Investigating Low-resource Machine Translation for English-to-Tamil

Akshai Ramesh[‡],Venkatesh Balavadhani Parthasarathy[‡], Rejwanul Haque and Andy Way
The ADAPT Centre, [‡]School of Computing
Dublin City University, Dublin, Ireland
akshai.ramesh2,venkatesh.balavadhaniparthasa2@mail.dcu.ie
rejwanul.haque,andy.way@adaptcentre.ie

Abstract

Statistical machine translation (SMT) which was the dominant paradigm in machine translation (MT) research for nearly three decades has recently been superseded by the end-to-end deep learning approaches to MT. Although deep neural models produce state-of-the-art results in many translation tasks, they are found to underperform on resource-poor scenarios. Despite some success, none of the present-day benchmarks that have tried to overcome this problem can be regarded as a universal solution to the problem of translation of many low-resource languages. In this work, we investigate the performance of phrase-based SMT (PB-SMT) and neural MT (NMT) on a rarely-tested low-resource language-pair, English-to-Tamil, taking a specialised data domain (software localisation) into consideration. In particular, we produce rankings of our MT systems via a social media platform-based human evaluation scheme, and demonstrate our findings in the low-resource domain-specific text translation task.

1 Introduction

In recent years, MT researchers have proposed approaches to counter the data sparsity problem and to improve the performance of NMT systems in low-resource scenarios, e.g. augmenting training data from source and/or target monolingual corpora (Sennrich et al., 2016a; Chen et al., 2019), unsupervised learning strategies in the absence of labeled data (Artetxe et al., 2018; Lample et al., 2018), exploiting training data involving other languages (Firat et al., 2017; Johnson et al., 2017), multi-task learning (Niehues and Cho, 2017), selection of hyperparameters (Sennrich and Zhang, 2019), and pre-trained language model fine-tuning (Liu et al., 2020). Despite some success, none of the existing benchmarks can be viewed as an overall solution as far as MT for low-resource language-pairs is concerned. For examples, the back-translation strategy of Sennrich et al. (2016a) is less effective in low-resource settings where it is hard to train a good back-translation model (Currey et al., 2017); unsupervised MT does not work well for distant languages (Marie and Fujita, 2018) due to the difficulty of training unsupervised cross-lingual word embeddings for such languages (Søgaard et al., 2018) and the same is applicable in the case of transfer learning too (Montoya et al., 2019). Hence, this line of research needs more attention from the MT research community. In this context, we refer interested readers to some of the papers (Bentivogli et al., 2016; Castilho et al., 2017) that compared PB-SMT and NMT on a variety of use-cases. As for low-resource scenarios, as mentioned above, many studies (e.g. Koehn and Knowles (2017); Östling and Tiedemann (2017); Dowling et al. (2018)) found that PB-SMT can provide better translations than NMT, and many found the opposite results (Casas et al., 2019; Sen et al., 2019; Sennrich and Zhang, 2019). Hence, the findings of this line of MT research have indeed yielded a mixed bag of results, leaving the way ahead unclear.

In Ramesh et al. (2020), we investigated the performance of PB-SMT and NMT systems on two rarely-tested under-resourced language-pairs, English-to-Tamil and Hindi-to-Tamil, taking a specialised data domain (software localisation) into account. In particular, in Ramesh et al. (2020), we carried out a comprehensive manual error analysis on the translations produced by our PB-SMT and NMT systems. This current work extends the work of Ramesh et al. (2020) in the following ways: (a) we present a social media platform-based human evaluation scheme for measuring the quality of translations generated by different MT systems, and (b) we select the PB-SMT and NMT systems of the English-to-Tamil translation task from Ramesh et al. (2020) and a commercial MT system, compare their performances, and produce rankings of the three MT systems in terms of the length of the sentences to be

translated using our proposed social media platform-based human evaluation scheme.

The remainder of the paper is organized as follows. Section 2 explains the experimental setup including the descriptions on our MT systems and details of the data sets used. Section 3 presents the results with discussions and analysis, while Section 4 concludes our work with avenues for future work.

2 Experimental Setups

2.1 The MT systems

This section provides an overview of the PB-SMT and NMT systems used for experimentation.[1] To build our PB-SMT systems we used the Moses toolkit (Koehn et al., 2007). We used a 5-gram language model trained with modified Kneser-Ney smoothing (Kneser and Ney, 1995) using the KenLM toolkit (Heafield et al., 2013). Our PB-SMT log-linear features include: (a) 4 translational features (forward and backward phrase and lexical probabilities), (b) 8 lexicalised reordering probabilities (*wbe-mslr-bidirectional-fe-allff*), (c) 5-gram LM probabilities, (d) 5 OSM features (Durrani et al., 2011), and (e) word-count and distortion penalties. The weights of the parameters are optimized using the margin-infused relaxed algorithm (Cherry and Foster, 2012) on the development set. For decoding, the cube-pruning algorithm (Huang and Chiang, 2007) is applied, with a distortion limit of 12.

To build our NMT systems, we used the OpenNMT toolkit (Klein et al., 2017). The NMT systems are Transformer models (Vaswani et al., 2017). The tokens of the training, evaluation and validation sets are segmented into sub-word units using Byte-Pair Encoding (BPE) (Sennrich et al., 2016b). Recently, Sennrich and Zhang (2019) demonstrated that commonly used hyper-parameter configurations do not provide the best results in low-resource settings. Accordingly, we carried out a series of experiments in order to find the best hyperparameter configuration for Transformer in our low-resource settings. In particular, we found that the following configuration lead to the best results in our low-resource translation settings: (i) the BPE vocabulary size: 8,000, (ii) the sizes of encoder and decoder layers: 4 and 6, respectively, (iii) learning-rate: 0.0005, (iv) batch size (token): 4,000, and (v) Transformer head size: 4. As for the remaining hyperparameters, we followed the recommended best set-up from Vaswani et al. (2017). The validation on the development set is performed using three cost functions: cross-entropy, perplexity and BLEU (Papineni et al., 2002). The early stopping criteria is based on cross-entropy; however, the final NMT system is selected as per highest BLEU score on the validation set. The beam size for search is set to 12.

[1] Note that we used the MT systems built by Ramesh et al. (2020) for our experiments.

2.2 Choice of Languages

In an attempt to test MT on low-resource scenarios, we chose English and an Indian language: Tamil. English and Tamil are Germanic and Dravidian languages, respectively, so the languages we selected for investigation are from different language families and morphologically divergent to each other. English is a less inflected language, whereas Tamil is a morphologically rich and highly inflected language. Our investigation is from a less inflected language to a highly inflected language. With this, we compare translation in PB-SMT and NMT with a translation-pair involving two morphologically divergent languages.

2.3 Data Used

This section presents the datasets used for MT system building (Ramesh et al., 2020). For experimentation we used data from three different sources: OPUS[2] (Tiedemann, 2012), WikiMatrix[3] (Schwenk et al., 2019) and PMIndia[4] (Haddow and Kirefu, 2020). Corpus statistics are shown in Table 1. We carried out experiments using two different setups: (i) in the first setup, the MT systems were built on a training set compiled from all data domains listed above; we call this setup MIXED, and (ii) in the second setup, the MT systems were built on a training set compiled only from different software localisation data from OPUS, *viz.* GNOME, KDE4 and Ubuntu; we call this setup IT. The development and test set sentences were randomly drawn from these localisation corpora.

We adopted a number of standard cleaning routines for removing noisy sentences from the training corpora (Ramesh et al., 2020). In order to perform tokenisation for English, we used the standard tool in the Moses toolkit. For tokenising and normalising Tamil sentences, we used the Indic NLP library.[5]

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

[3] https://ai.facebook.com/blog/wikimatrix/

[4] http://data.statmt.org/pmindia

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

Table 1: Data Statistics

		sents.	words [En]	words [Ta]
train sets	MIXED	222,367	5,355,103	4,066,449
	vocab		424,701	423,599
	avg. sent		25	19
	IT	68,352	448,966	407,832
	vocab		31,216	77,323
	avg. sent		7	6
devset		1,500	17,903	13,879
testset		1,500	16,020	12,925

3 Results and Discussion

3.1 Automatic Evaluation

We present the comparative performance of the PB-SMT and NMT systems in terms of the widely used automatic evaluation metric BLEU (Papineni et al., 2002). Additionally, we performed statistical significance tests using bootstrap resampling methods (Koehn, 2004). Sections 3.1.1 and 3.1.2 present the performance of the MT systems on the MIXED and IT setups, respectively.

3.1.1 The MIXED Setup

We show the BLEU scores on the test set in Table 2. The PB-SMT and NMT systems produce relatively low BLEU scores on the test set given the difficulty of the translation pairs. However, these BLEU scores underestimate the translation quality, given the relatively free word order in Tamil, and the fact that we have just a single reference translation set for evaluation. We see from Table 2 that PB-SMT surpassed NMT by a large margin in terms of BLEU, and found that the difference in the BLEU scores of the MT systems is statistically significant.

Table 2: The Mixed Setup.

PB-SMT	9.56
NMT	4.35

3.1.2 The IT Setup

This section presents the results obtained on the IT setup. The BLEU scores of the MT systems are reported in Table 3. When we compare the BLEU scores of this table with those of Table 2, we see a huge rise in terms of the BLEU scores for PB-SMT and NMT, and the improvements are found to be statistically significant.

As far as the IT setup is concerned, the PB-SMT system outperforms the NMT system statistically significantly, and we see an improvement of an absolute of 6.33 points (corresponding to 69.3% relative) in terms of BLEU on the test set. As discussed in Section 2.3, in the IT task, the MT systems were built exclusively on in-domain training data, and in the MIXED setup, the training data is composed of a variety of domains, i.e. religious, IT, political news. In a nutshell, when we compare PB-SMT and NMT, we see that PB-SMT is always the leading system across the training data setups (MIXED and IT).

Table 3: The IT Setup.

PB-SMT	15.47
NMT	9.14

3.2 Reasons for very low BLEU Scores

The BLEU scores reported in the sections above are very low. We looked at the translations of the test set sentences by the MT systems and compared them with the reference translations. We found that despite being good in quality, in many cases the translations were penalised heavily by the BLEU metric as a result of many *n*-gram mismatches with the corresponding reference translations. This happened mainly due to the nature of target language (Tamil) in question, i.e. Tamil is a free word order language. This is indeed responsible for the increase in non-overlapping *n*-gram counts. We also found that translations contain lexical variations of Tamil words of the reference translation, again resulting in the increase of the non-overlapping *n*-gram counts. We show some of such translations in Table 4.

(1)	src:	information
	hyp:	தகவல்
	ref:	அறிமுகம்
(2)	src:	file
	hyp:	கோப்பு
	ref:	பைல்
(3)	src:	authentication is required to change your own user data
	hyp:	பயனர் தரவை மாற்ற அனுமதி தேவை
	ref:	உங்களுடைய சொந்த பயனர் தரவை மாற்ற அனுமதி தேவை

Table 4: Translations that are good in quality were unfairly penalised by the BLEU metric.

3.3 The MT System Ranking

3.3.1 Evaluation Plan

We further assess the quality of our MT systems (the English-to-Tamil PB-SMT and NMT systems) via a

manual evaluation scheme. For this, we select our PB-SMT and NMT systems from the MIXED and IT setups. Additionally, we considered Google Translate (GT)[6] in this ranking task in order to compare it with PB-SMT and NMT. We randomly sampled a set of 100 source sentences from the test set (cf. Table 1), and their translations by the MT systems including GT. In order to conduct this evaluation, we developed a webpage that was made available online and accessible to the evaluators who ranked the MT systems according to their translation quality.

We placed the sentences of the test set into three sets based on the sentence length measure (source-side), i.e. number of words (nw)<=3, 3<nw<=9, and nw>9. We call these sets *sentence-length sets*. We recall Table 1 where the average sentence length of the English IT corpus is 7. This is the justification for our choice of sentence length range. We sampled 100 sentences from the test set in such a way that the sentences are equally distributed over the sentence-length sets. Thus, the first, second and third sentence-length sets contain 34, 33 and 33 sentences, respectively. The webpage displays 10 sentences together with the translations by the MT systems, which are taken from the sentence-length sets, with a minimum of 3 sentences from each set. The evaluators who are native speakers of Tamil with good knowledge of English were instructed to rank the MT systems as per the quality of the translations from best to worst. It was also possible that the evaluators could provide the same rank to more than one translation.

We disseminated the MT system ranking task via a variety of popular social media platforms, e.g. LinkedIn[7] and Facebook.[8] If we ask the evaluators to rank a large number of sentences, it is quite likely that they would not participate in the task. Even if some people might like to participate in the task, they may lose interest in the middle and quit. Therefore, we displayed translations in batches (i.e. 10 source sentences and their translations) on our webpage at any one time. We did not consider any partial submissions. We observed that a total of 38 and 60 evaluators participated in the task for the MIXED and IT setups, respectively. The submissions were then analysed to produce the final rankings of the MT systems. In order to measure agreement in judgement, we used Fleiss's Kappa.[9] The next section presents the ranking results.

3.3.2 Ranking Results

We adopted the idea of bilingual group pairwise judgements as in Papineni et al. (2002) in order to rank the MT systems. We take the pairwise scores of three MT systems and linearly normalise them across the three systems. We show our ranking results for the MIXED setup in the left half of Table 5. We see from the table that NMT is found to be the winner for first sentence-length set (nw<=3) followed by GT and PB-SMT. As for the other sentence-length-based sets, GT becomes the winner followed by PB-SMT and NMT. The same trend is observed when the systems are ranked ignoring the sentence-length measure. We recall Table 2 where we presented the BLEU scores of our English-to-Tamil MT systems (PB-SMT: 9.56 BLEU points and NMT: 4.35 BLEU points). Additionally, we evaluated GT on our test set in order to compare it with PB-SMT and NMT in this setting, and found that the GT MT system produced a 4.37 BLEU points on the test set. We see that PB-SMT is to the best choice and GT and NMT both are comparable if the MT systems are ranked according to the automatic evaluation scores. Therefore, the automatic evaluation results contradict the human ranking results above.

Using the submissions from the ranking task we also obtain the distributions of the translations by the PB-SMT, NMT and GT MT systems over the three ranking positions, which are shown in the upper graph of Figure 1. We see here that the majority of the translations that the evaluators tagged as 'best' (cf. 'first' in the upper graph of Figure 1) were from GT followed by NMT and PB-SMT. In case of the 'worst' position (cf. 'third' in the upper graph of Figure 1), we see that the majority of the translations are from the NMT systems followed by the PB-SMT and GT MT systems. When we look at the second position, we see that PB-SMT is the winner and NMT and GT are nearly neck-and-neck.

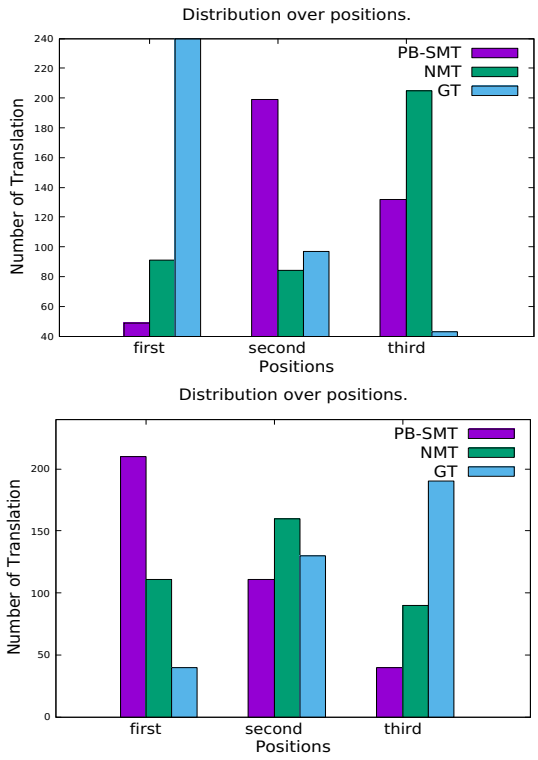
Table 5: Ranks of the MT Systems.

	Mixed setup			IT setup		
	NMT	PB	GT	NMT	PB	GT
s1 (nw<=3)	1st	3rd	2nd	1st	2nd	3rd
s2 (3<nw<=9)	3rd	2nd	1st	2nd	1st	3rd
s3 (nw>9)	3rd	2nd	1st	2nd	1st	3rd
test set	3rd	2nd	1st	2nd	1st	3rd

The ranking results for the IT setup are presented

[6] https://translate.google.com/
[7] https://www.linkedin.com/
[8] https://www.facebook.com/
[9] https://en.wikipedia.org/wiki/Fleiss%27_kappa

Figure 1: Distributions of translations over three positions (Mixed (top) and IT (bottom) setups).



in the right half of Table 5. This time, we see that NMT is the winner for first sentence-length set (nw<=3) followed by PB-SMT and GT. As for the other sentence-length-based sets and whole test set (100 sentences), PB-SMT becomes the winner followed by NMT and GT. The distributions of the translations by the MT systems over the three ranking positions are shown in the lower graph of Figure 1. We see that the majority of the translations that are tagged as 'best' were from PB-SMT followed by NMT and GT. In case of the 'worst' position, we see that the majority of the translations are from the GT system followed by the NMT and PB-SMT systems. When we look at the second position, we see that NMT is the winner and PB-SMT is not far behind, and the same is true for PB-SMT and GT too.

As for the first set of sentences (i.e. short sentences (nw<=3)), we observed that the translations by the NMT systems are found to be more meaningful compared to those by the other MT systems. This is true for both the MIXED and IT setups. As an example, the English sentence 'Nothing' is translated as எதுவும் இல்லை ('nothing') in Tamil by the NMT system, which, however, is translated as எதுவும் ('anything') in Tamil by the PB-SMT system.

On completion of our ranking process, we computed the inter-annotator agreements using Fleiss's Kappa for the three ranking positions first, second and third, which are 74.1, 58.4 and 67.3, respectively, for the MIXED setup and 75.3, 55.4 and 70.1, respectively, for the IT setup. A Kappa coefficient between 0.6-0.8 represents substantial agreement. In this sense, there is substantial agreement among the evaluators when they select positions for the MT systems.

4 Conclusion

In this paper, we investigated NMT and PB-SMT in resource-poor conditions. For this, we chose a specialised data domain (software localisation) for translation and a rarely-tested morphologically divergent low-resource language-pair, English-to-Tamil. We studied translations in two setups, i.e. training data compiled from (i) freely available variety of data domains (e.g. political news, Wikipedia), and (ii) exclusively software localisation data domains. In addition to an automatic evaluation, we randomly selected one hundred sentences from the test set, and ranked our MT systems via a social media platform-based human evaluation scheme. We also considered a commercial MT system, Google Translate, in this ranking task.

We found that use of in-domain data only at training has a positive impact on translation from English-to-Tamil. We looked at the translations produced by our MT systems and found that in many cases, the BLEU scores underestimate the translation quality mainly due to relatively free word order in Tamil. In this regard, both Shterionov et al. (2018) and Way (2018) note that BLEU may be under-reporting the difference in quality seen when using NMT systems, with the former attempting to measure the level of under-reporting using a set of novel metrics. Way (2018) reminds the MT community how important subjective evaluation is in MT and there is no easy replacement of that in MT evaluation. We refer the interested readers to Way (2019) who also drew attention to this phenomenon.

From our human ranking task we found that sentence-length could be a crucial factor for the performance of the NMT systems in low-resource scenarios, i.e. NMT turns out to be best-performing for very short sentences (number of words <= 3). This finding indeed does not correlate with the findings of our automatic evaluation process, where PB-SMT is found to be the best-performing, and GT and

NMT are comparable. This finding could be interest to translation service providers who use MT in their production for low-resource languages and may exploit the MT models based on the length of the source sentences to be translated.

GT becomes the winner followed by PB-SMT and NMT for the sentences of other lengths (number of words > 3) in the MIXED setup, and PB-SMT becomes the winner followed by NMT and GT for the sentences of other lengths (number of words > 3) in the IT setup. Overall, the human evaluators ranked GT as the first choice, PB-SMT as the second choice and NMT as the third choice MT systems in the MIXED setup. As for the IT setup, PB-SMT was the first choice, NMT was the second choice and GT was the third choice MT systems.

We believe that the findings of this work provide significant contributions to this line of MT research. In future, we intend to consider more languages from different language families. We also plan to include string-based MT evaluation metrics such as chrF (Popović, 2015) in our investigation, which have been shown to better reflect the actual performance improvement of NMT.

Acknowledgments

The ADAPT Centre for Digital Content Technology is funded under the Science Foundation Ireland (SFI) Research Centres Programme (Grant No. 13/RC/2106) and is co-funded under the European Regional Development Fund. This project has partially received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 713567, and the publication has emanated from research supported in part by a research grant from SFI under Grant Number 13/RC/2077.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3632–3642, Brussels, Belgium.

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus phrase-based machine translation quality: a case study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 257–267, Austin, Texas.

Noe Casas, José AR Fonollosa, Carlos Escolano, Christine Basta, and Marta R Costa-jussà. 2019. The TALP-UPC machine translation systems for WMT19 news translation task: pivoting techniques for low resource MT. In *Proceedings of the Fourth Conference on Machine Translation*, pages 155–162, Florence, Italy.

Sheila Castilho, Joss Moorkens, Federico Gaspari, Rico Sennrich, Vilelmini Sosoni, Panayota Georgakopoulou, Pintu Lohar, Andy Way, Antonio Valerio, Miceli Barone, and Maria Gialama. 2017. A Comparative Quality Evaluation of PBSMT and NMT using Professional Translators. In *Proceedings of MT Summit XVI, the 16th Machine Translation Summit*, pages 116–131, Nagoya, Japan.

Peng-Jen Chen, Jiajun Shen, Matthew Le, Vishrav Chaudhary, Ahmed El-Kishky, Guillaume Wenzek, Myle Ott, and Marc'Aurelio Ranzato. 2019. Facebook AI's WAT19 Myanmar-English translation task submission. In *Proceedings of the 6th Workshop on Asian Translation*, pages 112–122, Hong Kong, China.

Colin Cherry and George Foster. 2012. Batch tuning strategies for statistical machine translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 427–436, Montréal, Canada.

Anna Currey, Antonio Valerio Miceli Barone, and Kenneth Heafield. 2017. Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 148–156, Copenhagen, Denmark. Association for Computational Linguistics.

Meghan Dowling, Teresa Lynn, Alberto Poncelas, and Andy Way. 2018. SMT versus NMT: Preliminary comparisons for Irish. In *Proceedings of the AMTA 2018 Workshop on Technologies for MT of Low Resource Languages (LoResMT 2018)*, pages 12–20, Boston, MA.

Nadir Durrani, Helmut Schmid, and Alexander Fraser. 2011. A joint sequence translation model with integrated reordering. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1045–1054, Portland, Oregon, USA.

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

Barry Haddow and Faheem Kirefu. 2020. PMIndia–a collection of parallel corpora of languages of India. *arXiv preprint 2001.09907*.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified kneser-ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 690–696, Sofia, Bulgaria.

Liang Huang and David Chiang. 2007. Forest rescoring: Faster decoding with integrated language models. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 144–151, Prague, Czech Republic.

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

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

R. Kneser and H. Ney. 1995. Improved backing-off for m-gram language modeling. In *1995 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 181–184 vol.1.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 388–395, Barcelona, Spain.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, Williams College, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL 2007, Proceedings of the Interactive Poster and Demonstration Sessions*, pages 177–180, Prague, Czech Republic.

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.

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

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *arXiv preprint arXiv:2001.08210*.

Benjamin Marie and Atsushi Fujita. 2018. Unsupervised neural machine translation initialized by unsupervised statistical machine translation. *arXiv preprint arXiv:1810.12703*.

Héctor Erasmo Gómez Montoya, Kervy Dante Rivas Rojas, and Arturo Oncevay. 2019. A continuous improvement framework of machine translation for Shipibo-konibo. In *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*, pages 17–23, Dublin, Ireland. European Association for Machine Translation.

Jan Niehues and Eunah Cho. 2017. Exploiting linguistic resources for neural machine translation using multi-task learning. In *Proceedings of the Second Conference on Machine Translation*, pages 80–89, Copenhagen, Denmark.

Robert Östling and Jörg Tiedemann. 2017. Neural machine translation for low-resource languages. *CoRR*, abs/1708.05729.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL-2002: 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, PA. ACL.

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.

Akshai Ramesh, Venkatesh Balavadhani Parthasarathy, Rejwanul Haque, and Andy Way. 2020. An error-based investigation of statistical and neural machine translation performance on Hindi-to-Tamil and English-to-Tamil. In *Proceedings of the 7th Workshop on Asian Translation (WAT2020)*, Suzhou, China.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *arXiv preprint 1907.05791*.

Sukanta Sen, Kamal Kumar Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2019. IITP-MT system for Gujarati-English news translation task at WMT 2019. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 407–411, Florence, Italy. Association for Computational Linguistics.

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, Berlin, Germany.

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.

Dimitar Shterionov, Riccardo Superbo, Pat Nagle, Laura Casanellas, Tony O'dowd, and Andy Way. 2018. Human versus automatic quality evaluation of nmt and pbsmt. *Machine Translation*, 32(3):217–235.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788, Melbourne, Australia. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)*, pages 2214–2218, Istanbul, Turkey.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Andy Way. 2018. Quality expectations of machine translation. In S. Castilho, J. Moorkens, F. Gaspari, and S. Doherty, editors, *Translation quality assessment*, pages 159–178. Springer.

Andy Way. 2019. Machine translation: where are we at today? In Erik Angelone, Maureen Ehrensberger-Dow, and Gary Massey, editors, *The Bloomsbury Companion to Language Industry Studies*. Bloomsbury Academic Publishing.