JU-Saarland Submission in the WMT2019 English–Gujarati Translation Shared Task

Riktim Mondal[1,*], Shankha Raj Nayek[1,*], Aditya Chowdhury[1,*], Santanu Pal[2], Sudip Kumar Naskar[1], Josef van Genabith[2]
[1]Jadavpur University, Kolkata, India
[2]Saarland University, Germany
{riktimrules,shankharaj29,adityachowdhury21}@gmail.com
{santanu.pal,josef.vangenabith}@uni-saarland.de
sudip.naskar@cse.jdvu.ac.in

Abstract

In this paper we describe our joint submission (JU-Saarland) from Jadavpur University and Saarland University in the WMT 2019 news translation shared task for English–Gujarati language pair within the translation task subtrack. Our baseline and primary submissions are built using a Recurrent neural network (RNN) based neural machine translation (NMT) system which follows attention mechanism followed by fine-tuning using in-domain data. Given the fact that the two languages belong to different language families and there is not enough parallel data for this language pair, building a high quality NMT system for this language pair is a difficult task. We produced synthetic data through back-translation from available monolingual data. We report the automatic evaluation scores of our English–Gujarati and Gujarati–English NMT systems trained at word, byte-pair and character encoding levels where RNN at word level is considered as the baseline and used for comparison purpose. Our English–Gujarati system ranked in the second position in the shared task.

1 Introduction

Neural Machine translation (NMT) is an approach to machine translation (MT) that uses artificial neural network to directly model the conditional probability $p(y|x)$ of translating a source sentence (x_1,x_2,...,x_n) into a target sentence (y_1,y_2,...,y_m). NMT has consistently performed better than the phrase-based statistical MT (PB-SMT) approaches and has provided state-of-the-art results in the last few years. However, one of the major constraints of using supervised NMT is that it is not suitable for low resource language pairs. Thus, to use supervised NMT, low resource pairs need to resort to other techniques to increase the size of the parallel training dataset. In the WMT 2019 news translation shared task, one such resource scarce language pair is English-Gujarati. Due to insufficient volume of parallel corpora available to train an NMT system for these language pairs, creation of more actual/synthetic parallel data for low resources languages such as Gujarati, is an important issue.

In this paper, we described our joint participation of Jadavpur University and Saarland University in the WMT 2019 news translation task for English–Gujarati and Gujarati–English. The released training data set is completely different in-domain compared to the development set and the size is not anywhere close to the sizable amount of training data which is typically required for the success of NMT systems. We use additional synthetic data produced through back-translation from the monolingual corpus. This provides significant improvements in translation performance for both our English–Gujarati and Gujarati–English NMT systems. Our English–Gujarati system was ranked second in terms of BLEU (Papineni et al., 2002) and TER (Snover et al., 2006) in the shared task.

2 Related Works

Dungarwal et al. (Dungarwal et al., 2014) developed a statistical method for machine translation, where phrase based method for Hindi-English and factored based method for English-Hindi SMT system was used. They had shown improvements to the existing SMT systems using pre-procesing and post-processing components that generated morphological inflections correctly. Imankulova et al. (Imankulova et al., 2017) showed how back-translation and filtering from monolingual data can be used to build an effective translation system for a low-resourse language pair like Japanese-

*These three authors have contributed equally.

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers (Day 1) pages 308–313
Florence, Italy, August 1-2, 2019.

Dataset	Pairs
Parallel Corpora	192,367
Cleaned Parallel Corpora	64,346
Back-translated Data	219,654
Development Data	1,998
Gujarati Test Data	1,016
English Test Data	998

Table 1: Data Statistics of WMT 2019 English–Gujarati translation shared task.

Russian. Sennrich et al. (Sennrich et al., 2016a) shown how back-translation of monolingual data can improve the NMT system. Ramesh et al. (Ramesh and Sankaranarayanan, 2018) demonstrated how an existing model like bidirectional recurrent neural network can be used to generate parallel sentences for non-English languages like English-Tamil and English-Hindi, which belong to low-resource language pair, to improve the SMT and the NMT systems. Choudhary et al. (Choudhary et al., 2018) has shown how to build NMT system for low resource parallel corpus language pair like English-Tamil using techniques like word embeddings and Byte-Pair-Encoding (Sennrich et al., 2016b) to handle Out-Of-Vocabulary Words.

3 Data Preparation

For our experiments we used both parallel and monolingual corpus released by the WMT 2019 Organizers. We back-translate the monolingual corpus and use it as additional synthetic parallel corpus to train our NMT system. The detailed statistics of the corpus is given in Table 1.

We performed our experiments on two datasets, one using the parallel corpus provided by WMT 2019 for the Gujarati–English news translation shared task, and the other using the parallel corpus combined with back translated sentences from provided monolingual corpus (only News crawl corpus was used for back translation) for the same language pair.

Since the released parallel corpus was very noisy, containing redundant sentences, we cleaned the parallel corpus, the procedure of which is described in section 3.1.

In the next step we shuffle the whole corpus as it reduces variance and makes sure that our model overfits less. We then split the dataset into three parts: training, validation and test set. Shuffling is important in the splitting part too as it is important to choose the test and validation set from the same distribution and must be chosen randomly from the available data. Here, test set was also shuffled as this dataset was used for our internal assessment. After cleaning, we randomly selected 64,346 sentence pairs for training, 1,500 sentence pairs for validation and 1,500 sentences as test data. It is to be noted that our validation and test corpus is taken from the released parallel data to setup a baseline model. Later when WMT19 Organizers released the development set, we continued training our models by considering WMT19 development set as our test set and the new development set consisting of 3,000 sentences which were obtained after combining 1,500 sentences from the validation and the testing set (both were from the parallel corpus as stated above). While training our final model, the released development set was used. After cleaning it was obvious that the amount of training data is not enough to train a neural system for such a low resource language pair. Therefore, preparation for large volume of parallel corpus is required which can be produced either by manual translation by professional translators or scraping parallel data from the internet. However, these processes are costly, tedious and sometimes inefficient (in case of scraping from internet).

As the released data was insufficient, to generate more training data, we use back-translation. For back-translation we applied two methods, first, using unsupervised statistical machine translation as described in (Artetxe et al., 2018) and second, using Doc translation API[1] (The API uses Google translator as of April 2019). We have explained the extraction of sentences and the corresponding results using the above methods in section 4.2. The synthetic dataset which we have generated can be found here.[2]

3.1 Data Preprocessing

To train an efficient machine translation system, it is required to clean the available raw parallel corpus for the system to produce consistent and reliable translations. The released version of the raw parallel corpus consisted of redundant pairs which needs to be removed to obtain better results

[1] https://www.onlinedoctranslator.com/en/

[2] https://github.com/riktimmondal/Synthetic-Data-WMT19-for-En-Gu-Language-pair

as demonstrated in previous works (Johnson et al., 2017) which are of types as given below:

- The source is same for different targets.
- The source is different for the same target.
- Repeated identical sentence pair

The redundancy in the translation pairs makes the model prone to overfitting and hence prevents it from recognizing new features. Thus, one of the sentence pair is kept while the other redundant pairs are removed. Some sentence pairs had combinations of both language pairs which were also identified as redundant. These pairs strictly need elimination as the vocabularies of the individual languages consist of alphanumeric characters of the other language which results in inconsistent encoding and decoding during encoder-decoder application steps on the considered language pair. We tokenize the English side using Moses (Koehn et al., 2007) tokenizer and for Gujarati, we use the Indic NLP library tokenization tool[3]. Punctuation normalization was also done.

3.2 Data Postprocessing

Postprocessing, such as detokenization (Klein et al., 2017), punctuation normalization[4] (Koehn et al., 2007), was performed on our translated data (on the test set) to produce the final translated data.

4 Experiment Setup

We have explained our experimental setups in the next two sections. The first section contains the setup used for our final submission and the next section describes all the other supporting experimental setups. We use the OpenNMT toolkit (Klein et al., 2017) for our experiments.
We performed several experiments where the parallel corpus is sent to the model as space separated character format, space separated word format, and space separated Byte Pair Encoding (BPE) format (Sennrich et al., 2016b). For our final (i.e., primary) submission for the English–Gujarati task, the source input words were converted to BPE whereas the Gujarati words were kept as it is. For our Gujarati–English submission, both the source and the target were in simple word level format.

[3] http://anoopkunchukuttan.github.io/indic_nlp_library/

[4] punctuation_normalization.perl

4.1 Primary System description

Our primary NMT systems are based on attention-based uni-directional RNN (Cho et al., 2014) for Gujarati–English and bi-directional RNN (Cheng et al., 2016) for English–Gujarati.

hyper-parameter	Value
Model-type	text
Model-dtype	fp32
Attention-layer	2
Attention-Head/layer	8
Hidden-layers	500
Batch-Size	256
Training-steps	160,000
Source vocab-size	50,000
Target vocab-size	50,000
learning-rate	warm-up+decay*
global-attention function	softmax
tokenization-strategy	wordpiece
RNN-type	LSTM

Table 2: Hyper-parameter configurations for Gujarati–English translation using unidirectional RNN (Cho et al., 2014)), *learning-rate was initially set to 1.0.

Table 2 shows the hyper-parameter configurations for our Gujarati–English translation system. We initially trained our model with the cleaned parallel corpus provided by WMT 2019 up to 100K training steps. Thereafter, we fine-tune our generic model on domain specific corpus (containing 219K sentences back-translated using Doc Translator API) changing the learning rate to 0.5 and decay started from 130K training steps with a decay factor of 0.5 and keeping the other hyper-parameters same as mentioned in Table 2.

hyper-parameter	**Value**
Model-type	text
Model-dtype	fp32
Encoder-type	BRNN
Attention-layer	2
Attention-Head/layer	8
Hidden-layers	512
Batch-Size	256
Training-steps	135,000
Source vocab-size	26,859
Target vocab-size	50,000
learning-rate	warm-up+decay
global-attention function	softmax
tokenization-strategy	Byte-pair Encoding
RNN-type	LSTM

Table 3: Hyper-parameter configurations for English–Gujarati translation using bi-directional RNN (Cheng et al., 2016).

To build our English–Gujarati translation system, we initially trained a generic model like our

Gujarati–English translation system. However, in this case we use different hyper-parameter configurations as mentioned in Table 3. Additionally, here, we use byte-pair encoding on the English side with 32K merge operations. We do not perform BPE operation on the Gujarati corpus; we keep the original word format for Gujrati. Our generic model was trained with up to 100K training steps and then fine-tuned our model on domain specific parallel corpus having English side as BPE and Gujarati side as word level format. During fine-tuning, we reduce the learning rate from 1.0 to 0.25 and started decaying from 120K training steps with a decay factor of 0.5. The other hyper-parameter configurations remain unchanged. The respective hyperparameters used for the English–Gujarati task in our primary system submission were also tested for the reverse direction; however, it did not perform as good as the primary system and hence the final system is modified accordingly.

4.2 Other Supporting Experiments

In this section we describe all the supporting experiments that we performed for this shared task starting from Statistical MT to NMT with both supervised and unsupervised settings.

All the results and experiments discussed below are tested on the released development set (considering this as the test set). These models were not tested with the released test set as they provided poor BLEU scores on the development set.

We used uni-directional RNN having LSTM units trained on 64,346 pre-processed sentences (cf. Section 3) with 120K training steps and learning rate of 1.0. For English–Gujarati where input was space separated words for both sides, we achieved highest BLEU score of **4.15** after fine-tuning with 10K sentences selected from the cleaned parallel corpus whose total number of tokens(words) was exceeding 8.The BLEU score dropped to **3.56** while applying BPE on the both sides. For the other direction (Gujarati–English) of the language pair, we got highest BLEU scores of **5.13** and **5.09** at word level and BPE level respectively.

We also tried transformer-based NMT model (Vaswani et al., 2017) which however gave extremely poor results on similar experimental settings. The highest BLEU we achieved was **0.74** for Gujarati–English and **0.96** for English–Gujarati. The transformer model was trained until 100K training steps, with 64 batch size in a single GPU and positional encoding layers size was set to 2.

Since the the training data size was not enough, we used backtranslation to generate additional synthetic sentence pairs from the monolingual corpus released in WMT 2019. We initially used *monoses* (Artetxe et al., 2018), which is based on unsupervised statistical phrase based machine translation, to translate the monolingual sentences from English to Gujarati. We used 2M English sentences to train the monoses system. The training process took around 6 days in our modest 64 GB server. However, the results were extremely poor with a BLEU score of **0.24** for English–Gujarati and **0.01** for the opposite direction, without using preprocessed parallel corpus. Moreover, after adding preprocessed parallel corpus, the BLEU score dropped significantly. This motivated us to use online document translator, in our case Google translation API, for back-translating sentence pairs from the released monolingual dataset. The back-translated data was later combined with our preprocessed parallel corpus for our final model.

Additionally, we also tried a simple unidirectional RNN model on character level, however, this also fails to contribute in terms of improving performance. We have compiled all the results in table 4.

5 Primary System Results

Our primary submission for English–Gujarati using bidirectional RNN model with BPE at English side (see Section 4.1) and word format at Gujarati side gave the best result. On the other hand, the Gujarati-English primary submission, based on an uni-directional RNN model with both English and Gujarati in word format, gave the best result. Before submission, we performed punctuation normalization, unicode normalization, and detokenization for each runs. Table 5 shows the published results of our primary submissions on WMT 2019 Test set. Table 6 shows our hands on experimental results on the development set.

6 Conclusion and Future Work

In this paper, we applied NMT to one of the most challenging language pair, English–Gujarati, as the availability of parallel corpus is really scarce

Language pair	Model used	Tokenization Strategy	BLEU
EN-GU	RNN	Word	4.15
EN-GU	RNN	BPE	3.56
GU-EN	RNN	Word	5.13
GU-EN	RNN	BPE	5.09
EN-GU	Transformer	Word	0.96
GU-EN	Transformer	Word	0.74
EN-GU	Monoses	Word	0.24
GU-EN	Monoses	Word	0.01

Table 4: Results of supporting experiments.

Language pair	BLEU	BLEU-cased	TER	BEER2.0	characTER
EN-GU	21.9	21.9	0.688	0.529	0.647
GU-EN	12.8	11.8	0.796	0.422	0.891

Table 5: WMT 2019 evaluation for EN-GU and GU-EN on test set.

Language pair	BLEU	BLEU-cased
EN-GU	22.3	22.3
GU-EN	17.6	16.8

Table 6: WMT 2019 evaluation for EN-GU and GU-EN on development set released by WMT 2019.

for this language pair. In this scenario, collecting and preprocessing of data play very crucial role to increase the dataset as well as to obtain quality result using NMT. In this paper we show how increasing the parallel data through back-translation via Google translation API can increase the overall performance. Our primary result also exceeded Google translate (which gave a BLEU of 13.7) by a margin of around 8.0 absolute BLEU points. Our method is not just limited to English–Gujarati translation task; it can also be useful in various scarce-resource language pairs and domains.

We did not make use of any ensemble mechanism in this task, otherwise we could have achieved higher BLEU scores. Therefore, in future we will try to ensemble several models, increasing more useful back-translated data using existing state-of-the-art model. In future, we would also like to explore cross-lingual BERT (Devlin et al., 2018) to enhance the performance.

Acknowledgments

We want to thank Google Colab where we trained all our models. Santanu Pal is supported in part by the German research foundation (DFG) under grant number GE 2819/2-1 (project MMPE) and the German Federal Ministry of Education and Research (BMBF) under funding code 01IW17001 (project Deeplee). The responsibility for this publication lies with the authors. Sudip Kumar Naskar is partially supported by Digital India Corporation (formerly Media Lab Asia), MeitY, Government of India, under the Young Faculty Research Fellowship of the Visvesvaraya PhD Scheme for Electronics IT. We also want to thank the reviewers for their valuable input, and the organizers of the shared task.

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.

Yong Cheng, Shiqi Shen, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Agreement-based joint training for bidirectional attention-based neural machine translation. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, pages 2761–2767. AAAI Press.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder

for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Himanshu Choudhary, Aditya Kumar Pathak, Rajiv Ratan Saha, and Ponnurangam Kumaraguru. 2018. Neural machine translation for English-Tamil. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 770–775, Belgium, Brussels. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Piyush Dungarwal, Rajen Chatterjee, Abhijit Mishra, Anoop Kunchukuttan, Ritesh Shah, and Pushpak Bhattacharyya. 2014. The IIT Bombay Hindi-English translation system at WMT 2014. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 90–96, Baltimore, Maryland, USA. Association for Computational Linguistics.

Aizhan Imankulova, Takayuki Sato, and Mamoru Komachi. 2017. Improving low-resource neural machine translation with filtered pseudo-parallel corpus. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 70–78, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proc. ACL*.

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 ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. 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 on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sree Harsha Ramesh and Krishna Prasad Sankaranarayanan. 2018. Neural machine translation for low resource languages using bilingual lexicon induced from comparable corpora. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 112–119, New Orleans, Louisiana, USA. Association for Computational Linguistics.

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.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.