The AIC System for the WMT 2022 Unsupervised MT and Very Low Resource Supervised MT Task

Ahmad Shapiro, Mahmoud Tarek, Omar Khaled
Mohamed Fayed, Ayman Khalafallah, Noha Adly
Applied Innovation Center
{ahmad.shapiro, mahmoudtarek, omar.khaled, m.essam, a.khalafallah, nadly}@aic.gov.eg

Abstract

This paper presents our submissions to WMT 22 shared task in the Unsupervised and Very Low Resource Supervised Machine Translation tasks. The task revolves around translating between German ↔ Upper Sorbian (de ↔ hsb), German ↔ Lower Sorbian (de ↔ dsb) and Upper Sorbian ↔ Lower Sorbian (hsb ↔ dsb) in a both unsupervised and supervised manner. For the unsupervised system, we trained an unsupervised phrase-based statistical machine translation (UPBSMT) system on each pair independently. We pretrained a German-Slavic mBART model on the following languages Polish (pl), Czech (cs), German (de), Upper Sorbian (hsb), and Lower Sorbian (dsb). We then fine-tuned our mBART on the synthetic parallel data generated by the (UPBSMT) model along with authentic parallel data (de ↔ pl, de ↔ cs). We further fine-tuned our unsupervised system on authentic parallel data (hsb ↔ dsb, de ↔ dsb, de ↔ hsb) to submit our supervised low-resource system.

1 Introduction

Like most machine learning approaches, data can be considered the most important component of the recipe for modeling a solution for a given problem. Neural machine translation relies heavily on a large amount of training data to correctly model two languages and learn the mapping between them to produce semantically and syntactically right translations. However, machine translation is not available for the majority of the 7000 languages spoken on the earth. This is due to the fact that parallel corpora are scarce or non-existent. There have been several proposals to alleviate the issue of small amounts of parallel data such as pivot translation, multilingual training, and semi-supervised training which resulted in an acceptable performance.

Unsupervised machine translation became the go-to solution when lacking parallel data. The WMT 2022 Unsupervised MT Task focuses on two very low-resource languages: Upper Sorbian (HSB) and Lower Sorbian (DSB). Upper and Lower Sorbian are minority languages spoken in the federal states of Saxony and Brandenburg in Eastern Germany. With just 30,000 and 7,000 native speakers, working on these languages is an extreme low-resource task, with little prospect of ever approaching the number of resources available for languages with millions of speakers. However, because they are western Slavic languages, the Sorbian languages can benefit from Czech and Polish data (Libovický and Fraser, 2021).

In this paper, we describe our systems for translating between German ↔ Upper Sorbian (de ↔ hsb), German ↔ Lower Sorbian (de ↔ dsb), and Upper Sorbian ↔ Lower Sorbian (hsb ↔ dsb) in a both unsupervised and supervised manner.

We approach the task by combining two novel approaches for unsupervised machine translation. Influenced by (Artetxe et al., 2019), we start by developing unsupervised phrase-based statistical machine translation systems (UPBSMT) for all language pairs independently. In contrast to (Artetxe et al., 2019), (Lample and Conneau, 2019) relies on pre-training an XLM model on the source and target language to capture the translation signal instead of using (UPBSMT). Instead, we benefit from both the pre-training and UPBSMT. So, we pre-train an mBART model (Liu et al., 2020) on Polish, Czech, Upper Sorbian, Lower Sorbian, and German from scratch as we mentioned earlier that pl and cs are similar to dsb and hsb. We then fine-tune mBART on synthetic parallel data (de),(de ↔ hsb) and (hsb ↔ dsb) along with authentic parallel data (de ↔ pl, de ↔ cs).

We group pl, cs, dsb, hsb under one token $slavic$ while feeding it to the encoder. For our low-resource submission, we fine-tune the unsupervised model on the authentic parallel data provided by the task between (de ↔ hsb),(de ↔ dsb), and (hsb ↔ dsb).

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 1117–1121
Abu Dhabi, December 7–8, 2022. ©2022 Association for Computational Linguistics

Our unsupervised approach scored the highest BLEU in all directions except (de ↔ dsb) direction.

2 Related Work

The earliest approach on Unsupervised Machine Translation was introduced by (Ravi and Knight, 2011) where they frame the MT task as a decipherment task, treating the target language as cipher text of English. Their method is essentially the same approach taken by cryptanalysts and epigraphers when they use the source texts. They started by estimating the word translation probabilities using a devised Iterative EM algorithm, due to the huge consumption of memory since they operate on large-scale vocabularies. Followed by that, they propose a novel approach based on Bayesian Decipherment that outperformed the previous EM approach in all aspects. After that they build an n-gram translation table that was used to estimate an IBM Model 3 translation model, the highest BLEU (Papineni et al., 2002) score achieved was 19.3 on the Spanish-English OPUS subtitles data.

(Artetxe et al., 2019) provide a two-step solution to unsupervised machine translation. For step one, they start by building an UPBSMT system between source and target languages. Resulting in two translation models: source-to-target and target-to-source models. Using these models, they back-translate target monolingual data using the target-to-source model to generate $(\hat{src}, trg)$ pairs that will be used to train the source-to-target neural model. Similarly, they back-translate the source monolingual using the source-to-target model to generate $(src, \hat{trg})$ pairs that will be used to train the target-to-source neural model.

The second step is training two neural models, source-to-target and target-to-source, using the synthetic data generated from step 1 using iterative back-translation. The first iteration relies solely on data generated by UPBSMT, the following iterations substitute a percentage of synthetic data generated by UPBSMT by back-translated data from the neural model in the reverse direction. Until the whole training data is back-translated from the reverse model. In contrast, (Lample and Conneau, 2019) starts by pre-training an XLM encoder on Masked Language Modeling (MLM) task on the source and target languages.

After pre-training, they initialize an encoder-decoder model using the pre-trained XLM encoder. Their training step is composed of three tasks :

1. Denoising Auto encoding.
2. Cross Domain (Back-translation).
3. Adversarial Loss.

In our work, we combine the two methods. But instead of using neural iterative back-translation, we add authentic parallel data from related languages.

3 Approach

Inspired by (Artetxe et al., 2019) and (Lample and Conneau, 2019), we adapted a mixed approach to mitigate the weaknesses and combine the advantages of both methods. (Artetxe et al., 2019) use UPBSMT as an explicit initial translation signal to train two translation models from scratch on a translation task. But, UPBSMT's output is noisy and the translation model is trained from scratch without any denoising pre-training objective. In contrast, (Lample and Conneau, 2019) pre-trains an XLM encoder on MLM task and then use it to initialize a seq2seq model which will be trained to translate in an unsupervised manner as we discussed in Section 2. Although (Lample and Conneau, 2019) didn't train the translation model from scratch, they relied solely on the three training tasks discussed earlier to capture the cross-lingual translation signal in contrast to (Artetxe et al., 2019) who used an explicit cross-lingual translation signal.

We combine the best of both worlds by using UPBSMT as our initial translation signal to fine-tune a pre-trained mBART model on a multilingual translation task.

3.1 Unsupervised Phrase-based Statistical Machine Translation

We followed (Artetxe et al., 2018) approach to build an unsupervised phrase-based statistical machine translation system between the following pairs : $(de \rightarrow dsb)$, $(de \rightarrow hsb)$, $(dsb \rightarrow de)$, $(hsb \rightarrow de)$, $(hsb \rightarrow dsb)$ and $(dsb \rightarrow hsb)$.

Using the above models, we back-translated monolingual data of $lang_1$ to $\hat{lang_2}$ which will be used to train the reverse direction model as following :

1. de translated by $(de \rightarrow dsb)$ model, producing $(\hat{dsb}, de)$ pairs to train the $(dsb \rightarrow de)$ neural direction.
2. de translated by $(de \rightarrow hsb)$ model, producing $(\hat{hsb}, de)$ pairs to train the $(hsb \rightarrow de)$ neural direction.

3. dsb translated by $(dsb \rightarrow de)$ model, producing $(\hat{de}, dsb)$ pairs to train the $(de \rightarrow dsb)$ neural direction.

4. hsb translated by $(hsb \rightarrow de)$ model, producing $(\hat{de}, hsb)$ pairs to train the $(de \rightarrow hsb)$ neural direction.

5. dsb translated by $(dsb \rightarrow hsb)$ model, producing $(\hat{hsb}, dsb)$ pairs to train the $(hsb \rightarrow dsb)$ neural direction.

6. hsb translated by $(hsb \rightarrow dsb)$ model, producing $(\hat{dsb}, hsb)$ pairs to train the $(dsb \rightarrow hsb)$ neural direction.

3.2 German-Slavic mBART pre-training

Since Lower and Upper Sorbian are West-Slavic languages, their direct cousins in the West-Slavic family tree are Polish (pl) and Czech (cs). Polish and Czech are high-resource languages with a large-scale availability of both monolingual and parallel data. We pre-trained mBART model (Liu et al., 2020) from scratch on denoising auto-encoding objective on Polish (pl), Czech (cs), Upper Sorbian (hsb), Lower Sorbian (dsb), German (de).

3.3 mBART fine-tuning

Using the generated synthetic parallel data produced from UPBSMT step discussed in Section 3.1 along with authentic $(pl \rightarrow de)$ and $(cs \rightarrow de)$ from OPUS (Tiedemann, 2012). We fine-tuned our German-Slavic mBART on translation on $(pl \rightarrow de)$, $(cs \rightarrow de)$, $(de \leftrightarrow dsb)$, $(de \leftrightarrow hsb)$, $(hsb \leftrightarrow dsb)$. Taking advantage of the similarity between (pl, cs, dsb, hsb), we grouped those languages under one language token $(slavic)$ which is fed to the encoder of our mBART. This approach constructs our unsupervised submission.
For the low-resource submission, we further fine-tuned the resulted model on authentic parallel data provided by the task.

4 Experiments

In this section, we describe our experimental setup and results. Readers can refer to our GitHub Repository [1] for training scripts, checkpoints, hyper-paramters etc.

4.1 Data Pre-processing

We follow (Artetxe et al., 2019) cleaning approach as following :

1. `normalize-punctuation.perl` script from Moses library to normalize punctuations.

2. `remove-non-printing-char.perl` script from Moses library to remove non-printing characters.

3. Tokenizing using Moses Tokenizer.

4. Deduplication.

5. Cleaning by length, with minimum and maximum of 3 and 80 words respectively.

Language	Datasets	Sentences
Polish (pl)	europarl-v10	706,047
	news-crawl 2018 to 2021	12,653,333
	Total	13,359,380
Czech (cs)	europarl v10	669,676
	news-commentry v14-16	825,841
	news-crawl 2007 to 2021	109,599,883
	Total	111,095,400
German (de)	europarl v10	2,107,971
	news-commentry v14-16	1,259,790
	news-crawl 2007 to 2021	428,057,920
	Total	431,425,681
Upper Sorbian (hsb)	Witaj (2020)	222,027
	Sorbian-Insitute (2020)	339,822
	Task Data (2022)	436,579
	Total	998,428
Lower Sorbian (dsb)	Task Data (2021)	145,198
	Task Data (2022)	66,407
	Task Data : Wiki (2022)	8,814
	Total	220,419

Table 1: Monolingual Data sets used in our experiments

4.2 Unsupervised Statistical Machine Translation Data

We use monolingual data of German, Upper Sorbian and Lower Sorbian stated in Table 1. We used a 20MILL random sample from the German monolingual data. The output of the UPBSMT is synthetic parallel data that will be used to fine-tune the pre-trained mBART on the unsupervised translation task. The number of synthetic parallel data is shown in Table 2.

[1] `https://github.com/ahmadshapiro/WMT22`

Language	Sentences
dsb → de	19,486,715
de → dsb	155,683
hsb → de	19,486,715
de → hsb	873,794
hsb → dsb	873,794
dsb → hsb	155,683

Table 2: Synthetic Parallel Data Generated by UPBSMT

4.3 mBART Pre-training

We pretrained mBART on 32 V100 GPUs from scratch for less than 2 epochs (24hrs) on all monolingual data from Table 1. We learned 32k BPE codes using SentencePiece Library (Kudo and Richardson, 2018) on the concatenation of all monolingual data. This SentencePiece model will be used for the rest of neural experiments involving mBART. The average valid perplexity for all languages reached 4.16. We decided to stop training due to the time limit. All of our neural models were developed using FairSeq Framework (Ott et al., 2019).

4.4 mBART Fine-tuning (Unsupervised Submission)

We fine-tuned our pre-trained mBART on translation task using authentic parallel data of (pl-de, cs-de) shown in Table 3 along with all synthetic parallel data shown in Table 2. We grouped (hsb, dsb, pl, cs) under one token (slavic) which is passed to mBART encoder as we discussed earlier in Section 3.3. The training was done on 27 V100 GPUs for less than 1 epoch (24hrs).

4.5 mBART Fine-tuning (Low Resource Submission)

We further fine-tuned mBART on authentic parallel task data of (hsb-de, dsb-de, hsb-dsb) shown in Table 3 for 3 epochs to submit our supervised model.

5 Results

In this section, we present our results on the blind test set of WMT 22 workshop.

5.1 Unsupervised Submission

Our approach scored the highest BLEU in all pairs except the (de ↔ dsb) directions. This can be attributed to the fact of having multiple errors in the UPBSMT experiment on this specific pair. Due to the time limit, we had to use the un-tuned/corrupted models for this pair. In contrast, (de ↔ hsb) directions models scored almost 18.0 BLEU score. Surprisingly, this can reflect the importance of the UPBSMT component in our experiments, since hsb and dsb are hugely similar. But, due to an error in the UPBSMT training, the former hugely outperformed the latter. Results are reported in Table 4.

Language	Datasets	Sentences
pl-de	DGT JRC-Acquis MultiParaCrawl EUbookshop Europarl QED	12,375,574
cs-de	DGT JRC-Acquis MultiParaCrawl EUbookshop Europarl QED	12,427,403
hsb-de	Task Data (2020)	60,000
	Task Data (2021)	87,521
	Task Data (2022)	301,536
	Total	448,787
dsb-de	Task Data (2022)	40,193
hsb-dsb	Task Data (2022)	62,564

Table 3: Authentic Parallel Data sets from OPUS (Tiedemann, 2012) used in our experiments

Direction	BLEU
dsb → de	4.0
de → dsb	1.2
hsb → de	**18.0**
de → hsb	**17.9**
hsb → dsb	**35.9**
dsb → hsb	**44.2**

Table 4: Unsupervised results on Blind Test data of WMT22

5.2 Low Resource Submission

As shown in Table 5, further fine-tuning on authentic parallel data improved BLEU score in all directions even the corrupted (de ↔ dsb) directions. Our model was constantly improving through updates, but we had to stop the training due to time constraints. We didn't use any low-resource techniques such as back-translation, BPE dropout, etc.

Direction	**BLEU**
dsb → de	39.4
de → dsb	48.2
hsb → de	47.5
de → hsb	51
hsb → dsb	66.6
dsb → hsb	65.8

Table 5: Supervised results on Blind Test data of WMT22

6 Conclusion and Future Work

In this paper, we describe our submission to the WMT 2022 shared task of Unsupervised and Very Low Resource Supervised Machine Translation. We combined the advantages and mitigated the weaknesses of two novel unsupervised approaches along with pre-training a German-Slavic mBART model. Ablation studies for different components of our approach are left for future work.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. 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, and Eneko Agirre. 2019. An effective approach to unsupervised machine translation. *CoRR*, abs/1902.01313.

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. *CoRR*, abs/1901.07291.

Jindřich Libovický and Alexander Fraser. 2021. Findings of the wmt 2021 shared tasks in unsupervised mt and very low resource supervised mt. In *Proceedings of the Sixth Conference on Machine Translation*, pages 726–732, Online. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

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.

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, Portland, Oregon, USA. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218. Citeseer.