Linguistically Motivated Subwords for English-Tamil Translation

Dhar, Prajit; Bisazza, Arianna; van Noord, Gertjan

Published in:
Proceedings of the 5th Conference on Machine Translation (WMT)

Publication date:
2020

Linguistically Motivated Subwords for English-Tamil Translation: University of Groningen's Submission to WMT-2020

Prajit Dhar **Arianna Bisazza** **Gertjan van Noord**

University of Groningen

{p.dhar, a.bisazza, g.j.m.van.noord}@rug.nl

Abstract

This paper describes our submission for the English-Tamil news translation task of WMT-2020. The various techniques and Neural Machine Translation (NMT) models used by our team are presented and discussed, including back-translation, fine-tuning and word dropout. Additionally, our experiments show that using a linguistically motivated subword segmentation technique (Ataman et al., 2017) does not consistently outperform the more widely used, non-linguistically motivated SentencePiece algorithm (Kudo and Richardson, 2018), despite the agglutinative nature of Tamil morphology.

1 Introduction

In this paper we present the neural machine translation (NMT) systems submitted to the WMT-2020 English-Tamil (EN→TA) news translation task. This task is challenging mainly for two reasons:

1. Differing syntax: English is an Indo-European language which is fusional and SVO (Subject-Verb-Object). On the other hard, Tamil is part of the Dravidian language family and is a SOV language that is agglutinative. A good NMT system is expected to discern the various morphological forms on the Tamil target side.

2. Scarcity of training data: Prior to WMT-2020, there existed only a few corpora for parallel EN-TA sentences (Ramasamy et al., 2012; Germann, 2001). This left us with the choice of either only utilizing the low amount of parallel sentences or finding out ways of artificially enlarging the training data.

Through our submission we wish to provide solutions to the following questions:

- Is linguistically motivated subword segmentation beneficial for EN-TA translation?
- Can the addition of TA monolingual data compensate for the small amount of parallel EN-TA sentences despite the domain mismatch?
- Can fine-tuning on a corpus of Indian news improve quality on the WMT news translation task?

We start our paper with a short description of the Tamil language before delving into the various techniques adopted by our submitted NMT systems.

2 Tamil Language

Tamil is a Dravidian language spoken by around 80 million people. Tamil morphology is agglutinative and suffixal, i.e words are formed by suffixing morphemes to a lemma (Annamalai et. al 2014, cited in Sarveswaran et al. (2019)). Tamil suffixes can be either derivational (marking a change in PoS and/or meaning) or inflectional. In particular, nouns in Tamil are inflected for number, gender, case and animacy while verbs are inflected for tense, mood, aspect, negation, interrogation, information about emphasis, speaker perspective, sentience or rationality, and conditional and causal relations. Table 4 shows examples of the case forms in singular for the noun புத்தகம ’book’.

All the aforementioned statements substantiate the fact that Tamil morphology is highly complex. In fact, Ramasamy et al. (2012) identified 716 inflectional rules for nouns and 519 rules for verbs. Furthermore, designing a translation system for Tamil is challenging given the lack of training data (compare the sizes of Japanese and Tamil parallel datasets in WMT 2020, both agglutinative, however having vastly different training data; 25M sentences and 630k, respectively).

Proceedings of the 5th Conference on Machine Translation (WMT), pages 126–133
Online, November 19–20, 2020.

3 Previous Work

One of the earliest automatic translation systems for English→Tamil was by Germann (2001). They created a hybrid statistical/rule-based machine translation (SMT) system and trained it on only 5k EN-TA parallel sentences. Ramasamy et al. (2012) created SMT systems (phrase-based and hierarchical) that were trained on a much larger dataset of 190k parallel sentences. They also performed pre-processing steps involving morphological rules based on Tamil suffixes that improved upon the BLEU score of the baseline model (from 9.42 to 9.77). Their dataset (henceforth called UFAL) became the default benchmark for EN-TA translation systems until 2019, and we also use it in our experiments as an additional (general-domain) development set.

To the best of our knowledge, there have only been a handful of NMT systems trained on EN→TA. For the Indic languages multilingual tasks of WAT-2018, Sen et al. (2018), Dabre et al. (2018) and Ojha et al. (2018) reported BLEU scores for EN→TA. The Phrasal-based SMT system of Ojha et al. (2018) with a score of 30.53 BLEU outperformed the NMT systems of Sen et al. (2018) (11.88) and Dabre et al. (2018) (18.60), suggesting that the NMT systems were not suitable for translating a highly morphological language such as Tamil. However, the following year, Philip et al. (2019) outperformed Ramasamy et al. (2012) on the UFAL dataset with a BLEU score of 13.05. They report that techniques such as domain adaptation and back-translation can make training NMT systems on low-resource languages possible.

4 Datasets

For our constrained systems, we restrict ourselves to the datasets provided by WMT.

Parallel Table 1 presents the various parallel corpora along with their size and genre. The various corpora come from various sources and differ considerably in size. We also observe a very large difference in number of tokens between the two languages, with around 5 times more English tokens than Tamil tokens.

Monolingual Table 2 presents the monolingual Tamil corpora used in our experiments. Monolingual data is about 3 times larger than the parallel data in terms of tokens.

4.1 Pre-processing

For both parallel and monolingual data, the following steps are carried out sequentially:

- Sentences are tokenized and segmented by one of the segmentation algorithms described in the following section.
- Sentences longer than 150 tokens are removed.
- Sentences whose target to source ratio is below 0.7 are retained. This ratio is calculated based on the sentence lengths.
- Similar to Philip et al. (2019), a language match threshold is applied. Sentences rated 98% or higher are retained.
- Duplicate sentences are removed.

5 Methods

5.1 Segmentation

We compare two segmentation techniques: data-driven subwords and linguistically motivated subwords.

Subword segmentation refers to fully data-driven, non linguistically motivated segmentation algorithms (Sennrich et al., 2016c; Kudo and Richardson, 2018) that generate sub-words based on simpler frequency criteria to attain a predetermined vocabulary size. In our experiments we try out different vocabulary sizes as well as generating the subwords either individually for each language or jointly learning on both. The SentencePiece (**SP**) implementation (Kudo and Richardson, 2018) is used to perform this segmentation.

Linguistically Motivated Vocabulary Reduction (LMVR) is an unsupervised morphological segmentation algorithm based on Morfessor FlatCat (Kohonen et al., 2010; Grönroos et al., 2014) and proposed by Ataman et al. (2017). LMVR works by imposing an extra condition on the cost function of Morfessor so as to favour vocabularies of the desired size. When comparing regular Subword tokenization to LMVR, Ataman et al. (2017) report a +2.3 BLEU improvement on the English-Turkish translation task. Similar to SP, we need to set the vocabulary size prior to running the segmentation. LMVR models are trained separately for Tamil and English, which are then used to segment the respective datasets.

Name	Domain	EN Tokens(k)	TA Tokens(k)	Sentences(k)
Wikititles	Wikipedia	215	18	95
PMI	Political	707	87	40
UFAL	Mixed (News, Bible & Cinema)	3893	514	166
Koran	Religious	2366	586	92
MkB	Political (Speech)	104	15	6
PIB	Indian Press	1123	149	61
NLPC	Mixed	65	8	7
Wikimatrix	Mixed	2178	503	158
Total		10669	1885	625

Table 1: Approximate sizes (in thousands) of the Parallel Corpora used for training the NMT models

Name	Domain	TA Tokens(k)	Sentences(k)
Wikipedia Dumps	Wikipedia	4034	1669
News crawl	News	1496	709
PMI	Political	207	99
Total		5737	2477

Table 2: Approximate sizes (in thousands) of the Tamil Monolingual Corpora

5.2 Back-translation

In order to artificially increase the training data, we employ back-translation (BT) (Sennrich et al., 2016b). We consider two variations of this approach:

TaggedBT was presented by Caswell et al. (2019) and is similar to the original BT technique of Sennrich et al. (2016b), with the major difference being the addition of a special tag (here <BT>) in front of every back-translated English sentence. Caswell et al. (2019) had shown that this simple manoeuvre resulted in a higher BLEU score when compared to untagged BT based NMTs.

StupidBT Rather than performing actual BT which is expensive, Burlot and Yvon (2018) carry out the following:

1. Copy the target side data to the source side.

2. Prepend each token on the source side with a special id. For example, the token *tablet* becomes *bt_tablet*.

This simple and cost-effective technique was shown to perform almost on a par with regular BT on the English→French translation task.

5.3 Fine-tuning

Fine-tuning or transfer learning (Pan and Yang, 2010) is an effective technique to address a domain mismatch between the training set and the testset. While the testset consists of excerpts from newspapers, the training set consists of corpora with genres ranging from religious, political to movie subtitles. In fact, only a third of UFAL is news-oriented. A strategy to circumvent the domain mismatch is to fine-tune a pre-trained NMT system on a more domain specific dataset. Unfortunately the UFAL corpus is not domain tagged, so the news-only sentences cannot be easily retrieved.

We also excluded the PIB dataset due to its small size and large amount of almost identical sentences.

We hence perform fine tuning on the PMI dataset: This dataset consists of the sentences that were crawled from the Prime Minister of India's blog, with matters that are mostly political in nature. Despite the different content, we expect this corpus to be the closest in genre to the testset among the remaining parallel corpora.

5.4 Word Dropout

First introduced in Gal and Ghahramani (2016), the word dropout technique was modified by Sennrich et al. (2016a) to randomly drop tokens in-

stead of types during training. They reported an increase of 4-5 BLEU for the English↔Romanian language pair. Furthermore, Sennrich and Zhang (2019) report that introducing word dropout into NMT systems in low-resource settings leads to improvements in BLEU scores. We would hence like to investigate if the same improvements can be observed for EN-TA.

6 Experimental Setup

All our NMTs are developed using Fairseq (Ott et al., 2019). Following the architecture setup of Philip et al. (2019) the Transformer-Base implementation (BASE) is used, with slight changes to a few parameters, which are explained below. The encoder and decoder are both set to 5 layers with embedding dimension of 512 and 8 attention heads. The hidden layer dimension is 2048 and layer normalization is applied before each encoder and decoder layer. Other parameters were set as follows: dropout (0.001), weight decay (0.2) and batch size of 4k tokens. Our loss function is cross-entropy with label smoothing of 0.2. The model is trained for 100 epochs with early stopping criterion set to 3.

Segmentation The various segmentation algorithms are trained on the training data prior to the translation task. We report results with the following vocabulary sizes: 5k (source-target joint), 5k/5k, 10k/10k, 15k/15k and 20k/20k (source/target disjoint).

Back-Translation In order to perform BT, we first need to train a NMT model in the reverse direction, i.e. TA→EN. A Transformer based architecture is also used here. Our best configuration was: embedding and decoder having 6 layers, embedding layer having 512 dimensions and 6 attention heads with the rest of the parameters set as BASE. This model achieves a BLEU score of 18.27 on the UFAL TA-EN testset.

Fine-Tuning For the fine-tuning step, we take the pretrained BASE models and continue training them on the PMI dataset. An exhaustive search is done to find the best configurations for the fine tuning. The parameters with which we experimented are the learning rate, batch size, dropout and the value of label smoothing. Eventually we selected the following fine-tuning setup: learning rate of 0.002, batch size of 128, dropout of 0.3, label smoothing with factor of 0.3, and early stopping after 5 epochs without improvements.

Word Dropout Following Sennrich and Zhang (2019) we set the source word dropout to 0.3, i.e. the probability of a source word, in a batch, being dropped prior to training is 0.3.

7 Results

We report BLEU scores on three testsets: the UFAL testset (Ramasamy et al., 2012), half of the WMT2020 devset (DEV)[1] and the official WMT2020 testset. Given the rich morphology of Tamil, we also report CHRF scores (Popović, 2015) on the WMT2020 testset. We ran the program chrF++.py[2] with the arguments -nw 0 -b 3 to obtain the CHRF score.

From prior experimentation we found that a jointly trained SP model resulted in better BLEU when compared to separate training for each language, and hence perform the majority of SP experiments in Table 3 using a joint segmentation. On the other hand, LMVR being linguistically motivated is supposed to be trained independently for each language.

The last two contrastive experiments (Exp8.2 and Exp11.2) were run after the evaluation phase to better assess the impact of LMVR on translation quality in our best systems.

The following observations can be made based on the results:

Differences across testsets The trends are often inconsistent across testsets. Exp2 gave the highest BLEU score on UFAL (11.8) but a low BLEU score for DEV and WMT. On the other side, Exp11 (and Exp11.2) provided us the highest BLEU score on the official WMT testset, but a low 10.5 for UFAL. These variations could be attributed to the nature of the testsets and our training regime. Because we focused on improving our NMT systems to adapt to the news genre of WMT testset, this resulted in loss of translation accuracy of the UFAL testset, which was a mixture of three domains (one of them being news).

Effect of Back-translation Across both segmentation techniques, back-translation proved to be beneficial. Despite previously reported results, we found that fully fledged back-translation

[1] We randomly select one half of the WMT2020 devset for validation and use the remaining half for evaluation (DEV).

[2] https://github.com/m-popovic/chrF

	System	Segment.	Dict.size	BLEU UFAL	BLEU DEV	BLEU WMT	CHRF WMT
Exp1	BASE	SP	5k	11.2	8.5	5.1	42.8
Exp2	BASE +StupidBT	SP	5k	11.8	8.6	5.1	41.9
Exp3	BASE +TaggedBT	SP	5k	11.7	8.9	5.4	44.3
Exp6	BASE +TaggedBT	LMVR	5k/5k	11.1	9	5.6	40.1
Exp7	BASE +TaggedBT	LMVR	10k/10k	11.2	9.2	5.6	43.6
Exp8	BASE +TaggedBT	LMVR	15k/15k	11.1	9.3	6.0	48.1
Exp9	BASE +TaggedBT	LMVR	20k/20k	11.2	9.2	5.9	45.9
Exp11	BASE +TaggedBT+FT	LMVR	15k/15k	10.2	9.7	6.0	46.1
Exp13	BASE +TaggedBT+WD+FT	LMVR	15k/15k	10.7	**10.2**	**6.5**	**50.9**
Exp8.2	BASE +TaggedBT	SP	15k/15k	11.3	9.1	6.3	44.2
Exp11.2	BASE +TaggedBT+FT	SP	15k/15k	10.5	9.7	6.6	47.2

Table 3: English-Tamil results on three datasets: the general-domain UFAL (Ramasamy et al., 2012), our news development set (DEV) and the official WMT2020 news testset (WMT). Exp11 (in bold) was our official submission to WMT2020. SP refers to SentencePiece and LMVR to (Ataman et al., 2017). Dictionary size is given as one number for source-target joint segmentation, or as two numbers for source/target size when disjoint. FT and WD stand for fine-tuning and word dropout, respectively.

(TaggedBT) works considerably better than its cheaper approximation (StupidBT) on DEV, but not on the UFAL testset. While DEV reported increases of +0.3 (Exp2 vs. Exp3), a drop of -0.1 in BLEU was seen for UFAL. This could be due to the fact that Newscrawl was a major constituent of the monolingual corpora, that were used to train the TaggedBT systems. Also, when comparing a BASE system to one with TaggedBT (Exp1 vs. Exp3), we find an increase of +0.3 in BLEU. Given the DEV result, we decided to use fully fledged TaggedBT for the rest of our experiments.

SP vs. LMVR Based on our initial experiments, LMVR seemed to outperform SP. For instance, when comparing the TaggedBT systems with SP and LMVR (Exp3 vs. Exp9) we see a +0.5 increase in BLEU.

However, after the official submission, we performed additional contrastive experiments to account for LMVR having a much larger and disjoint vocabulary size (see Exp 8.2 vs. Exp8 and Exp11.2 vs. Exp11). In both settings, the linguistically motivated segmentation was actually outperformed by SentencePiece (+0.3 higher BLEU score on WMT). On the other hand, results were inconclusive when looking at the CHRF scores: namely, LMVR is much better than SP in the non fine-tuned system (Exp8 vs. Exp8.2), but slightly worse in the fine-tuned system (Exp11 vs. Exp11.2). These results seem to reveal a complex interplay between the effect of domain adaptation and the choice of an optimal segmentation strategy.

Effect of vocabulary size For our BT model with LMVR segmentation, we report the scores for four different vocabulary sizes (Exp6 to Exp9): among these, 15k for each language (Exp8) gives the best BLEU score of 9.3 on DEV. Therefore we use this size for the remaining experiments.

Effect of fine tuning When we compare models to their counterparts that were additionally fine-tuned, we observe a slight increase in the DEV BLEU score for the LMVR systems (compare Exp8 vs. Exp11) but unfortunately no effect on the WMT testset. This is probably due to the fact that the dataset on which we fine-tuned (PMI) was not close enough to the domain of the news translation testset.

Effect of word-dropout Word dropout was introduced to our best system, that is the one using TaggedBT and a LMVR vocabulary size of 15k/15k. The resulting system (Exp13) turned out to be our best performing system overall, but was not ready in time for the official submission. We find that the addition of word dropout resulted in a BLEU increase of +0.5 on DEV and WMT, and a large CHRF increase (+4.8) on WMT, which confirms the usefulness of this technique on a new lan-

Case	Case Marker	Tamil	SP	LMVR
Nominative	−∅	புத்தம் puththagam 'book'	புத்த+கம் puththa+gam	புத்த+கம் puththa+gam
Accusative	−அ −a	புத்தகம puththagama 'the book'	புத்தக+ம puththaga+ma	புத்த+க+ம puththa+ga+ma
Dative	−உக்கு −ukku	புத்தகமுக்கு puththagamukku 'to/for the book'	புத்தக+மு+க்கு puththaga+mu+kku	புத்த+கம+உக்கு puththa+gam+ukku
Genitive	−ஓட −ooda	புத்தகமோட puththagamooda 'the book's'	புத்தக+மோட puththaga+mooda	புத்த+க+மோ++ட puththa+ga+moo+da
Instrumental	−ஆல −aala	புத்தகமால puththagamaala 'by the book'	புத்தக+ம்+ஆல puththaga+m+aala	புத்த+க+மா++ல puththa+ga+maa+la
Sociative	−ஓட −ooda	புத்தகமோட puththagamooda 'along with the book'	புத்தக+மோட puththaga+mooda	புத்த+க+மோ++ட puththa+ga+moo+da
Locative	−ல −la	புத்தகம்ல puththagamla 'in the book'	புத்த+கம்+ல puththa+gam+la	புத்த+கம்+ல puththa+gam+la
Ablative	−லருந்து −larundhu	புத்தகம்லருந்து puththagamaala 'from the book'	புத்த+கம்+ல+ருந்து puththa+gam+la+rundhu	புத்த+கம்+ல+ருந்து puththa+gam+la+rundhu

Table 4: Different inflections of the Tamil singular noun புத்தகம். Columns SP and LMVR show the segmentations resulted by the SentencePiece (SP) and LMVR algorithms respectively.

guage pair.

8 Analysis

We also performed two small qualitative studies on the best systems based on segmentation. First, we compare how the segmentation algorithms segment the Tamil word புத்தகம் 'book'.

Secondly, using the example of the word *book* we observe how the systems translate the word to and from Tamil (Table 5).

Segmentation Table 4 shows how the word புத்தகம் and its various case forms are segmented by the segmentation techniques. The main differences that we observe are:

- LMVR and SP generated the same segmentation for three cases: nominative, locative and ablative.
- LMVR always generated segmentations with the base sub-word புத்த/puththa for all the case forms while SP generated the segments புத்த/puththa or புத்த/puththaga. This confirms the observations of Ataman et al. (2017), that LMVR produces more morphological segments.
- LMVR, on average, resulted in more segments per token than SP.

Translation Quality For the compound *comic-book*, the SP system translates it as நகைச்சுவை புக்/nakaiccuvai puk, i.e. *comedy book*, with the word புக்/puk being a direct transliteration for *book* and hence incorrect. On the other hand, LMVR provides the correct translation.

There were in total two occurrences of the nominative form of the புத்தகம்/puththagam, which were correctly translated by the two systems. The same was observed for the locative form புத்தகம்ல/puththagamla.

An example where both systems fail to translate the phrase *by the book* as in the sentence "I have every reason to believe they have done everything by the book and ...". The SP system provides

English	Tamil	SP	LMVR
comic-book	காமிக் புத்தக kaamik puththga	நகைச்சுவை புக் nakaiccuvai puk	காமிக் புத்தக kaamik puththga
book	புத்தகம் puththagam	புத்தகம் puththagam	புத்தகம் puththagam
in the book	புத்தகம்ல puththagamla	புத்தகம்ல puththagamla	புத்தகம்ல puththagamla
notebook	புத்தகத்தைத் puththagaththait	புத்தகட்டி puththagatti	குறிப்பேடு kurippetu
by the book	சட்டத்தைப் cattattaip	புத்தகமால் puththagamaal	விதிப்படி vithippati

Table 5: Qualitative Analysis of the Tamil word புத்தகம்/*puththagam* along with selected translations of the English word *book*

a grammatically correct form for book (ablative), it is however semantically incorrect. Meanwhile, the LMVR system generates the word விதிப்படி/vithippati meaning 'by rule' while the reference word has the meaning சட்டத்தைப்(bill).

Finally we observed with the word *notebook* that SP generated a non-existent word and LMVR provided an another translation for the English word *notebook*.

In the future, we aim to conduct in-depth analysis on what and which morphological features are captured by the NMT systems.

9 Conclusion

Although our results were not competitive with the other submissions for the EN-TA task, our paper presents the various settings that leads to an improvement in EN-TA translation. Mainly, we found that linguistically motivated subword segmentation (Ataman et al., 2017), which was previously shown to benefit translation from/into various non-Indian languages, does not consistently outperform the widely used SentencePiece segmentation despite the agglutinative nature of Tamil morphology. We also found that, for our English-Tamil systems, fully-fledged back-translation remains more competitive than its cheaper alternative (Burlot and Yvon, 2018). And finally, we observe a noticeable CHRF gain when adding word dropout (Sennrich et al., 2016a) to our best model.

Acknowledgements

Arianna Bisazza was funded by the Netherlands Organization for Scientific Research (NWO) under project number 639.021.646. We also would like to thank the Center for Information Technology of the University of Groningen for providing access to the Peregrine high performance computing cluster.

References

Duygu Ataman, Matteo Negri, Marco Turchi, and Marcello Federico. 2017. Linguistically motivated vocabulary reduction for neural machine translation from turkish to english. *The Prague Bulletin of Mathematical Linguistics*, 108(1):331 – 342.

Franck Burlot and François Yvon. 2018. Using monolingual data in neural machine translation: a systematic study. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 144–155, Brussels, Belgium. Association for Computational Linguistics.

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, Florence, Italy. Association for Computational Linguistics.

Raj Dabre, Anoop Kunchukuttan, Atsushi Fujita, and Eiichiro Sumita. 2018. NICT's participation in WAT 2018: Approaches using multilingualism and recurrently stacked layers. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation: 5th Workshop on Asian Translation: 5th Workshop on Asian Translation*, Hong Kong. Association for Computational Linguistics.

Yarin Gal and Zoubin Ghahramani. 2016. A theoretically grounded application of dropout in recurrent neural networks. In D. D. Lee, M. Sugiyama,

U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 1019–1027. Curran Associates, Inc.

Ulrich Germann. 2001. Building a statistical machine translation system from scratch: How much bang for the buck can we expect? In *Proceedings of the ACL 2001 Workshop on Data-Driven Methods in Machine Translation*.

Stig-Arne Grönroos, Sami Virpioja, Peter Smit, and Mikko Kurimo. 2014. Morfessor FlatCat: An HMM-based method for unsupervised and semi-supervised learning of morphology. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1177–1185, Dublin, Ireland.

Oskar Kohonen, Sami Virpioja, and Krista Lagus. 2010. Semi-supervised learning of concatenative morphology. In *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, pages 78–86, Uppsala, Sweden. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Atul Kr. Ojha, Koel Dutta Chowdhury, Chao-Hong Liu, and Karan Saxena. 2018. The RGNLP machine translation systems for WAT 2018. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation: 5th Workshop on Asian Translation: 5th Workshop on Asian Translation*, Hong Kong. Association for Computational Linguistics.

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*.

S. J. Pan and Q. Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

Jerin Philip, Shashank Siripragada, Upendra Kumar, Vinay Namboodiri, and C V Jawahar. 2019. Cvit's submissions to wat-2019. In *Proceedings of the 6th Workshop on Asian Translation*, pages 131–136, Hong Kong, China. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Loganathan Ramasamy, Ondřej Bojar, and Zdeněk Žabokrtský. 2012. Morphological processing for english-tamil statistical machine translation. In *Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages (MTPIL-2012)*, pages 113–122.

Kengatharaiyer Sarveswaran, Gihan Dias, and Miriam Butt. 2019. Using meta-morph rules to develop morphological analysers: A case study concerning Tamil. In *Proceedings of the 14th International Conference on Finite-State Methods and Natural Language Processing*, pages 76–86, Dresden, Germany. Association for Computational Linguistics.

Sukanta Sen, Kamal Kumar Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2018. IITP-MT at WAT2018: Transformer-based multilingual indic-English neural machine translation system. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation: 5th Workshop on Asian Translation: 5th Workshop on Asian Translation*, Hong Kong. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Edinburgh neural machine translation systems for WMT 16. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 371–376, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. 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. 2016c. 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.

Rico Sennrich and Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, Italy. Association for Computational Linguistics.