Contact Relatedness can help improve multilingual NMT: Microsoft STCI-MT @ WMT20

Vikrant Goyal, Anoop Kunchukuttan, Rahul Kejriwal, Siddharth Jain, Amit Bhagwat
STC India, Microsoft, Hyderabad
{vikgoyal, ankunchu, rakejriw, sija, amitb}@microsoft.com

Abstract

We describe our submission for the English→Tamil and Tamil→English news translation shared task. In this submission, we focus on exploring if a low-resource language (Tamil) can benefit from a high-resource language (Hindi) with which it shares contact relatedness. We show utilizing contact relatedness via multilingual NMT can significantly improve translation quality for English-Tamil translation.

1 Introduction

In recent years, Neural Machine Translation (Luong et al., 2015; Bahdanau et al., 2015; Johnson et al., 2017; Wu et al., 2018; Vaswani et al., 2017) (NMT) has become the most prominent approach to Machine Translation (MT) due to its simplicity, generality and effectiveness. In NMT, a single neural network often consisting of an encoder and a decoder is used to directly maximize the conditional probabilities of target sentences given the source sentences in an end-to-end paradigm. NMT models have been shown to surpass the performance of previously dominant statistical machine translation (SMT) (Koehn, 2009) on many well-established translation tasks. However, in order to obtain good translation quality, NMT systems tend to require very large parallel training corpora (Koehn and Knowles, 2017). Such corpora are not yet available for many language pairs.

The Indian subcontinent forms a *linguistic area* where languages from the Dravidian and Indo-Aryan families have been in **contact** for a long time leading to significant sharing of vocabulary and a convergence of linguistic features (Emeneau, 1956). Tamil is a major language from the Dravidian language family spoken in Southern India while Hindi is a widely spoken Indo-Aryan language. Kunchukuttan and Bhattacharyya (2020) estimate that lexical similarity between Hindi and Tamil to be around 27% in terms of character LCSR (Melamed, 1995), while multiple works have shown that language representations of Tamil and Hindi cluster in the same neighbourhood in a multilingual vector space (Kudugunta et al., 2019; Oncevay et al., 2020).

While English-Tamil parallel corpora is limited, more parallel corpora is available for English-Hindi. In this paper, we explore if English-Hindi can improve English-Tamil machine translation due to the similarities between Hindi and Tamil on account of contact relatedness. To this end, we train multilingual NMT models for English-Hindi and English Tamil (and vice-versa). Previous work has explored whether high-resource languages can transfer knowledge to genetically-related low-resource languages (Nguyen and Chiang, 2017; Dabre et al., 2017). In contrast, we explore if **contact relatedness** can benefit low resource languages. We further explore if reducing the divergence between Tamil and Hindi data by representing them in the same script is beneficial. In addition, we explored target agreement models, tagged and noisy back-translation in our submission.

2 Neural Machine Translation

Given a bilingual sentence pair (x, y), an NMT model learns its parameters θ by maximizing the log-likelihood $P(y|x;\theta)$, which is usually decomposed into the product of the conditional probability of each target word: $P(y|x;\theta) = \prod_{t=1}^{m} P_\theta(y_t|y_1, y_2, .., y_{t-1}, x; \theta)$, where m is the length of sentence y.

An encoder-decoder framework (Bahdanau et al., 2015; Luong et al., 2015; Gehring et al., 2017; Vaswani et al., 2017) is usually adopted to model the conditional probability $P(y|x;\theta)$. The encoder maps the input sentence x into a set of hidden representations h, and the decoder generates the target token y_t at position t using the previously generated target tokens $y_{<t}$ and the source representations h. Both the encoder and decoder can be implemented by different structure of neural models, such as RNN (LSTM/GRU) (Bahdanau et al., 2015; Luong et al., 2015), CNN (Gehring et al., 2017) and self-attention (Vaswani et al., 2017). Besides the basic component of the encoder and decoder, a source-target attention mechanism (Bahdanau et al., 2015) is usually adopted to selectively focus on the source representations when generating a target token.

The Transformer (Vaswani et al., 2017) model is the state-of-the-art NMT model relying completely on self-attention mechanism to compute representations of its input and output without using recurrent neural networks (RNN) or convolutional neural networks (CNN). In this work, we use the Transformer architecture in all of our NMT models. We use smaller capacity networks compared to *Transformer-base*, given the smaller size of the parallel data.

3 Multilingual Learning for NMT

The objective of multilingual learning for NMT is to construct a single model for translating to and from multiple languages. Multilingual models can help improve performance of low-resource languages by transferring from high-resource related languages they are trained jointly with. Firat et al. (2017) introduced a many-to-many system, which still relied upon separate encoders and decoders for each language along with a shared attention mechanism. In contrast, Johnson et al. (2017) introduced a “language flag”-based approach that shares the attention mechanism and a single encoder-decoder network to enable multilingual models. A language flag or token is part to the input sequence to indicate which direction to translate to. The decoder learns to generate the target given this input. This approach has been shown to be simple and effective and we use this in our multilingual models. As mentioned earlier, we train a joint Hindi,Tamil to English model as well as a joint English to Hindi,Tamil model. Our hypothesis is that the contact relatedness between Hindi and Tamil will help transfer knowledge from Hindi to Tamil effectively.

Tamil and Hindi use different scripts. However, it is possible to map almost all Devanagari (the script used for Hindi) characters to Tamil characters. This mapping is deterministic but lossy since the Tamil character set is smaller than the Devanagari character set. Such mapping will help to utilize the lexical similarity between the two languages directly. Hence, we convert all Hindi data to Tamil script during a pre-processing step. We also report results of our MultiNMT models without script conversion of Hindi to Tamil.

4 Backtranslation

Backtranslation (BT) (Sennrich et al., 2016a) is a widely used data augmentation method where the reverse direction is used to translate sentences from target-side monolingual data into the source language. This synthetic parallel data is combined with the actual parallel data to re-train the model leading to better language modelling on the target-side, regularization and target domain adaptation. Backtranslation is particularly useful for low-resource languages. We use backtranslation to augment our multilingual models. The backtranslation data is generated by multilingual models in the reverse direction, hence some implicit multilingual transfer is incorporated in the backtranslated data also.

4.1 Noisy and Tagged Backtranslation

Backtranslation typically uses beam search (Sennrich et al., 2016a) or just greedy search (Lample et al., 2018a,b) to generate synthetic source sentences. Both are approximate algorithms to identify the maximum a-posteriori (MAP) output, i.e. the sentence with the largest estimated probability given an input. Beam is generally successful in finding high probability outputs (Ott et al., 2018). However, MAP prediction can lead to less rich translations since it always favors the most likely alternative in case of ambiguity. Edunov et al. (2018) argue that this is also problematic for a data augmentation scheme such as backtranslation. Beam and greedy search focus on the head of the model distribution which results in very regular synthetic source sentences that do not properly cover the true data distribution. Following the approach proposed by Edunov et al. (2018), we apply noising to the beam search outputs. In particular, we transform source sentences with three types of noise: deleting words with probability 0.1, replacing words by a filler token with probability 0.1, and swapping words which is implemented as a random permutation over the tokens, drawn from the uniform distribution but restricted to swapping words no further than three positions apart.

Caswell et al. (2019) showed that main purpose of the synthetic noise is not to diversify the source but simply to indicate that the given source is synthetic. They proposed to prepend the input sequences of the synthetic data with a reserved token like <BT> to indicate that the given source is synthetic. In this paper, we experiment with both Noisy BT and Tagged BT.

5 Target Agreement

Due to the autoregressive structure, current NMT systems usually suffer from the so-called exposure bias problem (Bengio et al., 2015): during inference, true previous target tokens are unavailable and replaced by tokens generated by the model itself, thus mistakes made early can mislead subsequent translation, yielding unsatisfactory translations with good prefixes but bad suffixes. Such an issue can become severe as sequence length increases. Zhang et al. (2019) showed that the impact of this can be reduced by augmenting the training data with synthetic targets generated by a left-to-right (L2R) and a right-to-left (R2L) translation model. The directionality of the synthetic targets ensures that decoder input distribution becomes noisier (as happens at runtime) along one side of the target. The augmented data thus serves to reduce the divergence in the decoder input distribution. This is especially relevant to low-resource language scenarios where the model is not as robust to the decoder input distribution.

6 Experimental Settings

6.1 Dataset

We train our models only on the parallel data provided for the task (see Table 1 for dataset details). For backtraslation, we randomly selected 10M sentences from the newscrawl 2019 English monolingual corpora. For Tamil monolingual corpora, used the entire newscrawl corpus (0.7M), Wikipedia corpus and part of the CommonCrawl data made available for a consolidated corpus of 10M sentences. We use IIT-Bombay Hindi-English parallel corpora v2.0 (Kunchukuttan et al., 2018) containing 1.5M parallel sentences to build our multilingual models. We used UFAL's Tamil-English dev set containing 1,000 parallel sentences for tuning our models.

Dataset	# of Sentences
Wikititles	102,146
Wikimatrix	52,669
PMIndia	39,526
Tanzil (Koran)	93,540
NLPC_UOM	8,945
PIB (CVIT@IIITH)	60,836
MKB (CVIT@IIITH)	5,744
UFAL	166,871
Total	530,277

Table 1: Tamil-English parallel corpus statistics.

6.2 Data Processing

We use the Moses (Koehn et al., 2007) toolkit[1] for lowercasing, tokenization and cleaning the English side of the data. Both Tamil and Hindi data are first normalized and then tokenized. The Hindi data is mapped to Tamil script. We use the Indic NLP library[2] (Kunchukuttan, 2020) for text processing of the Indic languages. We remove all sentences of length greater than 80 words from our training corpus. In all cases, we use BPE subword segmentation (Sennrich et al., 2016b) with 32k merge operations. In case of mulitlingual models, we learn the BPE vocabulary jointly on the Hindi and Tamil data.

6.3 Training Details

For all of our experiments, we use the fairseq (Ott et al., 2019) toolkit[3]. We use the Transformer model with 4 layers in both the encoder and decoder, each with 512 hidden units. The word embedding size is set to 512 and 8 attention heads are used. The training is done in batches of maximum 2048 tokens at a time with dropout set to 0.2. We use the Adam (Kingma and Ba, 2015) optimizer to optimize model parameters with $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\epsilon = 1e - 9$ and we use the same learning rate schedule as Vaswani et al. (2017). We validate the model after each epoch via label smoothed cross entropy loss and perplexity on the development set. We train all our NMT models till convergence where convergence is determined by label smoothed cross entropy loss on the development set. After translation at the test time, we rejoin the translated BPE segments. Finally,

[1] https://github.com/moses-smt/mosesdecoder

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

[3] https://github.com/pytorch/fairseq

we evaluate the accuracy of our translation models using SacreBLEU (Post, 2018).

7 Results and Discussion

We report the SacreBLEU scores on the dev sets and test sets provided in the WMT20 News translation task for both the language directions: Tamil→English (Table 2) and English→Tamil (Table 3). We experimented with Multilingual (MultiNMT), Backtranslation (BT), Noisy Backtranslation (noisyBT) and Tagged Backtranslation (taggedBT) and Target Agreement (TA) models.

We observe that our multilingual models outperform the baseline bilingual models by significant margins. On the newstest2020 set, we see an improve of 2.9 and 1.5 BLEU points respectively in the Tamil→English and English→Tamil directions respectively. Note that the gains translating into multiple targets is lower. However, when backtranslated data is added we observe an improvement of 2.3 BLEU points in the en→ta quality. Note that the backtranslation data was generated via the multilingual ta→en model, hence there is an implicit benefit from multilinguality when using backtranslation. Backtranslation also provides a good improvement in the en→ta. We see that representing Indian language data in the same script was very beneficial for ta→en translation, while it did not help in the other direction. Having disjoint vocabularies during generation possibly helps the model learn distinct language models for Hindi and Tamil.

Our results indicate that Target Agreement and Noisy and Tagged Backtranslation schemes are not helpful in increasing the translation performance of the NMT models for the language pairs of our interest and requires more investigation on low resource language translation tasks. Further analysis is needed to understand why backtranslation variants and target agreement did not show improvements in our setting.

System	newsdev2020	newstest2020
Transformer baseline	10.4	10.0
MultiNMT (no script conversion)	12.5	12.2
MultiNMT	13.0	12.9
MultiNMT+BT	**19.1**	**14.2**
MultiNMT+noisyBT	18.3	14.2
MultiNMT+taggedBT	17.6	13.5
MultiNMT+BT+TA	19.1	14.2

Table 2: Tamil→English (Ta-En) experiment results.

System	newsdev2020	newstest2020
Transformer baseline	6.1	3.5
MultiNMT (no script conversion)	7.5	4.9
MultiNMT	7.4	5.0
MultiNMT+BT	**11.5**	**7.3**
MultiNMT+BT+TA	11.3	7.3

Table 3: English→Tamil (En-Ta) experiment results.

8 Conclusion

We believe contact relatedness can be utilized in the multilingual NMT framework for improving low-resource language translation. Our initial results confirm this for English-Tamil translation aided by English-Hindi data. In addition, we show, that the popular data augmentation methods like backtranslation further helps in increasing the translation performance of Multilingual NMT models.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICML*.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, pages 1171–1179.

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 53–63.

Raj Dabre, Tetsuji Nakagawa, and Hideto Kazawa. 2017. An Empirical Study of Language Relatedness for Transfer Learning in Neural Machine Translation. In *The 31st Pacific Asia Conference on Language, Information and Computation*.

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.

Murray B Emeneau. 1956. India as a lingustic area. *Language*.

Orhan Firat, Kyunghyun Cho, Baskaran Sankaran, Fatos T Yarman Vural, and Yoshua Bengio. 2017. Multi-way, multilingual neural machine translation. *Computer Speech and Language*, 45(C):236–252.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1243–1252. JMLR. org.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's 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. 2015. Adam: A method for stochastic optimization. In *ICML*.

Philipp Koehn. 2009. *Statistical machine translation*. Cambridge University Press.

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.

Sneha Kudugunta, Ankur Bapna, Isaac Caswell, and Orhan Firat. 2019. Investigating multilingual NMT representations at scale. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1565–1575, Hong Kong, China. Association for Computational Linguistics.

Anoop Kunchukuttan. 2020. The IndicNLP Library. `https://github.com/anoopkunchukuttan/indic_nlp_library/blob/master/docs/indicnlp.pdf`.

Anoop Kunchukuttan and Pushpak Bhattacharyya. 2020. Utilizing Language Relatedness to improve Machine Translation: A Case Study on Languages of the Indian Subcontinent. *arxiv pre-print 2003.08925*.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2018. The IIT Bombay English-Hindi Parallel Corpus. In *LREC*.

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

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

Minh-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.

I Dan Melamed. 1995. Automatic evaluation and uniform filter cascades for inducing n-best translation lexicons. In *Third Workshop on Very Large Corpora*.

Toan Q Nguyen and David Chiang. 2017. Transfer Learning across Low-Resource, Related Languages for Neural Machine Translation. In *International Joint Conference on Natural Language Processing*.

Arturo Oncevay, Barry Haddow, and Alexandra Birch. 2020. Bridging linguistic typology and multilingual machine translation with multi-view language representations. *arxiv pre-print 2004.14923*.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. In *ICML*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

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.

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, Fei Tian, Li Zhao, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2018. Adversarial neural machine translation. In *Asian Conference on Machine Learning*, pages 534–549.

Zhirui Zhang, Shuangzhi Wu, Shujie Liu, Mu Li, Ming Zhou, and Tong Xu. 2019. Regularizing neural machine translation by target-bidirectional agreement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 443–450.