Utilizing Monolingual Data in NMT for Similar Languages: Submission to Similar Language Translation Task

Jyotsana Khatri, Pushpak Bhattacharyya
Department of Computer Science and Engineering,
Indian Institute of Technology Bombay
{jyotsanak,pb}@cse.iitb.ac.in

Abstract

This paper describes our submission to Shared Task on Similar Language Translation in Fourth Conference on Machine Translation (WMT 2019). We submitted three systems for Hindi → Nepali direction in which we have examined the performance of a Recursive Neural Network (RNN) based Neural Machine Translation (NMT) system, a semi-supervised NMT system where monolingual data of both languages is utilized using the architecture by (Artetxe et al., 2017) and a system trained with extra synthetic sentences generated using copy of source and target sentences without using any additional monolingual data.

1 Introduction

In this paper, we present the submission for Similar Language Translation Task in WMT 2019. The task focuses on improving machine translation results for three language pairs Czech-Polish (Slavic languages), Hindi-Nepali (Indo-Aryan languages) and Spanish-Portuguese (Romance languages). The main focus of the task is to utilize monolingual data in addition to parallel data because the provided parallel data is very small in amount. The detail of task is provided in (Barrault et al., 2019). We participated for Hindi-Nepali language pair and submitted three systems based on NMT for Hindi → Nepali direction. We have utilized monolingual data of both languages and also trained an NMT system with copy data from both sides with no additional monolingual data.

The rest of the paper is organized as follows: We start with introduction to NMT, followed by a list of some of the existing methods for how to utilize monolingual data in NMT. A brief introduction to unsupervised and semi-supervised NMT is provided. We also describe in brief about two existing popular methods of training cross-lingual word embeddings in an unsupervised way. In Section 4.3 we describe our three submitted systems for the task.

2 Neural Machine Translation

Many architectures have been proposed for neural machine translation. Most famous one is RNN based encoder-decoder proposed in (Cho et al.), where encoder and decoder are both recursive neural networks, encoder can be bi-directional. After this attention based sequence to sequence models where attention is utilized to improve performance in NMT are proposed in (Bahdanau et al., 2014), (Luong et al., 2015). Attention basically instructs the system about which words to focus more, while generating a particular target word. Transformer based encoder-decoder architecture for NMT is proposed in (Vaswani et al., 2017), which is completely based on self-attention and positional encoding. This does not follow recurrent architecture. Positional encoding provides the system with information of order of words.

NMT needs lots of parallel data to train a system. This task basically focuses on how to improve performance for languages which are similar but resource scarce. There are many language pairs for which parallel data does not exist or exist in a very small amount. In past, to improve the performance of NMT systems various techniques like Back-Translation (Sennrich et al., 2016a), utilizing other similar language pairs through pivoting (Cheng et al., 2017) or transfer learning (Zoph et al., 2016), complete unsupervised architectures (Artetxe et al., 2017) (Lample et al., 2018) and many others have been proposed.

2.1 Utilizing monolingual data in NMT

There has been good amount of work done on how we can utilize monolingual data to improve performance of an NMT system. Back-Translation

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 3: Shared Task Papers (Day 2) pages 197–201
Florence, Italy, August 1-2, 2019.

was introduced by (Sennrich et al., 2016b), to utilize monolingual data of target language. This requires a translation system in opposite direction. In (Sennrich et al., 2016b), a method where empty sentences are provided in the input for target side monolingual data is also evaluated, back-translation performs better than this. In iterative Back-Translation, systems in both directions improve each other (Hoang et al., 2018), it is done in an incremental fashion. To generate back-translated data, current system in opposite direction is utilized. In (Currey et al., 2017) , target side monolingual data is copied to generate source synthetic translations and the system is trained by combining this synthetic data with parallel data. In (Zhang and Zong, 2016), source side monolingual data is utilized to iteratively generate synthetic sentences from the same model. In (Domhan and Hieber, 2017), there is a separate layer for target side language model in training, decoder utilize both source dependent and source independent representations to generate a particular target word. In (Burlot and Yvon, 2018), it is claimed that quality of back-translated sentences is important.

Recently many systems have been proposed for Unsupervised NMT, where only monolingual data is utilized. The Unsupervised NMT approach proposed in (Artetxe et al., 2017) follows an architecture where encoder is shared and decoder is separate for each language. Encoder tries to map sentences from both languages in the same space, which is supported by cross-lingual word embeddings. They fix cross-lingual word embeddings in the encoder while training, which helps in generating cross-lingual sentence representations in the same space.

The system with one shared encoder and two separate decoders with no parallel data is trained by iterating between Denoising and Back-Translation. Denoising tries to generate the correct sentence from noisy sentences, in that way the decoder is learning how to generate sentences in that particular language. These noisy sentences are created by shuffling words within a window. If the system is only trained with denoising then it may turn out to be a denoising auto-encoder. So they have also introduced back-translation in the training process to introduce translation task. Training is done by alternating between denoising and back-translation for mini-batches if parallel data is not available. In a semi-supervised setting if some amount of parallel data is available, training alternates between denoising, back-translation and parallel sentences. In (Lample et al., 2018), encoder and decoder both are shared between the languages. Training is done by alternating between denoising and back-translation. Initialization is performed using a system trained on word-word translated sentences which is performed using cross-lingual word embeddings trained using MUSE(Conneau et al., 2017). They also utilize a discriminator which tries to identify the language from the encoder representations, this leads to adversarial training.

2.2 Cross-lingual word embeddings

Cross-lingual word embeddings tries to map two word embedding spaces of different languages in the same space. The basic assumption for generating the cross-lingual word embeddings in most papers is that both the embedding spaces must be isometric. Cross-lingual word embeddings is generated by learning a linear transformation which minimizes the distances between words given in a dictionary. There are many methods proposed for training cross-lingual word embeddings in an unsupervised way. While training cross-lingual word embeddings in an unsupervised manner there is no dictionary available, only the monolingual embeddings are available. In (Artetxe et al., 2018), cross lingual word embeddings are generated following a series of steps which involves: normalization of the embeddings so they can be used together to utilize for a distance metric, unsupervised initialization using normalized embeddings, self-learning framework using adversarial training where it iterates between creating the dictionary and finding the optimal mapping, and some weighting refinement over this. Through these steps a transformation of these spaces to a common space is learnt. In (Conneau et al., 2017) an adversarial training process is followed where discriminator tries to correctly identify the language using its representation and the mapping matrix W tries to confuse the discriminator.

3 System Overview

This section describes the specification of the systems submitted in detail. We have submitted sys-

tems for Hindi-Nepali language pair in Hindi → Nepali direction. Hindi and Nepali both are Indo-Aryan languages and are very similar to each other. They share a significant portion of the vocabulary and similar word orders. The three submitted systems are:

- A pure RNN based NMT system
- Semi-supervised RNN based NMT system
- Utilization of copied data in RNN based NMT

First system is pure RNN based NMT system. To train this we have utilized only parallel corpora.
Second system is trained using a semi-supervised NMT system where monolingual data from both languages is utilized. We have utilized architecture proposed in (Artetxe et al., 2017) where encoder is shared and decoders are separate for each language and model is trained by alternating between denoising and back-translation. This architecture can also be utilized for completely unsupervised setting.
Third system is also a pure RNN based NMT system where additional parallel data (synthetic data) is created by copying source side sentences to target side and target side sentences to source side, but we do this only for the available parallel sentences, no additional monolingual data is utilized. In this way the amount of available data becomes three times of the original data. All the data is combined together, shuffled and then provided to the NMT system, there is no identification provided to distinguish between parallel data and copy data.
To train all three systems we have utilized the implementation of (Artetxe et al., 2017).

4 Experimental Details

4.1 Dataset

We have utilized monolingual corpora of both languages in our primary system. The dataset details are given in Table 1. Hindi-Nepali parallel data is provided in the task, which contains 65505 sentences. Hindi monolingual corpora is IITB Hindi monolingual corpora (Kunchukuttan et al., 2018). Nepali monolingual sentences are created using the monolingual data of Wikipedia and CommonCrawl provided for Parallel corpus filtering task [1] by separating each sentence using | and keeping sentences of length 500 and less.

Dataset	Number of sentences
Hindi-Nepali Parallel Data	65,505
IITB Hindi Monolingual Corpora	45,075,242
Nepali Monolingual corpora	6,688,559

Table 1: Dataset details

4.2 Preprocessing

Sentences are preprocessed using tokenization and Byte Pair Encoding (BPE). Sentences are tokenized for both hindi and nepali using IndicNLP[2] library. This tokenized data is preprocessed using BPE. Number of merge operations for BPE is set to 20000 for both languages and learnt separately for each language. The results may improve if we learn BPE jointly because both languages are similar. Byte pair Encoding is learnt using the implementation by (Sennrich et al., 2016b).
Monolingual embeddings are trained using FastText[3] (Bojanowski et al., 2017) using bpe applied monolingual data for both languages. The dimension of embeddings is set to 100. Cross-lingual embeddings are created using VecMap (Artetxe et al., 2018).

4.3 System detail

Table 2 reports BLEU score for the test and dev data for all three systems. We have not utilized dev data while training. We have used encoder and decoder with 2 layers, 600 hidden units each, GRU cells, batch size of 50 and maximum sentence length of 50. Adam optimizer is used with learning rate 0.0002. We have trained all three systems with fixed 300000 iterations. The number of sentences in test and dev data is 1567 and 3000 respectively. The BLEU score for test data is provided by task organizers and for dev data BLEU score is calculated using multi-bleu.pl from moses toolkit (Koehn et al., 2007).

[1] http://www.statmt.org/wmt19/parallel-corpus-filtering.html

System	Test	Dev
Basic	3.5	4.6
With Monolingual Data	2.8	3.27
With copy data	2.7	4.38

Table 2: Experimental results (BLEU scores)

4.4 Results

As it is clear from the results in Table 2 that the system with only parallel data is performing better than when we are utilizing monolingual data. To answer why this is happening, a study of size and quality of monolingual data, the study of ratio of monolingual and parallel data provided to the system is required. The intuition behind using copied data with parallel data is, that both the languages are similar and this may provide more data to the system. But the results show the system is getting confused as we are providing all the data together without any distinguishing mark between parallel and copied sentences. For the same sentence both original translation and its copy is given in the output which may be causing confusion.

5 Summary

In this paper we have explained about systems submitted for Similar Language Translation task in WMT 2019. We have reported results for a semi-supervised technique which utilizes denoising and back-translation. We have utilized lots of monolingual data together with available parallel data for training a neural machine translation system which share encoder and have separate decoders for each language, in a semi-supervised setting. A study of size and quality of monolingual data is required to analyze the performance which is left as future work. We have also explained results for utilizing copied data with parallel data and compared both the above mentioned techniques with a pure RNN based NMT system.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798.

[2] https://github.com/anoopkunchukuttan/indic_nlp_library

[3] https://fasttext.cc/

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2017. Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041*.

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

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (wmt19). In *Proceedings of the Fourth Conference on Machine Translation, Volume 2: Shared Task Papers*, Florence, Italy. Association for Computational Linguistics.

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.

Franck Burlot and François Yvon. 2018. Using monolingual data in neural machine translation: a systematic study. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 144–155.

Yong Cheng, Qian Yang, Yang Liu, Maosong Sun, and Wei Xu. 2017. Joint training for pivot-based neural machine translation. In *IJCAI*, pages 3974–3980.

Kyunghyun Cho, Bart van Merriënboer Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation.

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

Anna Currey, Antonio Valerio Miceli Barone, and Kenneth Heafield. 2017. Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 148–156.

Tobias Domhan and Felix Hieber. 2017. Using target-side monolingual data for neural machine translation through multi-task learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1500–1505.

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.

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.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2018. The iit bombay english-hindi parallel corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, et al. 2018. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049.

Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.

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)*, volume 1, pages 86–96.

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)*, volume 1, pages 1715–1725.

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.

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.

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