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

Aditya Vyawahare* Rahul Tangsali*
aditya.vyawahare07@gmail.com rahuul2001@gmail.com
Aditya Mandke† Onkar Litake† Dipali Kadam‡
amandke@ucsd.edu olitake@ucsd.edu 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 Google1 and Bing2, 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 rˇec) and Upper Sorbian (hornjoserbska rˇeˇc) 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 Fairseq3 library for the same, which is a sequence-to-sequence learning toolkit for neural

1https://translate.google.com/
2https://www.bing.com/translator
3https://fairseq.readthedocs.io/en/latest/
Table 1: Statistics of the dataset used for supervised training

<table>
<thead>
<tr>
<th>Supervised</th>
<th>de-dsb</th>
<th>dsb-de</th>
<th>dsb-hsb</th>
<th>hsb-dsb</th>
<th>de-hsb</th>
<th>hsb-de</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parallel</td>
<td>40,194</td>
<td>40,194</td>
<td>62,565</td>
<td>62,565</td>
<td>70,000</td>
<td>70,000</td>
</tr>
</tbody>
</table>

Table 2: Statistics of the dataset used for unsupervised training

<table>
<thead>
<tr>
<th>Unsupervised</th>
<th>de</th>
<th>dsb</th>
<th>hsb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monolingual</td>
<td>53,309</td>
<td>1,45,198</td>
<td>2,22,027</td>
</tr>
</tbody>
</table>

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 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 project5. 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 fastBPE6 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.

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-preprocess8, 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 sacremoses9
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. Byte-pair encoding algorithm computes the unique set of words used in the corpus (after the normalization and pre-tokenization 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 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 50 epochs, and the total training time for every model was around 5 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-
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).

<table>
<thead>
<tr>
<th>Training Approach</th>
<th>de-dsb</th>
<th>dsb-de</th>
<th>dsb-hsb</th>
<th>hsb-dsb</th>
<th>de-hsb</th>
<th>hsb-de</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised</td>
<td>BLEU: 20.8 chrF: 44.1</td>
<td>BLEU: 25.4 chrF: 51.3</td>
<td>BLEU: 49.1 chrF: 65.5</td>
<td>BLEU: 50.7 chrF: 66.9</td>
<td>BLEU: 25.7 chrF: 49.1</td>
<td>BLEU: 29.7 chrF: 53.8</td>
</tr>
<tr>
<td>Unsupervised</td>
<td>BLEU: 0.2 chrF: 8.1</td>
<td>BLEU: 0.1 chrF: 5.0</td>
<td>BLEU: 10.4 chrF: 48.6</td>
<td>BLEU: 9.3   chrF: 44.2</td>
<td>BLEU: 0.5 chrF: 14.3</td>
<td>BLEU: 0.3 chrF: 13.6</td>
</tr>
</tbody>
</table>

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 separate 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 stochastic 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-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 follows 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 pre-processing techniques and linguistic methods to improve the usefulness of the final training data to be fed to the models. We plan to implement back-translation to improve current translation accuracy, have longer pre-training, and implement other pretrained models such as mBERT and XLM.

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

References


