The IIIT-H Gujarati-English Machine Translation system for WMT19

Vikrant Goyal
IIIT Hyderabad
vikrant.goyal@research.iiit.ac.in

Dipti Misra Sharma
IIIT Hyderabad
dipti@iiit.ac.in

Abstract

This paper describes the Neural Machine Translation system of IIIT-Hyderabad for the Gujarati→English news translation shared task of WMT19. Our system is based on encoder-decoder framework with attention mechanism. We experimented with Multilingual Neural MT models. Our experiments show that Multilingual Neural Machine Translation leveraging parallel data from related language pairs helps in significant BLEU improvements upto 11.5, for low resource language pairs like Gujarati-English.

1 Introduction

Neural Machine Translation (Luong et al., 2015; Bahdanau et al., 2014; Johnson et al., 2017; Wu et al., 2017; Vaswani et al., 2017) has been receiving considerable attention in the recent years, given its superior performance without the demand of heavily hand crafted engineering efforts. NMT often outperforms Statistical Machine Translation (SMT) techniques but it still struggles if the parallel data is insufficient like in the case of Indian languages.

The bulk of research on low resource NMT has focused on exploiting monolingual data or parallel data from other language pairs. Some recent methods to improve NMT models that exploit monolingual data ranges from back-translation (Sennrich et al., 2015a), dual NMT (He et al., 2016) to Unsupervised MT models (Lample et al., 2017; Artetxe et al., 2017; Lample et al., 2018). Transfer Learning is also a promising approach for low resource NMT which exploits parallel data from other language pairs (Zoph et al., 2016; Nguyen and Chiang, 2017; Kocmi and Bojar, 2018). Typically it is achieved by training a parent model in a high resource language pair, then using some of the trained weights as the initialization for a child model and further train it on the low-resource language pair. Other promising approach for improving translation performance for low resource languages is Multilingual Neural Machine Translation. It has been shown that exploiting data from other language pairs & joint training helps in improving the translation performance of NMT models. (Ha et al., 2016; Firat et al., 2016; Johnson et al., 2017).

This paper describes the NMT system of IIIT-H for WMT19 evaluation. We participated in the Gujarati→English news translation task. We used an attention-based encoder-decoder model as our baseline system and used Byte Pair Encoding (BPE) to enable open vocabulary translation. We then leverage Hindi-English parallel corpus in a multilingual setting so as to improve our baseline system. We basically combined Hindi-English and Gujarati-English parallel corpus and use it as our training corpus. Our multilingual system is similiar to Johnson et al. (2017) but we don't use any artificial token at the start of source sentences to indicate the target language. The reason is trivial, that is we have only English as our target language. We also provide results of our experiments conducted post WMT19 shared task involving Transformer models.

2 Neural MT Architecture

Our NMT model consists of an encoder and a decoder, each of which is a Recurrent Neural Network (RNN) as described in (Luong et al., 2015). The model directly estimates the posterior distribution $P_\theta(y|x)$ of translating a source sentence $x = (x_1, .., x_n)$ to a target sentence $y = (y_1, .., y_m)$ as:

$$P_\theta(y|x) = \prod_{t=1}^{m} P_\theta(y_t|y_1, y_2, .., y_{t-1}, x) \qquad (1)$$

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers (Day 1) pages 191–195
Florence, Italy, August 1-2, 2019. ©2019 Association for Computational Linguistics

Each of the local posterior distribution $P(y_t|y_1,_2,..,y_{t-1},x)$ is modeled as a multinomial distribution over the target language vocabulary which is represented as a linear transformation followed by a softmax function on the decoder's output vector $\tilde{h}_t^{dec}$:

$$c_t = AttentionFunction(h_{1:n}^{enc}, h_t^{dec}) \quad (2)$$

$$\tilde{h}_t^{dec} = tanh(W_o[h_t^{dec}; c_t]) \quad (3)$$

$$P(y|y_1, y_2, .., y_{t-1}, x) = softmax(W_s\tilde{h}_t^{dec}; \tau) \quad (4)$$

where c_t is the context vector, h^{enc} and h^{dec} are the hidden vectors generated by the encoder and decoder respectively, AttentionFunction(. , .) is the attention mechanism as shown in (Luong et al., 2015) and [. ; .] is the concatenation of two vectors.

An RNN encoder first encodes x to a continuous vector, which serves as the initial hidden vector for the decoder and then the decoder performs recursive updates to produce a sequence of hidden vectors by applying the transition function f as:

$$h_t^{dec} = f(h_{t-1}^{dec}, [\tilde{h}_{t-1}^{dec}; e(y_t)]) \quad (5)$$

where e(.) is the word embedding operation. Popular choices for mapping f are Long-Short-Term Memory (LSTM) units and Gated Recurrent Units (GRU), the former of which we use in our models.

An NMT model is typically trained under the maximum log-likelihood objective:

$$\max_\theta J(\theta) = \max_\theta \mathrm{E}_{(x,y)\sim D}[\log P_\theta(y|x)] \quad (6)$$

where D is the training set. Our NMT model uses a bi-directional RNN as an encoder and a uni-directional RNN as a decoder with global attention (Luong et al., 2015) .

3 Multilingual Neural Machine Translation

Most of the practical applications in Machine Translation have focused on individual language pairs because it was simply too difficult to build a single system that translates to and from many language pairs. But Neural Machine Translation was shown to be an end-to-end learning approach and was quickly extended to multilingual machine translation in several ways. In Dong et al. (2015), the authors modify the attention-based encoder-decoder approach by introducing separate decoder and attention mechanism for each target language. In Zoph and Knight (2016), multi-source translation was proposed where the model has different encoders and different attention mechanisms for different source languages. In Firat et al. (2016), the authors proposed a multi-way multilingual NMT model using a single shared attention mechanism but with multiple encoders/decoders for each source/target language. In this paper, we adopted the approach proposed in Johnson et al. (2017), where a single NMT model is used for multilingual machine translation. We used Hindi-English as our assisting language pair and combined it with Gujarati-English parallel data to form a multi source translation system.

4 Experimental setup

4.1 Dataset

In our experiments, we use the Gujarati-English training data provided by the organisers namely Wiki Titles, Bible corpus, Localisation Opus, Wikipedia corpus & crawled corpus. It consists of around 155K parallel sentences. We used newsdev2019 as our development corpus. For building our multilingual model, we used IIT-Bombay parallel data (Kunchukuttan et al., 2017) as our Hindi-English parallel corpus. The top level statistics of the data used is provided in Table 1.

Table 1: Statistics of our processed parallel data.

Dataset	Sentences	Tokens
IITB Hi-En Train	15,28,631	21.5M / 20.3M
Gu-En Train	1,55,767	1.68M / 1.58M
Gu-En Dev	1,997	51.3K / 47.4K
Gu-En Test	1,998	51.5K / 47.5K

4.2 Data Processing

We used Moses (Koehn et al., 2007) toolkit for tokenization and cleaning the English side of the data. Gujarati and Hindi sides of the data is first normalized with Indic NLP library[1] followed by tokenization with the same library. As our preprocessing step, we removed all the sentences of length greater than 80 from our training corpus.

4.3 Subword Segmentation for NMT

Neural Machine Translation relies on first mapping each word into the vector space, and tradi-

[1] https://anoopkunchukuttan.github.io/indic_nlp_library/

tionally we have a word vector corresponding to each word in a fixed vocabulary. Addressing the problem of data scarcity and the hardness of the system to learn high quality representations for rare words, (Sennrich et al., 2015b) proposed to learn subword units and perform translation at a subword level. With the goal of open vocabulary NMT, we incorporate this approach in our system as a preprocessing step. In our early experiments, we note that Byte Pair Encoding (BPE) works better than UNK replacement techniques. For our baseline system, we learn separate vocabularies for Hindi and English each with 32k merge operations. For our multilingual model, we learn a joint vocabulary for Hindi and Gujarati & a separate vocabulary for English. With the help of BPE, the vocabulary size is reduced drastically and we no longer need to prune the vocabularies. After the translation, we do an extra post processing step to convert the target language subword units back to normal words. We found this approach to be very helpful in handling rare word representations.

4.4 Script Conversion

India is a linguistically rich country having 22 constitutional languages, written in different scripts. Indian languages are highly inflectional with a rich morphology, default sentence structure as subject object verb (SOV) and relatively free word order. Many of them are structurally similar, also called as sibling languages. Hindi & Gujarati languages are such siblings. That is why, we have chosen Hindi as an assisting language for our multilingual model.

Although, there are many linguistic similarities between Gujarati & Hindi, both of these languages are written in different scripts. So, to make a strong multilingual NMT model, we converted the script of the Gujarati side of the parallel corpus to Hindi (Devanagari script). We used Indic NLP Library's transliteration script for this purpose. We found this approach to be very helpful in enabling better sharing between languages on the encoder side. BPE also enhances the usage of script conversion technique. We used script conversion only with our additional Multilingual NMT experiments based on Transformer architecture.

4.5 Training Details

The structure of our NMT model is same as in Luong et al. (2015), an RNN based encoder-decoder model with Global Attention mechanism. We used an LSTM based Bi-directional encoder and a unidirectional decoder. We kept 4 layers in both the encoder & decoder with embedding size set to 512. The batch size was set to 64 and a dropout rate of 0.3. We used Adam optimizer (Kingma and Ba, 2014) for our experiments. Our multilingual model is trained with all the same hyperparameters as our baseline model except that the training data is a combination of Hindi-English & Gujarati-English parallel data.

5 Results

In this section, we report the BLEU (Papineni et al., 2002) scores on the test sets provided in WMT19. Our simple NMT model which is an attention-based LSTM encoder-decoder model achieves a BLEU score of 6.2 on the test set. Our multilingual model which is trained with the help of Hindi-English parallel corpus attains a BLEU score of 9.8, showing a gain of +3.6 BLEU points on the same test set.

Table 2: WMT19 evaluation of our systems

System	BLEU
encoder-decoder + attention	6.2
Multilingual model	**9.8(+3.6)**

6 Additional Transformer Experiments

In this section, we present a set of experiments and results post WMT19 shared task involving the Transformer (Vaswani et al., 2017) architecture. We used the Transformer-Base architecture in this set of experiments with the rest of the pipeline being kept same as described before. We used 6 layers in both the encoder decoder with embedding size set to 512. The batch size was 2048 tokens & a dropout of 0.3. We used Adam optimizer for our experiments. During inference time, we averaged the checkpoints of the model at different epochs to obtain better results than a single checkpoint. In the multilingual Transformer experiments, we employ script conversion technique for its merits described before.

In table 3, we provide the results of our Transformer experiments and also compare it to other systems submitted to WMT19.

Table 3: Our Transformer models vs other systems at WMT19

System	BLEU
Transformer	4.28
Multilingual Transformer	15.78 (+11.5)
+ Averaging	**16.49 (+0.71)**
NICT (Unsupervised MT)	9.6
NICT (Transfer Learning)	18.6
NEU (WMT19 Best)	26.5

7 Conclusion & Future Work

We believe that NMT is a promising approach for Machine Translation for low resource languages. But we need various techniques to handle the data scarcity problem. Transfer Learning and Multilingual Machine Translation are two important areas of research that tackles this problem. In this paper, we showed that how Multilingual MT models are more effective than the individually trained MT models for a low resource language pair. We presented our results on the Gujarati→English language pair and achieved significant BLEU improvements. The Multilingual NMT model we presented in this paper is a many-to-one model. In future, we will work on building effective one-to-many Multilingual NMT systems.

References

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*.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. 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)*, volume 1, pages 1723–1732.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. *arXiv preprint arXiv:1601.01073*.

Thanh-Le Ha, Jan Niehues, and Alexander Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. *arXiv preprint arXiv:1611.04798*.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tieyan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *Advances in Neural Information Processing Systems*, pages 820–828.

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

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Tom Kocmi and Ondřej Bojar. 2018. Trivial transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1809.00357*.

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 ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2017. The iit bombay english-hindi parallel corpus. *arXiv preprint arXiv:1710.02855*.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. *arXiv preprint arXiv:1804.07755*.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Toan Q Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. *arXiv preprint arXiv:1708.09803*.

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*, pages 311–318. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015a. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015b. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

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.

Lijun Wu, Yingce Xia, Li Zhao, Fei Tian, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2017. Adversarial neural machine translation. *arXiv preprint arXiv:1704.06933*.

Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. *arXiv preprint arXiv:1601.00710*.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*.