to fairseq-preprocess. We used sacremoses[9]

[4]https://github.com/facebookresearch/XLM
[5]https://opus.nlpl.eu/
[6]https://github.com/glample/fastBPE
[7]https://www.witaj-sprachzentrum.de/
[8]https://fairseq.readthedocs.io/en/latest/command_line_tools.html
[9]https://github.com/alvations/sacremoses

tokenizer, which helps us to tokenize and normalize text according to our needs. fairseq-preprocess binarizes the training data and builds vocabularies from the text of that particular language.

For unsupervised learning, we applied some additional preprocessing, which consisted of using XLM-Moses tokenizer. The XLM-Moses tokenizer performs the following steps: removing unicode punctuations, normalizing punctuations (punctuations will be removed from the utterances during training) and removing any non-printing characters. We also applied byte-pair encoding (Sennrich et al., 2016) to our data, where we use the inbuilt script provided by fastBPE[10]. Byte-pair encoding algorithm computes the unique set of words used in the corpus (after the normalization and pretokenization steps are completed), and then builds the vocabulary by taking all the symbols used to write those words. Byte-pair encoding algorithm application includes learning the BPE codes from the training dataset, then applying the same on the training, validation and test datasets. Also, we get the training vocabulary once we have obtained the codes after training. Finally, we apply XLM preprocessing provided by Facebook XLM, to get the final preprocessed data.

4 Model Description

4.1 Supervised Training

We trained transformer models from scratch using the 'transformer' architecture provided by open-source toolkit Fairseq. Fairseq provides multiple state-of-the-art architectures to build translation models. Transformers use self-attention along with an encoder-decoder approach to train (Sutskever et al., 2014b). Encoders extract features from input sentences, and decoders use those features to produce output translations. The encoder in the transformer consists of multiple encoder blocks. Input sentences pass through encoder blocks, and the outputs of the last encoder block become the inputs to the transformer decoder. The decoder also consists of multiple decoder blocks, and feature information is received from the encoder by each block of the decoder.

4.2 Unsupervised Training

For unsupervised training, a masked language model (MLM) is implemented using data from both languages (source and target). We use the XLM model for easier implementation of the MLM objective. Masked prediction is implemented during the training steps, along with denoising autoencoding (Vincent et al., 2008), which involves reconstructing the original text data from a corresponding noisy version. We use an encoder-decoder transformer model, consisting of 12 layers in total (6 each to encoder and decoder), and is similar to the XLM architecture. We transfer the masked language model trained encoder transformer to the aforementioned encoder-decoder translation model.

5 Experiments

5.1 Training Details

For training the models we used the fairseq, a sequence model toolkit written in Pytorch (Paszke et al., 2019) developed by Facebook Artificial Intelligence Research (FAIR) team. We trained our models on the Nvidia Tesla K80 GPU, which has a 13GB RAM capacity.

For supervised learning, we trained our models on 50 epochs, and the total training time for every model was around 2 hours. We used the Adam optimizer (Kingma and Ba, 2014) for enhancing training performance, with corresponding beta coefficients set to 0.9 and 0.98. Label smoothing rate (Paszke et al., 2019) for the model is set to 0.1 (Label smoothing encourages the model to produce a finite output, which may lead to better generalization and prevent overfitting). Clip threshold of gradients is set to 0 (Zhang et al., 2019). A dropout (Srivastava et al., 2014) of 0.2 for input features is specified in the architecture. Maximum number of tokens in a batch is set to 4096 during training. Learning rate (Igiri et al., 2021) for the model is set to 0.0005.

For the unsupervised training, we train our models on 20 epochs, taking about five hours to train. Input words to the model are randomly shuffled during training, 3 at a time (Malkin et al., 2021). A word dropout of 0.1 is specified. 8 attention heads are taken in each layer of the encoder. Overall dropout and attention dropout of 0.1 is specified. 1000 tokens are taken per batch, and a batch-size of 32 is taken for training the models. Dimension of the embedding layer in the model was set to 1024. Sequence length of 256 is specified during training. We use the GeLU activation function (Hendrycks and Gimpel, 2016) in this model, instead of the typi-

[10]https://github.com/glample/fastBPE

Training Approach	de-dsb		dsb-de		dsb-hsb		hsb-dsb		de-hsb		hsb-de	
	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF
Supervised	20.8	44.1	25.4	51.3	49.1	65.5	50.7	66.9	25.7	49.1	29.7	53.8
Unsupervised	0.2	8.1	0.1	5.0	10.4	48.6	9.3	44.2	0.5	14.3	0.3	13.6

Table 3: Scores received on the translations obtained by performing supervised and unsupervised approaches for the WMT22-Unsupervised and Very Low Resource Supervised Task (de: German, dsb: Lower Sorbian, hsb: Upper Sorbian).

cal ReLU function used. Here too, Adam optimizer was used, with corresponding beta coefficients set to 0.9 and 0.98.

5.2 Evaluation Metrics

The BLEU (Papineni et al., 2002) and chrF (Popović, 2015) metrics were used for evaluation of the generated translations. The same metrics were used for evaluation on the shared task leaderboard.

BLEU stands for Bilingual Language Understudy. BLEU algorithm is used to evaluate machine translation quality. BLEU metric is language independent, and is easy to understand and compute. Higher the BLEU, better are the translations.

chrF stands for "character n-gram F-score". Informally, it measures the amount of overlap of short sequences of characters (n-grams) between the MT output and the reference. According to (Mathur et al., 2020) , chrF is "is technically the macro-average of n-gram statistics over the entire test set".

6 Results

For the results, please refer to Table 3. Table contains the BLEU and chrF scores to the translations that we obtained on all six language pairs (de-dsb, dsb-de, dsb-hsb, hsb-dsb, de-hsb, hsb-de), by both supervised and unsupervised approaches. These scores are obtained from our submissions to the leaderboard for the Unsup-Very Low Sup Shared task.

7 Related Work

Machine translation has been a pivotal field of research in the natural language processing domain. With rule-based and statistical machine translation methods proposed in the past decades, neural machine translation has surpassed these conventional methods by achieving state-of-the-art accuracies with each year. In 2014, Bahdanau (Bahdanau et al., 2015) proposed the base paper for neural machine translation. According to the paper, the encoder part of the model encodes the input sentence into a fixed length vector, from which the decoder generates the translation. The encoder and decoder parts could be neural architectures such as a simple RNN, LSTM (Sherstinsky, 2020), Bidirectional RNN (Schuster and Paliwal, 1997), GRU (Chung et al., 2014), etc. With the introduction of transformers (Vaswani et al., 2017) and self-attention in training neural networks, NMT research got a substantial boost.

German, being pretty high resource in nature; there has been significant work carried out in German in NMT. The Workshop on Machine Translation (WMT) has a significant contribution to the same. Minh-Thang Luong (Luong et al., 2015) demonstrate two seperate attention mechanisms (global and local attention) for bidirectional translations between English and German, gaining an increase of 5.0 in the BLEU score over non-attention based techniques. Macketanz (Macketanz et al., 2021) present the result of applying a fine-grained test suite on the outputs of 36 state-of-the-art machine translation systems between English and German, which were submitted to the Sixth Conference on Machine Translation. Xu (Xu et al., 2021) proposed BiBERT, a bilingual BERT model which helped in achieving state-of-the-art translation performance compared to other published papers till date, and that too without implementing backtranslation (Edunov et al., 2018). The paper also proposes a stoicastic layer selection method which helps in improving translation performance.

Sorbian family of languages have started receiving attention with regards to NMT research in the past few years. Li and team (Li et al., 2020) worked on supervised machine translation for a few language pairs, which included German-Upper Sorbian translations. They experimented with document-enhanced NMT, XLM pretrained language model enhanced NMT, etc. Their primary submissions won the first place in the German to Upper Sorbian Translation directions.

Knowles and team (Knowles et al., 2020) worked on implementing ensemble learning in transformer models for German-Upper Sorbian, built using BPE-dropout, lexical modifications and backtranslation.

Pertaining to unsupervised learning: Chronopoulou (Chronopoulou et al., 2020) propose the LMU Munich System for the WMT 2020 Unsupervised Machine Translation task, which involves using a pretrained monolingual model and finetuning it on both German and Upper Sorbian. Finally, the system uses backtranslation, and uses the pseudo-parallel data obtained to finetune the model further. Finally, the paper ensembles the best best-performing systems and give state-of-the-art scores on unsupervised translations between German and Upper Sorbian. Edman (Edman et al., 2021) implement transformer encoder-decoder architectures for unsupervised NMT from German to Lower Sorbian. The system has three modifications from the conventional methodology- training followes a bilingual approach, instead of a multilingual system approach. Secondly, a novel method is introduced for building the vocabulary of an unseen language. Finally, experimentation is done with the order of implementation of online and offline backtranslation. The paper received first place in the Unsupervised Machine Translation Task for WMT 2021.

8 Conclusion

Thus, we have implemented supervised and unsupervised neural machine translation approaches for translation between language pairs consisting of German(de), Lower Sorbian(dsb), and Upper Sorbian(hsb). We utilized different architectures for implementing the same. Our future plans include training these models with much larger corpora on computationally-efficient machines to obtain better evaluation metric scores and use high-end GPUs for practical training. We plan to use better preprocessing techniques and linguistic methods to improve the usefulness of the final training data to be fed to the models. We plan to implement backtranslation to improve current translation accuracy, have longer pre-training, and implement other pretrained models such as mBERT and XLM.

Acknowledgements

We are grateful to Dr. Geetanjali Kale and Dr. P. T. Kulkarni for their constant guidance.

We also thank SCTR's Pune Center for Analytics with Intelligent Learning for Multimedia Data for their continuous support.

References

Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi. 2017. Understanding of a convolutional neural network. In *2017 International Conference on Engineering and Technology (ICET)*, pages 1–6.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

G. Bebis and M. Georgiopoulos. 1994. Feed-forward neural networks. *IEEE Potentials*, 13(4):27–31.

Alexandra Chronopoulou, Dario Stojanovski, and Alexander Fraser. 2021. Improving the lexical ability of pretrained language models for unsupervised neural machine translation. In *NAACL-HLT*, pages 173–180.

Alexandra Chronopoulou, Dario Stojanovski, Viktor Hangya, and Alexander Fraser. 2020. The LMU Munich System for the WMT 2020 Unsupervised Machine Translation Shared Task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1084–1091, Online. Association for Computational Linguistics.

Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*.

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.

M. Durrell. 2006. *Germanic Languages*, pages 53–55.

Lukas Edman, Ahmet Üstün, Antonio Toral, and Gertjan van Noord. 2021. Unsupervised translation of German–Lower Sorbian: Exploring training and novel transfer methods on a low-resource language. In *Proceedings of the Sixth Conference on Machine Translation*, pages 982–988, Online. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*,

pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Chris Chinenye Emezue and Femi Pancrace Bonaventure Dossou. 2020. FFR v1.1: Fon-French neural machine translation. In *Proceedings of the The Fourth Widening Natural Language Processing Workshop*, pages 83–87, Seattle, USA. Association for Computational Linguistics.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus).

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9:1735–80.

Phil Howson. 2017. Upper sorbian. *Journal of the International Phonetic Association*, 47(3):359–367.

Chinwe Igiri, Anyama Uzoma, and Abasiama Silas. 2021. Effect of learning rate on artificial neural network in machine learning. *International Journal of Engineering Research*, 4.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

Rebecca Knowles, Samuel Larkin, Darlene Stewart, and Patrick Littell. 2020. NRC systems for low resource German-Upper Sorbian machine translation 2020: Transfer learning with lexical modifications. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1112–1122, Online. Association for Computational Linguistics.

Bin Li, Yixuan Weng, Fei Xia, and Hanjun Deng. 2022. Towards better chinese-centric neural machine translation for low-resource languages. *ArXiv*, abs/2204.04344.

Zuchao Li, Hai Zhao, Rui Wang, Kehai Chen, Masao Utiyama, and Eiichiro Sumita. 2020. SJTU-NICT's supervised and unsupervised neural machine translation systems for the WMT20 news translation task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 218–229, Online. Association for Computational Linguistics.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Vivien Macketanz, Eleftherios Avramidis, Shushen Manakhimova, and Sebastian Möller. 2021. Linguistic evaluation for the 2021 state-of-the-art machine translation systems for German to English and English to German. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1059–1073, Online. Association for Computational Linguistics.

Nikolay Malkin, Sameera Lanka, Pranav Goel, and Nebojsa Jojic. 2021. Studying word order through iterative shuffling.

Courtney Mansfield, Ming Sun, Yuzong Liu, Ankur Gandhe, and Björn Hoffmeister. 2019. Neural text normalization with subword units. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, pages 190–196, Minneapolis, Minnesota. Association for Computational Linguistics.

Nitika Mathur, Tim Baldwin, and Trevor Cohn. 2020. Tangled up in bleu: Reevaluating the evaluation of automatic machine translation evaluation metrics.

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 (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

M. Schuster and K.K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

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

Alex Sherstinsky. 2020. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Physica D: Nonlinear Phenomena*, 404:132306.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014a. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112, Cambridge, MA, USA. MIT Press.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014b. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Ayisha Tabassum and Dr. Rajendra R. Patil. 2020. A survey on text pre-processing & feature extraction techniques in natural language processing.

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.

Theo Vennemann. 2008. Lombards and consonant shift: A unified account of the high germanic consonant shift. pages 213–256.

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*, ICML '08, page 1096–1103, New York, NY, USA. Association for Computing Machinery.

Haoran Xu, Benjamin Durme, and Kenton Murray. 2021. Bert, mbert, or bibert? a study on contextualized embeddings for neural machine translation. pages 6663–6675.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. 2019. Why gradient clipping accelerates training: A theoretical justification for adaptivity.