Language model pre-training and transfer learning for very low resource languages

Jyotsana Khatri[†], **Rudra Murthy V**[‡], **Pushpak Bhattacharyya**[†]
[†] Center for Indian Language Technology (CFILT)
Department of Computer Science and Engineering
IIT Bombay, India.
[‡]IBM Research, Bangalore, India.
{jyotsanak,pb}@cse.iitb.ac.in, rmurthyv@in.ibm.com

Abstract

This paper describes our submission for the shared task on Unsupervised MT and Very Low Resource Supervised MT at WMT 2021. We submitted systems for two language pairs: German ↔ Upper Sorbian (de ↔ hsb) and German ↔ Lower Sorbian (de ↔ dsb). For de ↔ hsb, we pretrain our system using MASS (Masked Sequence to Sequence) objective and then finetune using iterative back-translation. We perform final finetuning using the provided parallel data for translation objective. For de ↔ dsb, no parallel data is provided in the task, we use final de ↔ hsb model as initialization of the de ↔ dsb model and train it further using iterative back-translation, using the same vocabulary as used in the de ↔ hsb model.

1 Introduction

Transformer based architecture (Vaswani et al., 2017) has become the de-facto approach for training NMT models. These models have achieved good performance for resource rich languages. NMT models are usually data hungry and require lot of parallel data to get trained. However, many low-resource languages have very little or no parallel data to train a NMT model. For low resource language pairs, unsupervised MT (Artetxe et al., 2018; Lample et al., 2018; Lample and Conneau, 2019; Song et al., 2019), and transfer learning (Zoph et al., 2016a) have proven to be helpful in improving the translation performance. Unsupervised MT has gained a lot of attention in the past 3 years as it utilizes only monolingual data to train a NMT system.

In this paper, we present our system for shared task on Unsupervised MT and Very Low Resource Supervised MT at WMT2021. The task covers three languages pairs German (de) ↔ Lower Sorbian (dsb), German (de) ↔ Upper Sorbian (hsb), and Russian (ru) ↔ Chuvash (ch). We submitted systems for de ↔ hsb and de ↔ dsb. For de ↔ dsb there is no parallel data provided but for de ↔ hsb, there is small parallel data.

Summary of our submitted systems:

- We use language model pretraining using MASS (Song et al., 2019) objective to pretrain a model for de ↔ hsb using shared encoder, shared decoder, and shared vocabulary, which is followed by finetuning using iterative back-translation. The final model is finetuned using parallel data with translation objective.
- For de ↔ dsb, our model is trained using provided monolingual dsb and de data using iterative back-translation. The model is initialized using the final model of de ↔ hsb.

2 Related Work

Supervised NMT using transformer based architectures (Vaswani et al., 2017) has achieved high translation accuracy for high resource languages like English-French and English-German. Supervised NMT requires lots of parallel data to get trained. For low resource languages (which does not have large amount of parallel data) the performance of NMT systems is usually poor. We briefly describe some literature on Unsupervised MT and transfer learning.

2.1 Unsupervised NMT

Unsupervised MT gained quite a lot of attention of researchers because of its ability to train MT system without using any parallel data. The research in Unsupervised MT started with techniques which are based on statistical decipherment (Ravi and Knight, 2011; Dou and Knight, 2012, 2013; Dou et al., 2014, 2015). The approaches proposed in Artetxe et al. (2018); Lample et al. (2017) are majorly based on unsupervised cross-lingual embeddings, denoising auto-encoders, and iterative

Proceedings of the Sixth Conference on Machine Translation (WMT), pages 995–998

back-translation. Later, some approaches of Unsupervised SMT have been proposed where a phrase table is constructed using bilingual embeddings and training is performed using language model and distortion model (Artetxe et al., 2019; Lample et al., 2018).

State of the art approaches of Unsupervised NMT are based on cross-lingual language model pretraining followed by iterative back-translation (Lample and Conneau, 2019; Song et al., 2019; Lewis et al., 2019). All these models differ with respect to pretraining objective. Lample and Conneau (2019) pretrain encoder and decoder separately using masked language modeling objective, while Song et al. (2019) pretrains encoder and decoder together using MASS (masked sequence to sequence) objective. Lewis et al. (2019) pretrains encoder and decoder using an objective similar to MASS but here the decoder is supposed to predict the whole sentence rather than only predicting the masked span of tokens.

2.2 Transfer Learning

Transfer learning have proven to be helpful for low resource languages (Zoph et al., 2016b; Dabre et al., 2017; Nguyen and Chiang, 2017). In Zoph et al. (2016b), authors use a model trained on one language pair as the initialization of the model for another language pair, they do not consider or do anything with the vocabulary. Gheini and May (2019) proposed to create a universal vocabulary before starting the training of the parent model. The transfer learning works best if the language pairs are related (Dabre et al., 2017). Aji et al. (2020) shows that the internal layers are most important in transfer learning.

3 System Overview

In this section, we describe the details of the submitted systems to shared task on Unsupervised MT and Very Low Resource Supervised MT at WMT 2021. We report results for our 2 types of models:

- **Language model pretraining using MASS objective**: For de ↔ hsb, we pretrain our model using MASS objective and then finetune it using iterative back-translation. Final finetuning is performed using parallel data of de ↔ hsb provided in the task.
- **Transfer learning**: For de ↔ dsb, we use the final model of de ↔ hsb to initialize the model of de ↔ dsb and train it further using iterative back-translation using monolingual data of de and dsb.

To train our models, we use shared encoder-decoder transformer architecture. We also use shared vocabulary of both source and target languages. For de ↔ dsb, we use the same vocabulary as used in de ↔ hsb model without considering the vocabulary mismatch.

4 Experiments

In this section, we describe the experimental setup and the hyper-parameters used.

4.1 Data and Preprocessing

For de ↔ hsb, we use monolingual data of hsb provided in the task and we use a subset (equal to the size of the hsb monolingual data) of news-crawl-2020 dataset downloaded from WMT[1] provided in the WMT news translation task for de monolingual data, and also use the parallel data provided in the task. For de ↔ dsb, we use monolingual data of dsb provided in the task together with a subset (equal to the size of the dsb data) of news-crawl-2020 dataset provided in WMT news translation task for de monolingual data.

We tokenize using Moses tokenizer (Koehn et al., 2007). We use fastBPE[2] to learn BPE (Byte pair encoding) (Bojanowski et al., 2017) with 32k BPE codes over the combined tokenized data of both languages. For de ↔ dsb, we use the same vocabulary and codes learnt for de ↔ hsb.

4.2 Experimental Setup

We use 6 layers in the encoder and decoder with 8 attention heads and 1024 embedding dimension. We use Adam (Kingma and Ba, 2015) optimizer. We use, a warm-up phase of 4000 steps with initial learning rate starting from $1e^{-7}$ to $1e^{-4}$, in the warm-up phase learning rate is increased linearly and then starts to decrease with inverse square root learning rate schedule. We use mini-batches of size 2000 tokens and set the dropout to 0.1 (Gal and Ghahramani, 2016). Maximum sentence length is set to 100 after applying BPE. At the time of decoding, we set beam size to 1. For experiments, we are using MASS[3] codebase.

[1] http://statmt.org/wmt21/translation-task.html

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

[3] https://github.com/microsoft/MASS

The pretraining is performed for 100 epochs for both de ↔ hsb. de ↔ hsb model is further finetuned using iterative back-translation for 60 epochs and then trained using parallel data for 60 epochs. de ↔ dsb model is further finetuned for iterative back-translation using the final de ↔ hsb model for 60 epochs. Epoch size is set to .2M sentences.

4.3 Results and Discussion

Lang Pair	Train	Valid	Test
de-hsb	147521	2000	2000
de-dsb	0	601	602

Table 1: Parallel data (Number of sentences)

Language	Train
hsb	695721
dsb	145198

Table 2: Monolingual data (Number of sentences) (We use equal amount of german data from news-crawl2020 as dsb and hsb to train their respective models)

Lang Pair	Our system	Best system
de-hsb	60.2	66.3
hsb-de	60.1	67.7
de-dsb	6.4	29.9
dsb-de	5.9	33.5

Table 3: Results: BLEU scores for our system and highest scoring system in the task

All the results are shown in 3. We achieve BLEU score of 60.2 and 60.1 for de → hsb and hsb → de respectively. Using the final model of de ↔ hsb as initialization of the model for de ↔ dsb, we achieve BLEU score of 6.4 and 5.9 for de → dsb and dsb → de respectively even with using the same vocabulary of de ↔ hsb. The percentage of vocabulary overlap (the percentage of de ↔ dsb vocabulary that is present in de ↔ hsb vocabulary) is 68.21 after applying BPE which makes the transfer learning work. After MASS pretraining and iterative back-translation (without using any parallel data), the BLEU scores are 4.74 and 4.92 for de → hsb and hsb → de respectively. We are able to achieve above BLEU scores without using any parallel data because of the similarity between de and hsb. The percentage of vocabulary overlap between de and hsb (the percentage of vocabulary of de present in hsb) is 60.73, which makes them highly similar languages. Similarly, the percentage of vocabulary overlap between de and dsb (the percentage of vocabulary of de present in dsb) is 54.26. The vocabulary here refers to the number of unique tokens after applying BPE.

5 Conclusion

In this paper, we study the impact of language model pretraining together with iterative back-translation for very low resource language pair i.e. de ↔ hsb. We also study the impact of transfer learning from de ↔ hsb to de ↔ dsb. In future, we plan to filter bad back-translated data while training for de ↔ dsb using iterative back-translation and also different transfer learning techniques to improve the performance for de ↔ dsb.

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.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. An effective approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. *ArXiv*, abs/1710.11041.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

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.

Qing Dou and Kevin Knight. 2012. Large scale decipherment for out-of-domain machine translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages

266–275, Jeju Island, Korea. Association for Computational Linguistics.

Qing Dou and Kevin Knight. 2013. Dependency-based decipherment for resource-limited machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1668–1676, Seattle, Washington, USA. Association for Computational Linguistics.

Qing Dou, Ashish Vaswani, and Kevin Knight. 2014. Beyond parallel data: Joint word alignment and decipherment improves machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 557–565.

Qing Dou, Ashish Vaswani, Kevin Knight, and Chris Dyer. 2015. Unifying Bayesian inference and vector space models for improved decipherment. 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 836–845, Beijing, China. Association for Computational Linguistics.

Yarin Gal and Zoubin Ghahramani. 2016. A theoretically grounded application of dropout in recurrent neural networks. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS, pages 1027–1035, Red Hook, NY, USA. Curran Associates Inc.

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

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 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.

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

Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. 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.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

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.

Sujith Ravi and Kevin Knight. 2011. Deciphering foreign language. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 12–21.

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.

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.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016a. Transfer learning for low-resource neural machine translation. *ArXiv*, abs/1604.02201.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016b. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575.