Study on Unsupervised Statistical Machine Translation for Backtranslation

Anush Kumar[1], Nihal V. Nayak[2], Aditya Chandra[1] and Mydhili K. Nair[1]
[1]Department of Information Science and Engineering, Ramaiah Institute of Technology
[2]Department of Computer Science, Brown University
anush070@gmail.com, nihalnayak@brown.edu,
adihyachndrt@gmail.com, mydhili.nair@msrit.edu

Abstract

Machine Translation systems have drastically improved over the years for several language pairs. Monolingual data is often used to generate synthetic sentences to augment the training data which has shown to improve the performance of machine translation models. In our paper, we make use of an Unsupervised Statistical Machine Translation (USMT) to generate synthetic sentences. Our study compares the performance improvements in Neural Machine Translation model when using synthetic sentences from supervised and unsupervised Machine Translation models. Our approach of using USMT for backtranslation shows promise in low resource conditions and achieves an improvement of 3.2 BLEU score over the Neural Machine Translation model.

1 Introduction

Neural Machine Translation systems with encoder-decoder architecture have significantly improved the state-of-the-art for several language pairs (Bahdanau et al., 2015; Vaswani et al., 2017). A majority of the systems in WMT18 used a Transformer approach with varying number of encoder-decoder layers (Bojar et al., 2018).

Supervised Neural Machine Translation systems are data-hungry as they require huge amounts of aligned parallel corpora (Koehn and Knowles, 2017). Additionally, obtaining parallel corpora is expensive and requires expert knowledge in both - source and target languages. To overcome this bottleneck, recent research has focused on unsupervised machine translation where only monolingual corpus is required, eliminating the need for a bilingual parallel corpora. The approaches in unsupervised machine translation have shown promise (Lample et al., 2018; Artetxe et al., 2018, 2019). Nonetheless, the performance of traditional Neural Machine Translation is still better than Unsupervised Machine Translation.

Often monolingual data is used to further improve the performance of a Neural Machine Translation model (Sennrich et al., 2016a; Currey et al., 2017). Backtranslation is one such popular method in which monolingual data is utilized to improve the model performance. In this technique, a model is initially trained in one of the directions (say target to source) and the trained model is used to translate a monolingual corpora to obtain synthetic sentences in the source language. These synthetic sentences are then included in the training set and a new model is trained from source to target. We take advantage of Unsupervised Machine Translation model for backtranslation (i.e. to generate synthetic sentences).

In our paper, we use Unsupervised Machine Translation Model (USMT) to generate the synthetic sentences for Russian-English language pair. Our aim is to improve the performance of the machine translation model using synthetic sentences from USMT model. Additionally, we look provide information on the settings and scenarios which benefit from USMT model and NMT model based backtranslation. To the best of our knowledge, there has been no study on using unsupervised machine translation models for backtranslation.

Our experiments indicated an improvement of 3.2 BLEU points for Russian to English language pair in a low resource setting while using synthetic sentences generated from an USMT model. However, we observed that NMT based backtranslation is superior when sufficient data is available and significantly improves the overall model performance.

Our paper is organized as follows - In the next section, we discuss related works in the field that have motivated us to carry out this study. Next, we

Proceedings of Recent Advances in Natural Language Processing, pages 578–582,
Varna, Bulgaria, Sep 2–4, 2019.
https://doi.org/10.26615/978-954-452-056-4_068

Dataset	Type	Domain	No. of Sentences
News-Commentary-v14 (Ru-En)	Parallel	News	290866
Newscrawl 2018 (Ru)	Monolingual	News	8669559
Newscrawl 2017 (Ru)	Monolingual	News	8233907
Newscrawl 2018 (En)	Monolingual	News	18113311

Table 1: Statistics about the data

describe the datasets that we used for our experiments. Next, we describe our experimental setup for carrying out the experiments. We then analyze the results for NMT, USMT and other models on Russian - English dataset. We conclude the paper with discussion on future work.

2 Related Works

In this section, we will describe the main ideas and experiments using backtranslation and monolingual data.

Backtranslation was popularized by Sennrich et al. (2016a) where they improved the state-of-the-art in several language pairs. They trained a target to source machine translation model on the aligned parallel corpus. The model was later used to translate target-side monolingual sentences to the source language. These new synthetic sentences were added to the training set and a new model is trained.

On similar lines, Zhang and Zong (2016) use the source-side monolingual data to train the NMT model. They build a baseline NMT model and then use that model to generate the synthetic parallel data. They experiment on self-training as well as multitask learning.

He et al. (2016) introduced a dual learning mechanism for neural machine translation. They train translation models in both directions and use monolingual data to provide feedback on the quality of the translation. Their main contribution was to treat machine translation as a reinforcement learning problem.

Hoang et al. (2018) proposed a simple technique. They do show that iterative backtranslation improves the performance of the system on large datasets. However, they observe that the improvements in low resource datasets were not significant.

As mentioned earlier, obtaining aligned parallel corpora is expensive and cumbersome. Recent efforts in Unsupervised Machine Translation indicate that it is possible to develop a competitive system using only monolingual corpus (Lample et al., 2018; Artetxe et al., 2018).

In our study, we use an unsupervised statistical machine translation system based on Artetxe et al. (2018) [1]. In their paper, they describe a novel method to build a statistical machine translation model using monolingual data without any parallel data. The main idea is to learn word embeddings for each language independently and use linear transformations to bring them to shared space. These embeddings are used to generate the phrase table and then, the SMT is fine-tuned on a synthetic training set. In our work, we make use of this model to generate more synthetic data to be added to the parallel corpora.

3 Data

We perform our experiments on the Russian - English language pair. For training the supervised model (both the back-translated model and re-trained model), we use the Russian - English parallel corpus from WMT 19 [2]. To train our unsupervised model, we used monolingual corpora for Russian [3] and English [4] from Newscrawl 2017 and 2018 datasets. For English, we consider only the Newscrawl 2018 as the monolingual data where as for Russian we combine both Newscrawl 2017 and 2018 for the monolingual data.

We provide more details regarding the data in Table 1.

3.1 Preprocessing

For normalizing punctuation in the parallel corpora, we use the default scripts provided by Moses [5]. We then perform true-casing followed byte-

[1] We use this method as it did not require huge amounts of resource to train compared to other methods such as Artetxe et al. (2019)

[2] http://www.statmt.org/wmt19/index.html

[3] http://data.statmt.org/news-crawl/ru/

[4] http://data.statmt.org/news-crawl/en/

[5] https://github.com/moses-smt/mosesdecoder/tree/master/scripts

pair encoding while training the NMT [6] (Sennrich et al., 2016b).

3.2 Postprocessing

We remove byte pair encodings, detrucase and detokenize all the translated sentences before validating the predictions against the test set.

4 Experimental Setup

We describe our experimental setup in this section.

For our unsupervised statistical machine translation model, we use Monoses (Artetxe et al., 2018). We use the defaults for training the unsupervised model from the Monoses code. We train the model on both Russian to English (ru-en) and English to Russian (en-ru) monolingual corpus.

Monoses frameworks trains the model in 8 steps. Steps 1-7 involves training the word embeddings and bringing them to a shared space to build an initial phrase table. In step 8, 2M sentences are generated through backtranslation in both the directions. The phrase table is fine-tuned for 3 iterations using the synthetic sentences to obtain the final model.

We use OpenNMT for training the supervised machine translation models (Klein et al., 2017). We use the default architecture in OpenNMT, which includes an LSTM layer for encoding and another LSTM layer for decoding. Furthermore, we do not use the pretrained word embeddings. For both NMT and retraining the NMT model, we use the same architecture with default values.

To train the USMT model and NMT models, we used a system with 4x vCPUS, NVIDIA Tesla K80 GPU and 61 GB of RAM. The USMT model took about 2 weeks to complete training where as the NMT model usually took about 4-5 hours.

5 Experiments

Our primary aim is to show that we can use an unsupervised machine translation model to improve the performance of NMT systems. At the same time, we want to have a fair validation and investigate numerous scenarios where our approach performs well.

Therefore, in this experiment, we use monolingual data to generate synthetic sentences using both NMT and USMT separately and then, each of them is used to augment the training data for the Neural Machine Translation model. For example, we use back-translated Russian monolingual corpus to generate synthetic English sentences. These English sentences are added to the training corpus (with varying training corpora sizes) with English as the source and Russian as the target. We then train an NMT model from scratch and report our performance on an unseen test set.

In the case USMT, we backtranslate the monolingual corpus using the USMT model and augment the sampled training data. The training data is used to build the NMT model.

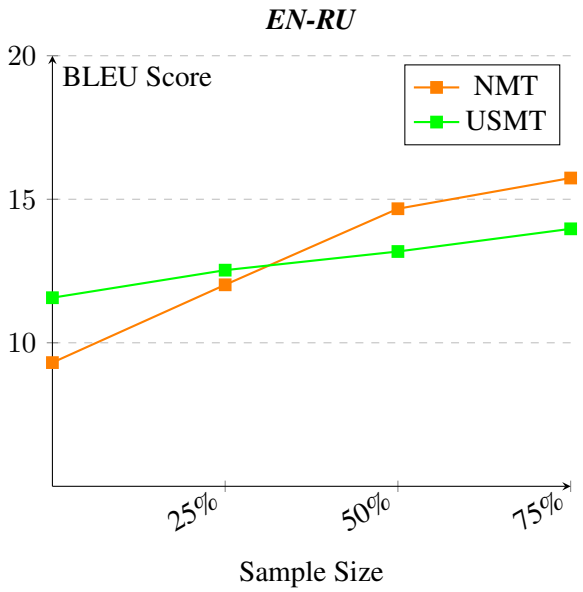

Figure 1: BLEU scores for NMT backtranslation and USMT backtranslation for EN-RU

6 Results

We perform all the experiments mentioned in the Table 2. Each row has a key which corresponds to the type of training corpora. The training corpora includes Only Parallel corpora, Only Monolingual Corpora and a sampled combination of the Parallel Corpora and the backtranslated Monolingual Corpora.

We obtain the baseline NMT model using the parallel corpus without any additional synthetic sentences. As mentioned earlier, we train the parallel corpora using the default encoder-decoder architecture from OpenNMT. Furthermore, we train a baseline USMT model using the monolingual corpora of English and Russian. We report the results in Table 2.

From the baselines of NMT and USMT, it is very clear that NMT outperforms USMT when

[6] https://github.com/rsennrich/subword-nmt/tree/master/subword_nmt

Training Corpus	ru-en		en-ru	
	NMT	USMT	NMT	USMT
Only Parallel Corpora	**17.65**	-	**15.54**	-
Only Monolingual Corpora	-	**15.58**	-	**8.07**
10% (∼29K) + Backtranslated Corpora	13.17	**16.34**	9.31	**11.57**
25% (∼72.5K) + Backtranslated Corpora	14.81	**17.17**	12.02	**12.63**
50% (∼145K) + Backtranslated Corpora	**19.63**	18.46	**14.67**	13.18
75% (∼217K)+ Backtranslated Corpora	**20.17**	19.34	**15.74**	13.97

Table 2: The BLEU scores for the different experiments. The training corpus with x% indicates the percentage of aligned training pairs randomly sampled from the parallel corpora.

there is sufficient data available.

To train the NMT backtranslation models, we sample the parallel corpora in batches of 10%, 25%, 50%, and 75% and train the NMT backtranslation model i.e. NMT model in the reverse direction (target to source). The synthetic data is generated by translating the monolingual corpora in English to Russian and Russian to English respectively. In Table 2, we refer to the synthetic sentences as Backtranslated Corpora. These sentences are combined with the sampled parallel corpora and retrained in the correct direction. In the case of USMT, we directly translate the monolingual corpora and include these synthetic sentences as a part of the training for NMT model in the correct direction.

Our results indicate the following - It can be seen that in critically low resource scenarios, the USMT backtranslated model performs better than the NMT backtranslated model (By 2.4 to 3.2 BLEU points for Russian to English and 0.61 to 2.26 BLEU points for English to Russian). However, the performance of the NMT system dramatically increases with the availability of parallel data. This shows that USMT as a backtranslation model works well mostly in low resource settings.

We can infer that the quality of the backtranslation model has a significant impact on the performance of the model. Additionally, we can see that the NMT model with small amount parallel data combined USMT model improves over the USMT baseline performance.

7 Discussion

In our future experiments, we would like to investigate the effect on lexical properties such as Named Entities and numbers in the predictions. We would also like to experiment our approach with newer techniques from Unsupervised Machine Translation (Artetxe et al., 2019). Additionally, we would like to extend our approach to other languages to comprehensively test our hypothesis.

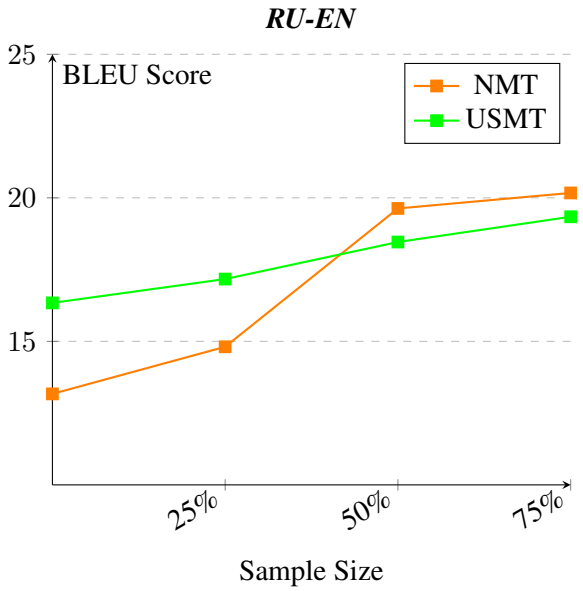

Figure 2: BLEU scores for NMT backtranslation and USMT backtranslation for RU-EN

Our experiments show that Unsupervised Statistical Machine Translation models can be used as a means of obtaining backtranslations to improve the performance of supervised machine translation models. We also note that the improved performances due unsupervised machine translation models are restricted to low resource scenarios. The performance of NMT model with NMT backtranslated sentences is superior when compared to the NMT model with USMT backtranslated sentences. In conclusion, our study helps in identifying the settings which benefit from USMT and NMT backtranslation models.

8 Code

To facilitate reconstruction of our paper, we are releasing the code - https://github.com/anush6/USMT_For_Backtranslation

9 Acknowledgement

We would like to thank our anonymous reviewers for their helpful feedback and comments.

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*, Brussels, Belgium. Association for Computational Linguistics.

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 (Volume 1: Long Papers)*, Florence, Italy. Association for Computational Linguistics.

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*.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Anna Currey, Antonio Valerio Miceli Barone, and Kenneth Heafield. 2017. Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the EMNLP 2017 Second Conference on Machine Translation (WMT17)*, Copenhagen, Denmark.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 820–828. Curran Associates, Inc.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24, Melbourne, Australia. Association for Computational Linguistics.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

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 (EMNLP)*.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545, Austin, Texas. Association for Computational Linguistics.