The NITS-CNLP System for the Unsupervised MT Task at WMT 2020

Salam Michael Singh **Thoudam Doren Singh** **Sivaji Bandyopadhyay**
Center for Natural Language Processing (CNLP) and Dept. of Computer Science & Engg.
National Institute of Technology Silchar, India
{salammichaelcse,thoudam.doren,sivaji.cse.ju}@gmail.com

Abstract

We describe NITS-CNLP's submission to WMT 2020 unsupervised machine translation shared task for German language (de) to Upper Sorbian (hsb) in a constrained setting i.e, using only the data provided by the organizers. We train our unsupervised model using monolingual data from both the languages by jointly pre-training the encoder and decoder and fine-tune using backtranslation loss. The final model uses the source side (de) monolingual data and the target side (hsb) synthetic data as a pseudo-parallel data to train a pseudo-supervised system which is tuned using the provided development set(dev set).

1 Introduction

This paper provides the system description of the unsupervised neural machine translation system for German to Upper Sorbian submitted by the Center for Natural Language Processing of National Institute of Technology, Silchar, India (NITS-CNLP) in the WMT 2020 shared task for Unsupervised and Very Low Resource machine translation for German and Upper-Sorbian language pair. Specifically, we made our primary submission for the unsupervised task in $de \rightarrow hsb$ direction. We use the data provided by the organisers only i.e, in a constrained manner. Our unsupervised neural machine translation (UNMT) system first pre-trains a transformer (Vaswani et al., 2017) based encoder and decoder model using masked sequence to sequence (MASS) pre-training (Song et al., 2019) and fine-tune using the back-translation (Sennrich et al., 2016a) loss. The final model trained using MASS objective is then used to translate the source side (M_{de}) monolingual data into a synthetic target side data ($M^{'}_{hsb}$)and then train a pseudo-supervised model using $\{M_{de}$,$M^{'}_{hsb}\}$ from scratch.

The remaining of the paper is arranged in following manner: Section 2 gives a brief background of an unsupervised MT. Section 3 describes the data preprocessing. In Section 4, we describe our UNMT system. The results and analysis are shown in Section 5. Finally, Section 6 concludes the paper.

2 Background

NMT (Kalchbrenner and Blunsom, 2013; Cho et al., 2014; Bahdanau et al., 2014) has become the de-facto MT system in recent times achieving near human level translation quality for many language pair however at the cost of millions of bi-text data. Unfortunately, bi-text data for many languages is scarce or non-existent. Unsupervised MT (Lample et al., 2018a; Artetxe et al., 2018b) is one of the techniques to handle the bi-text unavailability by exploiting monolingual data (Sennrich et al., 2016a). Primitive unsupervised MT first maps the monolingual data into a common cross-lingual shared vector embedding space (Conneau et al., 2017; Artetxe et al., 2017) and infer a bilingual dictionary from this shared space using adversarial training (Lample et al., 2018a) or through self learning (Artetxe et al., 2018b) and further improve the model through a combination of de-noising auto-encoder and iterative or on-the-fly back-translation. Subsequently, this principle has been applied in SMT (Lample et al., 2018b; Artetxe et al., 2018a) or a combination of NMT and SMT (Marie and Fujita, 2018; Ren et al., 2019) to further improve the unsupervised MT. However, in this work, we follow a newer approach of cross-lingual language model pretraining (Lample and Conneau, 2019; Song et al., 2019) which has shown to be a stronger initialization for unsupervised MT than the cross-lingual shared vector embedding space.

3 Data and Preprocessing

This section is further divided into two subsections briefing the data description and the preprocessing steps used.

Proceedings of the 5th Conference on Machine Translation (WMT), pages 1139–1143

Corpus		Sentences
mono	de (News Crawl)	5 **M**
	hsb	756.3 **K**
dev/test	de	2 **K**
	hsb	2 **K**

Table 1: Statistics of the monolingual and the dev/test set.

3.1 Data Description

We use a randomly sampled 5**M** monolingual corpus for German side from News Crawl[1] dataset, while we use all the available monolingual data[2] and the parallel side[3] of Upper Sorbian[4] as the combined monolingual data for the same and summing up 756,271 number of sentences. For tuning and evaluation[5], we use the provided devtest[6] data with 2000 sentences for both the dev and test files as shown in Table 1.

3.2 Preprocessing

We use `Moses`(Koehn et al., 2007) toolkit for preprocessing the data. The corpus underwent removal of non-printing characters and tokenization. For the Upper Sorbian, we used Czech `(cs)` language code for tokenization as Upper Sorbian `(hsb)` language code is unavailable in `Moses` toolkit[7] and considering the relatedness of these languages[8].

The above preprocessing is used by MASS pretrain and MASS finetune models while the pseudo-supervised model uses the raw data and learns a Sentencepiece BPE. The details are described in Section 4.2.

4 UNMT System

Our UNMT system is a pipeline of encoder-decoder pretraining and fine-tuning using MASS (Song et al., 2019) and using the synthetic data generated (M'_{hsb}) from the source monolingual data (M_{de}) to train a forward model from scratch. This section is further divided into two subsections, first describing the MASS pretraining and fine-tuning and second, the transformer based forward ($\overrightarrow{f}$) pseudo-supervised model using the pseudo-parallel ($\{M_{de}, M'_{hsb}\}$) data by inducing Lample et al. (2018a) style noise (word drop, word shuffle and word blank) upon the input data.

```
--mass_steps 'de,hsb'
--encoder_only false
--emb_dim 1024 --n_layers 6
--n_heads 8 --dropout 0.1
--attention_dropout 0.1
--gelu_activation true
--tokens_per_batch 3000
--optimizer adam_inverse_sqrt,
  beta1=0.9,beta2=0.98,lr=0.0001
--word_mass 0.5 --min_len 5
```

Table 2: `MASS` pretraining parameters

4.1 MASS Pretrain and Finetune

We use the `MASS` toolkit[9] to pretrain a cross-lingual language model using the masked sequence to sequence objective. Initially, the corpus are segmented into subword units using BPE(Sennrich et al., 2016b). A joint BPE is learnt over the monolingual data of both the languages (German and Upper Sorbian) and the vocabulary is limited to 60,000 shared vocabulary tokens.

MASS Pretraining: The BPE tokenized monolingual data is used to pretrain the encoder and decoder jointly by the cross lingual MASS objective and the training is done for 100 epochs. The parameters for the `MASS` pretraining is shown in Table 2.

MASS Fine-tuning: The pretrained model is capable to generate translations but it is merely a copy task. So, in order to make the model more robust, it is further fine-tuned using the loss objective of back-translation. The fine-tuning is halted after the 10th epoch before being converged due to resource limitation. The parameters for fine-tuning is listed in Table 3.

4.2 Pseudo-Supervised NMT

We follow Marie et al. (2019) style of using the pseudo-parallel data generated from a previous

[1] `http://data.statmt.org/news-crawl/de/`

[2] `http://www.statmt.org/wmt20/unsup_and_very_low_res/`

[3] `http://www.statmt.org/wmt20/unsup_and_very_low_res/train.hsb-de.hsb.gz`

[4] The parallel side of Upper Sorbian is allowed for Unsupervised task.

[5] We use *newstest2020* test set for the submission.

[6] `http://www.statmt.org/wmt20/unsup_and_very_low_res/devtest.tar.gz`

[7] `https://github.com/moses-smt/mosesdecoder`

[8] Both Czech and Upper Sorbian belongs to Western Slavic language branch.

[9] `https://github.com/microsoft/MASS`

```
--bt_steps 'de-hsb-de,hsb-de-hsb'
--encoder_only false
--emb_dim 1024 --n_layers 6
--n_heads 8 --dropout 0.1
--attention_dropout 0.1
--gelu_activation true
--tokens_per_batch 2000
--optimizer adam_inverse_sqrt,
  beta1=0.9,beta2=0.98,lr=0.0001
--eval_bleu true
```

Table 3: MASS finetuning parameters

model to train a forward pseudo-supervised model. In our case, we first generate a synthetic data (M'_{hsb}) from the source monolingual data (M_{de}) using beam search decoding with a beam size of 10 from the MASS fine tuned model. Unlike Marie et al. (2019) where back translation was applied, we use forward translation from the source side monolingual (He et al., 2020) data to generate synthetic data. The synthetic data is detokenized, and we learn a joint subword BPE from the raw M_{de} and M'_{hsb} using Sentencepiece (Kudo and Richardson, 2018) and limit the shared vocabulary to 10 **K** units.

Noisy Pseudo-Supervised NMT: We add perturbations or noise, specifically we apply word dropout, word shuffle and word blank to our synthetic data. This kind of perturbation is found to be effective for overcoming the local minima by enforcing local smoothness (He et al., 2020; Shen et al., 2019). We train our pseudo-supervised NMT in a pseudo self-training approach by leveraging the source side monolingual data. This self-training is partial in the sense that we only use the pseudo-parallel data which lacks any sort of real labelled data for a single iteration.

The pseudo-supervised NMT is trained from scratch using Fairseq (Ott et al., 2019) toolkit[10] i.e, we do not use the previous models weights rather we apply random weight initialization for our new model. The model is trained for 300 **K** update steps. We follow Guzmán et al. (2019) style transformer architecture of 5 encoder and decoder layers, 512 embedding dimension, the feed-forward hidden dimension is 2048 with 4 multi-head attentions[11]. The rest of the parameters are listed in Table 4. We

```
--encoder-normalize-before
--decoder-normalize-before
--dropout 0.3 --relu-dropout 0.3
--attention-dropout 0.3
--label-smoothing 0.2
--criterion label_smoothed_
  cross_entropy
--weight-decay 0.0001
--lr-scheduler inverse_sqrt
--min-lr 1e-9 --max-tokens 4000
--warmup-updates 4000
--warmup-init-lr 1e-7
--optimizer adam --lr 0.0005
--adam-betas '(0.9, 0.98)'
--share-all-embeddings
```

Table 4: Pseudo-supervised NMT training parameters

make our primary submission of the test source generated using a beam search decoding with beam size of 5 and a length penalty of 1.2.

5 Result

The official automatic evaluation uses the the following metrics: BLEU (Papineni et al., 2002), TER (Snover et al., 2006), BEER (Stanojević and Sima'an, 2014), and CharactTER (Wang et al., 2016). Our primary submission (NITS-CNLP), the pseudo-supervised NMT achieves a cased BLEU of 15.4 and 15.8 as the uncased BLEU score on the *newstest2020* blind-test data. The scores are reported in Table 5. We also present the sample input-output of our primary system (NITS-CNLP) from two randomly selected test sentences from the matrix[12] in Table 6. We also report the Sacrebleu score of our various settings with the released test set (non blind test) in Table 7.

6 Conclusion

We report here the system description for our submission to the WMT 2020 shared task of Unsupervised MT for German-Upper Sorbian language pair. We submit our pipelined architecture of masked sequence to sequence pretraining along with finetuning and a pseudo-supervised model in German to Upper Sorbian direction. We observe that the performance of an unsupervised model improves significantly over the base MASS pretraining and

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

[11] We have used 4 attention heads instead of 8 as in Guzmán et al. (2019)

[12] http://matrix.statmt.org/matrix/output/1920?run_id=7785

System	BLEU	BLEU-cased	TER	BEER 2.0	CharactTER
NITS-CNLP	15.8	15.4	0.668	0.489	0.604

Table 5: BLEU, BLEU-cased, TER, BEER 2.0 and CharactTER scores of our final primary system NITS-CNLP for the German → Upper Sorbian language using blindtest (newstest2020).

Source-1	Möchten Sie erfahren, wie sich bei uns die Unterrichtsräume mit Leben füllen?
Reference-1	Chceće wědźeć, kak so pola nas wučbne rumnosće ze žiwjenjom pjelnja?
NITS-CNLP	Časće zhonić, kak so pola nas wučbnych rumow z žiwami čuje?
Source-2	Rächt euch nicht selbst, sondern gebt Raum dem Zorn Gottes.
Reference-2	Njewjećće so sami, ale dajće městno Božemu hněwu.
NITS-CNLP	Njech wam sam, ale pomha rumnosć Božeje służby.

Table 6: Sample input-output excerpted from the matrix primary submission of NITS-CNLP.

System	BLEU
MASS-PT	2.3
MASS-FT	8.1
PSNMT	14.5

Table 7: BLEU, scores of our three systems using the released test set: MASS-pretrain (MASS-PT), MASS-finetune (MASS-FT) and Pseudo Supervised NMT (PSNMT) for German → Upper Sorbian language.

finetuning after using the synthetic data to train a pseudo-supervised model using a very naive way of self-training i.e, we have just used a single iteration of our forward training. Synthetic data is the *de-facto* for any modern semi-supervised MT system and in this experiment we show that synthetic data in an unsupervised MT is effective and also emphasised the importance of a pseudo-supervised MT model as a refinement step to an unsupervised MT.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. 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)*, pages 451–462, Vancouver, Canada. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3632–3642, Brussels, Belgium. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018b. Unsupervised neural machine translation. In *Proceedings of the Sixth International Conference on Learning Representations*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar. Association for Computational Linguistics.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Junxian He, Jiatao Gu, Jiajun Shen, and Marc'Aurelio Ranzato. 2020. Revisiting self-training for neural sequence generation. In *Proceedings of ICLR*.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709, Seattle, Washington, USA. Association for Computational Linguistics.

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

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations (ICLR)*.

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.

Benjamin Marie and Atsushi Fujita. 2018. Unsupervised neural machine translation initialized by unsupervised statistical machine translation. *arXiv preprint arXiv:1810.12703*.

Benjamin Marie, Haipeng Sun, Rui Wang, Kehai Chen, Atsushi Fujita, Masao Utiyama, and Eiichiro Sumita. 2019. NICT's unsupervised neural and statistical machine translation systems for the WMT19 news translation task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 294–301, Florence, Italy. Association for Computational Linguistics.

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.

Shuo Ren, Zhirui Zhang, Shujie Liu, Ming Zhou, and Shuai Ma. 2019. Unsupervised neural machine translation with smt as posterior regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 241–248.

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.

Jiajun Shen, Peng-Jen Chen, Matt Le, Junxian He, Jiatao Gu, Myle Ott, Michael Auli, and Marc'Aurelio Ranzato. 2019. The source-target domain mismatch problem in machine translation. *arXiv preprint arXiv:1909.13151*.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, volume 200. Cambridge, MA.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. In *International Conference on Machine Learning*, pages 5926–5936.

Miloš Stanojević and Khalil Sima'an. 2014. BEER: BEtter evaluation as ranking. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 414–419, Baltimore, Maryland, USA. Association for Computational Linguistics.

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.

Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016. Character: Translation edit rate on character level. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 505–510.