IITP-MT System for Gujarati-English News Translation Task at WMT 2019

Sukanta Sen, Kamal Kumar Gupta, Asif Ekbal, Pushpak Bhattacharyya
Department of Computer Science and Engineering
Indian Institute of Technology Patna
{sukanta.pcs15,kamal.pcs17,asif,pb}@iitp.ac.in

Abstract

We describe our submission to WMT 2019 News translation shared task for Gujarati-English language pair. We submit constrained systems, i.e, we rely on the data provided for this language pair and do not use any external data. We train Transformer based subword-level neural machine translation (NMT) system using original parallel corpus along with synthetic parallel corpus obtained through back-translation of monolingual data. Our primary systems achieve BLEU scores of 10.4 and 8.1 for Gujarati→English and English→Gujarati, respectively. We observe that incorporating monolingual data through back-translation improves the BLEU score significantly over baseline NMT and SMT systems for this language pair.

1 Introduction

In this paper, we describe the system that we submit to the WMT 2019[1] news translation shared task (Bojar et al., 2019). We participate in Gujarati-English language pair and submit two systems: English→Gujarati and Gujarati→English. Gujarati language belongs to Indo-Aryan language family and is spoken predominantly in the Indian state of Gujarat. It is a low-resource language as only a few thousands parallel sentences are available, which are not enough to train a neural machine translation (NMT) system as well statistical machine translation (SMT) system. Gujarati-English is a distant language pair and they have different linguistic properties including syntax, morphology, word order etc. English follows subject-verb-object order while Gujarati follows subject-object-verb order.

NMT (Kalchbrenner and Blunsom, 2013; Cho et al., 2014; Sutskever et al., 2014; Bahdanau et al., 2015) has recently become dominant paradigm for machine translation (MT) achieving state-of-the-art on standard benchmark data sets for many language pairs. As opposed to SMT, NMT systems are trained in an end-to-end manner. Training an effective NMT requires a huge amount of high-quality parallel corpus and in absence of that, an NMT system tends to perform poorly (Koehn and Knowles, 2017). However, back-translation (Sennrich et al., 2016) has been shown to improve NMT systems in such a situation. In this work, we train a SMT system and an NMT system for both English→Gujarati and Gujarati→English using the original training data. SMT systems are also used to generate synthetic parallel corpora through back-translation of monolingual data from English news crawl and Gujarati Wikipedia dumps. These corpora along with the original training corpora are used to improve the baseline NMT systems. All the SMT and NMT systems are trained at subword level.

Our SMT systems are standard phrase-based SMT systems (Koehn et al., 2003), and NMT systems are based on Transformer (Vaswani et al., 2017) architecture. Experiments show that NMT systems achieve BLEU (Papineni et al., 2002) scores of 10.4 and 8.1 for Gujarati→English and English→Gujarati, respectively, outperforming the baseline SMT systems even in the absence of enough-sized parallel data.

Rest of the paper is arranged in following manner: Section 2 gives brief introduction of the Transformer architecture that we used for NMT training, Section 3 describes the task, Section 4 describes the submitted systems, Section 5 gives various evaluation scores for English-Gujarati translation pair, and finally, Section 6 concludes the work.

[1] http://www.statmt.org/wmt19/translation-task.html

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers (Day 1) pages 407–411
Florence, Italy, August 1-2, 2019. ©2019 Association for Computational Linguistics

2 Transformer Architecture

Recurrent neural network based encoder-decoder NMT architecture (Cho et al., 2014; Sutskever et al., 2014; Bahdanau et al., 2015) deals with input/output sentences word-by-word sequentially, which prevents the model from parallel computation. Vaswani et al. (2017) came up with a highly parallelizable architecture called Transformer which uses the self-attention to better encode a sequences. Self-attention is used in the architecture to calculate attention between a word and the other words in the sentence itself. Encoder and decoder both are stack of 6 identical layers. Each layer in encoder has two sub-layers: *i.* multi-head self attention mechanism and *ii.* position wise feed forward network. Each sub-layer is associated with residual connections, followed by layer normalization. Multi-head attention computes the attention multiple times for each word. Since their is no sequence to sequence encoding, positional encoding is used to encode the sequence information.

3 Task Description

This task focuses on translating news domain corpus and this year, Gujarati language is introduced for the first time in a WMT shared task. Gujarati is a low-resource language and not many results have been reported in machine translation involving this language. Also, there was no standard test set for this language pair. So introduction of this language pair will help in further research for this language pair.

As Gujarati does not have enough parallel data, the data that are provided for this shared task are mainly from WikiTitles which consists of only 11,671 parallel titles. Apart from that, few publicly available domain specific parallel data that are provided are: Bible corpus (Christodouloupoulos and Steedman, 2015); a localization extracted from OPUS[2]; parallel corpus extracted from Wikipedia; crawled corpus produced for this task; and monolingual Wikipedia dumps.

4 System Description

We participated in Gujarati-English pair only and we submit for both directions: English→Gujarati and Gujarati→English. As Gujarati is a low-resource language and only a little amount of parallel data is available, we explore the back-translation technique for this pair. Also our models are based on Transformer as it has become state of the art for machine translation for many language pairs. We train systems at subword level. For back-translation, we train a phrase-based SMT (Koehn et al., 2003) system for each system in reverse direction. Using these SMT systems, monolingual sentences (for both Gujarati and English) are translated to create synthetic parallel data having original monolingual sentences at target and translated sentences at source side. These synthetic parallel data, along with the original parallel data are used to train a transformer based NMT system for each direction.

4.1 Dataset

Sources	#Sentences
Parallel	
Bible	7,807
govin-clean.gu-en.tsv	10,650
opus.gu-en.tsv	107,637
wikipedia.gu-en.tsv	18,033
wikititles-v1.gu-en.tsv	11,671
Total	155,798
Monolingual	
Gujarati (Wikipedia dump)	382,881
English (News crawl)	1,000,000

Table 1: Training data sources and number of sentences.

The datasets that we use for training are shown in the Table 1, which combine to a total of 155,798 parallel sentences. These parallel data are compiled from different sources. The compiled datasets are Bible[3], govin-clean.gu-en.tsv[4], opus.gu-en.tsv[5], wikipedia.gu-en.tsv[6] and wikititles-v1.gu-en.tsv[7]. We use *newsdev2019* for tuning the model, which has 1,998 parallel sentences.

[2] http://opus.nlpl.eu

[3] http://data.statmt.org/wmt19/translation-task/bible.gu-en.tsv.gz

[4] http://data.statmt.org/wmt19/translation-task/govin-raw.gu-en.tsv.gz

[5] http://data.statmt.org/wmt19/translation-task/opus.gu-en.tsv.gz

[6] http://data.statmt.org/wmt19/translation-task/wikipedia.gu-en.tsv.gz

[7] http://data.statmt.org/wikititles/v1/wikititles-v1.gu-en.tsv.gz

System	BLEU	BLEU-cased	TER	CharactTER
		English→Gujarati		
PBSMT	5.2	5.2	0.987	0.782
Transformer	4.0	4.0	1.005	0.884
Transformer + Synth	8.1	8.1	0.919	0.763
		Gujarati→English		
PBSMT	7.3	6.3	0.883	0.817
Transformer	5.5	5.1	0.905	0.859
Transformer + Synth	10.4	9.4	0.828	0.774

Table 2: BLEU scores of different SMT and NMT based systems. Synth: Synthetic data

Apart from these parallel data, we use monolingual English (news crawl) and Gujarati (Wikipedia dumps) sentences for synthetic parallel data creation. After training two models i.e. English→Gujarati and Gujarati→English using the parallel data mentioned in Table 1, English and Gujarati monolingual sentences are back translated respectively.

4.2 Experimental Setup

We train phrase based statistical system (PBSMT) (Koehn et al., 2003) as well as Transformer (Vaswani et al., 2017) based neural system for comparing their performance under low-resource conditions. In addition to that, PBSMT are used to genrate synthetic parallel data. PBSMT systems are trained only on original training data, while neural based models are trained on original training data (*Transfomer* in Table 2), and also with synthetic parallel data in addition to original data (*Transfomer+Synth* in Table 2). Synthetic parallel data are obtained through back-translation of a target monolingual corpus into source using PBSMT system. We use Moses (Koehn et al., 2007) toolkit for PBSMT training and Sockeye (Hieber et al., 2017) toolkit for NMT training. Some preprocessing of data is required before using it for experiment. English data is tokenized using moses tokenizer, and truecased. For tokenizing Gujarati data, we use indic_nlp_library[8]. After tokeninzation and truecasing, we subword (Sennrich et al., 2015) all original data. We apply 10,000 BPE merge operations over English and Gujarati data independently.

For back-translation of monolingual data, two PBSMT models English→Gujarati and Gujarati→English are trained over original available parallel subworded corpora. 4-gram language model is trained using KenLM (Heafield, 2011). For word alignment, we use GIZA++ (Och and Ney, 2003) with grow-diag-final-and heuristics. Model is tuned with Minimum Error Rate Training (Och, 2003). After these two models are trained, monolingual subworded data from both English and Gujarati are back-translated using English→Gujarati and Gujarati→English PBSMT model, respectively. We merge the back translated data with original parallel data to have larger parallel corpora for Gujarati→English and English→Gujarati translation directions.

Finally, with the augmented parallel corpora, we train one Transformer based NMT model for each direction. We use the following hyperparameters values of Sockeye toolkit: 6 layers in both encoder and decoder, word embedding size of 512, hidden size of 512, maximum input length of 50 tokens, Adam optimizer, word batch size 1000, attention type is dot, learning rate of 0.0002. The rest of the hyper-parameters are set to the default values in Sockeye. We use early stopping criteria for terminating the training on the validation set of 1,998 parallel sentences.

5 Results

The official automatic evaluation uses the following metrics: BLEU (Papineni et al., 2002), TER (Snover et al., 2006), CharactTER (Wang et al., 2016). The official scores are shown in the Table 2. Phrase-base SMT (PBSMT) obtains BLEU scores of 5.2 and 7.3 for English→Gujarati and Gujarati→Englsih, respectively. Whereas, baseline NMT (*Transformer*) obtains lower BLEU scores of 4.0 and 5.5 for the same directions. Though, SMT systems outperforms baseline NMT systems trained using small amount of original parallel data only. We observe from the Table 2 that Transformer with synthetic (*Transformer +*

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

Gujarati→English

Ave.	Ave. z	System
64.8	0.210	NEU
61.7	0.126	UEDIN
59.4	0.100	GTCOM-Primary
60.8	0.090	CUNI-T2T-transfer
59.4	0.066	aylien-mt-multilingual
59.3	0.044	NICT
51.3	−0.189	online-G
50.9	−0.192	IITP-MT
48.0	−0.277	UdS-DFKI
47.4	−0.296	IIITH-MT
41.1	−0.598	Ju-Saarland

English→Gujarati

Ave.	Ave. z	System
73.1	0.701	HUMAN
72.2	0.663	online-B
66.8	0.597	GTCOM-Primary
60.2	0.318	MSRA-CrossBERT
58.3	0.305	UEDIN
55.9	0.254	CUNI-T2T-transfer
52.7	−0.079	Ju-Saarland-clean-num-135-bpe
35.2	−0.458	IITP-MT
38.8	−0.465	NICT
39.1	−0.490	online-G
33.1	−0.502	online-X
33.2	−0.718	UdS-DFKI

Table 3: Preliminary official results of WMT 2019 news translation task for Gujarati-English pair. Systems ordered by DA score z-score; systems within a cluster are considered tied; lines indicate clusters according to Wilcoxon rank-sum test $p < 0.05$; grayed entry indicates resources that fall outside the constraints provided.

Synth) data obtained through back-translation of monolingual data, outperforms the baseline SMT systems with a margin of 2.9 and 3.1 BELU points. Also, as a result of augmenting back-translated data with original training data, we obtain improvement of of 4.7 and 5.3 BLEU points over baseline NMT for English→Gujarati and Gujarati→English, respectively. The official preliminary human evaluation results are shown in the Table 3.

6 Conclusion

In this paper, we described our submission to the WMT 2019 News translation shared task for Gujarati-English language pair. This is the first time Gujarati language is introduced in a WMT shared task. We submit Transformer based NMT systems for English-Gujarati language pair. Since the number of parallel sentences in training set are very less and many sentences have length of only 2-3 tokens, BLEU scores for English-Gujarati pair using only available parallel corpus are very low (4.0 and 5.1 for English→Gujarati and Gujarati→English, respectively). So we use monolingual sentences for both languages to create synthetic parallel data through back-translation, and merged them with original parallel data. We obtained improved BLEU scores of 8.1 and 10.4, respectively.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of the 3rd International Conference on Learning Representation (ICLR 2015).*

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Christof Monz, Mathias Müller, and Matt Post. 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.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the Properties of Neural Machine Translation: Encoder-decoder Approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: the Bible in 100 languages. *Language resources and evaluation*, 49(2):375–395.

Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197. Association for Computational Linguistics.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2017. Sockeye: A toolkit for neural machine translation. *arXiv preprint arXiv:1712.05690.*

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent Continuous Translation Models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, pages 1700–1709.

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.

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.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54. Association for Computational Linguistics.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 160–167. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.

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, Philadelphia, Pennsylvania.

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany*.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, volume 200.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Proceedings of Advances in neural information processing systems (NIPS 2014)*, pages 3104–3112.

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.

Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016. Character: Translation edit rate on character level. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, volume 2, pages 505–510.