CVIT-MT Systems for WAT-2018

Jerin Philip[†], Vinay P. Namboodiri[‡] and C.V. Jawahar[†]
[†] CVIT, IIIT Hyderabad, [‡] IIT-Kanpur
jerin.philip@research.iiit.ac.in
vinay.pn@cse.iitk.ac.in, jawahar@iiit.ac.in

Abstract

This document describes the machine translation system used in the submissions of IIIT-Hyderabad (CVIT-MT) for the WAT-2018 English-Hindi translation task. Performance is evaluated on the associated corpus provided by the organizers. We experimented with convolutional sequence to sequence architectures. We also train with additional data obtained through backtranslation.

1 Introduction

Innovations in Neural Machine Translation (NMT) have led to success in many machine translation tasks, often outperforming Statistical Machine Translation (SMT) techniques. Similar to many other language pairs, NMT based approaches have been attempted for the English-Hindi language pair as well (e.g. the WAT-2017 submission (Wang et al., 2017)). Hindi continue to remain as a low resource language demanding further attention from Natural Language Processing (NLP, Machine Learning ML and other related communities. The Hindi-English pair has limited availability of sentence level aligned bitext as parallel corpora.

Lack of sufficient data for Indian languages motivated us to explore techniques that can help in low-resource situations. Recent works (such as (Edunov et al., 2018; Lample et al., 2018)) point to the use of iterative backtranslation to improve translation in low resource languages or under the unavailability of parallel corpora.

This paper describes an overview of the submission from IIIT Hyderabad (CVIT-MT) in WAT-2018(Nakazawa et al., 2018) for for the Hindi-English and English-Hindi translation tasks of the mixed domain tasks. In Section 2, we describe the components constituting our pipeline, following which in Section 3 we provide the details of the data used and procedure used for the training. Section 4 summarizes our results for WAT-2018. Finally in Section 5 we include additional results using newer architectures. We conclude our observations in Section 6.

2 System Description

In this section, we describe the details associated with the tokenization, architecture and data augmentation. These are the three components that helped in obtaining superior results on the corpus provided by the organizers of WAT-2018.

2.1 Tokenization

A popular method of addressing rare-words without compromising coverage of the entire corpus was Byte Pair Encoding (BPE) (Sennrich et al., 2016b), which used a deterministic greedy compression based algorithm to bring the vocabulary down to a finite feasible value.

SentencePiece (Kudo, 2018) builds on top of byte pair encoding. Unlike BPE, which is agnostic to language, SentencePiece gives the most likely derivation of a sentence composed of subword units. This setting reduces to character level in case a completely unknown sentence/word is provided, and the translation model also learns to transliterate. We use SentencePiece for its merits mentioned above.

2.2 Convolutional Sequence to Sequence Learning

In our submission, we employ the Convolutional Sequence to Sequence architecture (CONVS2S) (Gehring et al., 2017). CONVS2S follows an encoder decoder architecture. This has the advantage of being faster than the popular Recurrent Neural Network (RNN) based encoder decoder architectures with attention. This is because the context is built through multiple inputs stacking k convolution blocks ($O(\frac{n}{k})$) with the ability to build in parallel representations for multiple parts of the sentence, unlike through time in the RNN ($O(n)$).

A 1-D convolutional filter of width w with two channels at the output sliding over the embeddings of the text inputs constitute a basic convolutional block. Output of one channel builds up context representation and the other is used to enable gating through Gated Linear Units (GLUs) (Dauphin et al., 2017). The encoder is constructed by stacking k of the above setup, creating a receptive field controlled by w and k. The decoder is similar to the encoder in architecture, with a fully connected layer projecting output to vocabulary size.

2.3 Backtranslation

Backtranslation is a widely tried and tested data augmentation method, proposed for aiding NMT in languages low on parallel resources using available monolingual data by Sennrich *et. al* (2016a). The method works by first training a model in the low to high resource direction followed by using this model on monolingual data. The process provides more authentic sentences in the resource-scarce language and close approximation of its translation in the high resource language. It has been empirically shown that synthetic data alone generated through backtranslation can attain upto 83% of the performance using proper bitext (Edunov et al., 2018).

In the next section, we describe how the components explained above are implemented and used in training - including generating dataset, preprocessing and filtering the training samples, hyperparameters of the architectures in place and evaluations.

3 Experimental Setup

3.1 Dataset

In our experiments, we use the training data provided by organizers. In addition, we also use data obtained from translated Hindi content available on Internet. Top level statistics of the data used are provided in Table 1.

Dataset	Pairs	Tokens	
		hi	en
IITB train	1,492,827	22.2M	20.6M
IITB train†	923,377	20.3M	18.9M
National News	2,495,129	41.2M	39.0M
Backtranslated	5,653,644	77.5M	91.9M
IITB dev	505	10,656	10,174
IITB test	2,507	49,394	57,037

Table 1: Descriptions of the corpora used, IITB train† is a filtered version of the IITB train corpus.

The training corpus provided by the organizers, hereafter denoted by IITB-corpus consists of data from mixed domains. There are roughly 1.5M samples in training data from diverse sources, while the development and test sets are from newspaper crawls. In addition to this, monolingual data collected by the organizers from several sources are used in our backtranslation enabled attempts at training an NMT system. There are 45M samples in the monolingual corpus provided.

We enhanced the training data with additional pairs, but automatically translated. Note that no manual translation was used to create additional data. We obtain 2.5M Hindi sentences automatically translated to English from newspapers and similar resources, obtained from Internet. This data is some what domain specific. They are primarily, from news articles related to national news. This is mentioned as National News in Table 1.

We also create a parallel corpus through backtranslation using the organizers monolingual Hindi data hereafter denoted by Backtranslated, the details of which are also included in Table 1 and the methods of creation elaborated in Section 3.3.

3.2 Data Processing

We train separate SentencePiece models using official implementation available online [1] with vocabulary restricted to 8000 units to function as a learned tokenizer for both English and Hindi. We use the unigram model, which gives language aware tokenization.

To filter any noisy content from IITB corpus, *langdetect*[2] and removed every pair which had probability of being in the respective language less than 0.95. This gave us roughly 0.92M pairs for training, from IITB corpus and is indicated as IITB train† in Table 1. English data is kept true-cased, which we found to have better results consistently with our NMT model.

3.3 Training

In our experiments we use the fairseq [3] toolkit. For the tasks in this submission we use the CONVS2S model.

The encoder and decoder embeddings have a dimension of 512. The hidden units in the encoder and decoder are also 512 dimensional, following Gehring *et. al* (2017). We use convolutional filters of width 3 and 20 layers stacked for both the encoder and decoder. A dropout with probability 0.1 is put in-place right after the embeddings layer for better generalization. The training is run in batches of maximum 4000 tokens at a time, which is on an average 140 sample sentences per batch. The model is trained to minimize the categorical cross-entropy loss at the token level using Nestorov accelerated gradient descent. Decoding is performed through beam search with a beam width of 10.

We run training using four NVIDIA 1080Ti-s until validation loss hasn't improved for 3 epochs straight. The training time was roughly 2 days and stopping around 30-40 epochs.

We keep our model hyperparameters constant as specified across experiments and work with different combinations of corpora created from augmenting the National News dataset and official parallel corpora. For creating the Backtranslated corpus, we use a model trained to translate from Hindi to English using both National News and IITB corpus. We filter the obtained pairs using confidence of translation obtained from the beam-score and further to pairs with a length between 10 and 30 tokens.

3.4 Evaluations

We report Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002), Rank-based Intuitive Bilingual Evaluation Score (RIBES) (Isozaki et al., 2010), Adequacy-fluency metrics (AM-FM) (Banchs et al., 2015) for all our attempts and scores from WAT-2018 human evaluations(Human in Table 2) when available.

BLEU is computed as the geometric mean of unigram, bigram, trigram and 4-gram precision multiplied by a brevity penalty (BP). BLEU ranges from 0 to 1, but the values reported in Tables 2 and 3 are in percentages. RIBES, also giving a value in $[0, 1]$ was proposed to tackle shortcomings of BLEU in distant language pairs, where changes in word ordering deteriorates BLEU.

4 Discussions

The results using our systems for WAT-2018 are presented in Table 2 (see some additional results in Table 3). The first part of the table consists of results on combinations of datasets and augmentations. All values are for models trained from scratch. In the second part, the current leader board is indicated for comparison. Note that entries in this part don't correspond to a single submission, but the values corresponding to the best in the respective metric.

Our submission based on the combination National News and IITB corpus tops human evaluation in Hindi to English, and ranks second in English to Hindi. We demonstrate the possibility of distilling knowledge of online available sources into a usable translation model. We successfully use the CONVS2S architecture along with SentencePiece to obtain results comparable to the top submissions. Our experiments also indicates data augmentation using backtranslation positively works for the Hindi-English pair.

5 Additional Transformer Experiments

In this section, we present a set of experiments and results post WAT-2018 involving the Transformer

[1] https://github.com/google/sentencepiece

[2] https://github.com/Mimino666/langdetect

[3] http://github.com/pytorch/fairseq (formerly fairseq-py)

Dataset	en-hi				hi-en			
	BLEU	RIBES	AM-FM	Human	BLEU	RIBES	AM-FM	Human
IITB train†	13.25	0.695113	0.647220	-	11.83	0.675462	0.572900	-
National News	18.77	0.748008	0.697630	-	19.53	0.745764	0.614260	-
+IITB train†	19.69	0.758365	0.699810	69.50	20.63	0.751883	0.623240	72.25
Backtranslated	16.77	0.714197	0.664330	50.50	-	-	-	-
2017 Best	21.39	0.749660	0.688770	64.50	22.44	0.750921	0.629530	68.25
2018 Best	20.28	0.761582	0.704220	77.00	17.80	0.731727	0.611090	67.25

Table 2: Quantitative results of translating English to Hindi and vice versa.

Architecture (Vaswani et al., 2017). Two variants of the architecture - Transformer Base and Transformer Big outperformed then state of the art CONVS2S models in the WMT German-English and French-English translation tasks.

We used the Transformer-Base architecture in further experiments with the National News + IITB corpus where CONVS2S performed the best, with the rest of the pipeline being kept same as described before. We went with the default hyperparameters provided by *fairseq* framework - which did not give us impressive results.

Following Popel and Bojar (2018), we modified the hyperparameters for initial warm-up steps of 16000 without any learning rate decay, starting from a learning rate of 0.25, followed by an exponential decay of learning rate. We also had to enable delayed gradient updates (Ott et al., 2018) to simulate a larger batch on smaller GPU before the model demonstrated any learning. During inference time, we averaged checkpoints of the model at different epochs once the loss on the development set had plateaued to obtain better results than a single checkpoint.

Architecture	BLEU	RIBES	AM-FM
CONVS2S	19.69	0.758365	0.699810
Transformer	21.10	0.771549	0.712200
+Averaging	**21.57**	**0.773923**	**0.712110**

Table 3: Transformer-Base vs CONVS2S on National News + IITB corpus, for English to Hindi direction.

In Table 3, we compare the performance of the transformer with that CONVS2S. Consistent with observations in languages like German-English and French-English, the transformer network produces better results than CONVS2S on all metrics. The averaged model performs the best in all metrics in English to Hindi translation task, at the time of writing this paper.

6 Observations

We believe that NMT is a promising approach for Indian language machine translation for obtaining reasonably accurate solutions. Our initial results reported here confirms this. In addition, we believe, the popular data augmentation methods are effective and feasible for many low-resource machine translation settings. We see the direct utility of the advances in NMT for many western language pairs on English-Hindi in terms of ideas and architectures. At the same time, we also believe, there is much more to do for making them effective on Indian languages.

Acknowledgments

We thank the organizers for systematically setting up this task, and for the very useful resources. We also thank the larger language processing group at IIIT Hyderabad for the encouragement, support and insights.

References

Rafael E Banchs, Luis F D'Haro, and Haizhou Li. 2015. Adequacy-fluency metrics: Evaluating MT in the continuous space model framework. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 23(3):472–482.

Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. 2017. Language Modeling with Gated Convolutional Networks. In *International Conference on Machine Learning*, pages 933–941.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding Back-Translation at Scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional Sequence to Sequence Learning. In *International Conference on Machine Learning*, pages 1243–1252.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 944–952. Association for Computational Linguistics.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations (ICLR)*.

Toshiaki Nakazawa, Shohei Higashiyama, Chenchen Ding, Raj Dabre, Anoop Kunchukuttan, Win Pa Pa, Isao Goto, Hideya Mino, Katsuhito Sudoh, and Sadao Kurohashi. 2018. Overview of the 5th workshop on asian translation. In *Proceedings of the 5th Workshop on Asian Translation (WAT2018)*, Hong Kong, China, December.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 1–9. 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*, pages 311–318. Association for Computational Linguistics.

Martin Popel and Ondřej Bojar. 2018. Training Tips for the Transformer Model. *The Prague Bulletin of Mathematical Linguistics*, pages 43–70.

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.

Boli Wang, Zhixing Tan, Jinming Hu, Yidong Chen, et al. 2017. XMU neural machine translation systems for WAT 2017. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 95–98.