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

Akshai Ramesh
Dublin City University
akshai.ramesh2@mail.dcu.ie

Venkatesh B P
Dublin City University
venkatesh.balavadhani2@mail.dcu.ie

Abstract

Statistical machine translation (SMT) was the state-of-the-art in machine translation (MT) research for more than two decades, but has since been superseded by neural MT (NMT). Despite producing state-of-the-art results in many translation tasks, neural models 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 performance of phrase-based SMT (PB-SMT) and NMT on two rarely-tested low-resource language-pairs, English-to-Tamil and Hindi-to-Tamil, taking a specialised data domain (software localisation) into consideration. This paper demonstrates our findings including the identification of several issues of the current neural approaches to low-resource domain-specific text translation.

The **link** to this work can be found here [1]

1 Introduction

As of today, NMT [2, 3] represents the state-of-the-art in MT research given their ability to produce better translations than the previous mainstream PB-SMT approaches [4]. The NMT approaches are often labeled as *data-hungry* as their learning task heavily relies on large amounts of training data, on the order of a few million sentence pairs, in order to produce reasonable quality translations. However, most of the world's language-pairs are low-resource or extremely low-resource. This situation becomes even worse if one takes a specialised domain (e.g. software, health, agriculture) into consideration for translation. As for the specialised domains, not only parallel corpora are scarce but monolingual corpora too. In this perspective, many high-resource languages can also be labelled as low-resource or extremely low-resource. Hence, the data-demanding nature of NMT is really a concerning matter for those language-pairs that do not have enough parallel training data. In this context, many recent studies have demonstrated that NMT performs poorly for many under-resource language-pairs, so much so that in many cases, NMT is found to be inferior to PB-SMT [5, 6, 7].

2 Related Works

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 [8, 9], unsupervised learning strategies in the absence of labeled data [10], exploiting training data involving other languages [11], multi-task learning [12], selection of hyperparameters [13], and pre-trained language model fine-tuning [14]. 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 [8] is less effective in low-resource settings where it is hard to train a good back-translation model [15]; unsupervised MT does not work well for distant languages [16], and the same is applicable in the case of transfer

learning too [17]. Hence, this line of research needs more attention from the MT research community. In particular, investigating low-resource MT taking a variety of criteria (e.g. multifaceted error analysis, testing untested language-pairs) into consideration may unravel lacunas of the state-of-the-art neural models. In this context, we refer the interested readers some of the papers [18, 19] that compared PB-SMT and NMT on a variety of use-cases. As for low-resource scenarios, as mentioned above, many studies [5, 6, 7] found that PB-SMT can provide better translations than NMT. In contrast to these findings, however, many studies have demonstrated that NMT is better than PB-SMT in low-resource situations [13, 20, 21]. Hence, the findings of this line of MT research have yielded indeed a mixed bag of results, where way ahead unclear.

In this work, we thoroughly investigate the performance of PB-SMT and NMT models on two rarely-tested under-resourced language-pairs, English-to-Tamil and Hindi-to-Tamil, taking a specialised data domain (software localisation) into account.

3 CRISP-DM Methodology

3.1 Business Understanding

We put forward the following questions that needs to be answered:
1. Investigate the performance of PB-SMT and NMT system with low resource scenario.
2. Evaluate the translation efficiency of MT systems on custom dataset containing only IT terms.
3. To examine the significance of byte pair encoding technique with morphologically rich and complex languages.
4. To find out the different types of errors in MT systems and classify them into specific categories.

3.2 Data Understanding

This section presents our datasets. For experimentation we used data from three different sources: OPUS1 (Tiedemann, 2012), WikiMatrix2 (Schwenk et al., 2019) and PMIndia3 (Haddow and Kirefu, 2020). Corpus statistics for English-to-Tamil 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 translation task IT-1, and (ii) in the second setup, the MT systems were built on a training set compiled from different software localisation data from OPUS, viz. GNOME, KDE4 and Ubuntu; we call this translation task IT-2. The development and test set sentences were randomly drawn from these localisation corpora.

Table 1: Data Statistics: Hindi-to-Tamil

		sents.	words [Hi]	words [Ta]
train sets	IT-1	1,00,047	1,705,034	1,196,008
	vocab		104,564	284,921
	avg. sent		17	14
	IT-2	48,461	3,54,426	2,76,514
	vocab		31,258	67,069
	avg. sent		8	7
devset		1,500	10,903	7,879
testset		1,500	9,362	6,748

Corpus statistics for Hindi-to-Tamil are shown in Table 2. We followed the training setups of the English-to-Tamil task for this task.

Table 2: Data Statistics: English-to-Tamil

		sents.	words [En]	words [Ta]
train sets	IT-1	222,367	5,355,103	4,066,449
	vocab		424,701	423,599
	avg. sent		25	19
	IT-2	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

3.3 Data Preparation

In order to remove noise from the data sets, we adopted the following measures. We observed that the corpora of one language (say, Hindi) contains sentences of other languages (e.g. English), so we use a language identifier[1] in order to remove such noise. Then, we adopted a number of standard cleaning routines for removing noisy sentences, e.g. removing sentence-pairs that are too short, too long or which violate certain sentence-length ratios. In order to perform tokenisation for English, we used the standard tool [2] in the Moses toolkit. For tokenising Hindi and Tamil sentences, we used the Indic NLP library.[3] The English corpora is lowercased. Note that Hindi and Tamil are unicase languages. The Hindi and Tamil texts were also normalised using the

[1] cld2: https://github.com/CLD2Owners/cld2
[2] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/tokenizer.perl
[3] https://github.com/anoopkunchukuttan/indic_nlp_library

Indic NLP library. BPE is regarded as the benchmark strategy for reducing data sparsity for NMT. Therefore, we built our NMT engines on both word and subword-level training corpora to test BPE's effectiveness on low resource translation tasks.

3.4 Modeling

To build our PB-SMT systems we used the Moses toolkit. We used a 5-gram language model trained with modified Kneser-Ney smoothing. 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, and (d) word-count and distortion penalties. The weights of the parameters are optimized using the margin-infused relaxed algorithm [22] on the development set. For decoding, the cube-pruning algorithm [23] is applied, with a distortion limit of 12.

To build our NMT systems, we used the OpenNMT toolkit. In our experiments we followed the recommended best set-up by building a transformer model from [3]. The tokens of the training, evaluation and validation sets are segmented into sub-word units using Byte-Pair Encoding (BPE) [24]. Since English, Hindi and Tamil are written in Roman, Devanagari and Brahmi scripts, respectively, and have no overlapping characters, BPE is applied individually on the source and target languages. We set the BPE vocabulary size to 2,000 as in [13]. We set the learning-rate to 0.0005 and batch size (token) to 4,000. The latter three setups were found to be helpful in low-resource MT [13].

3.5 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).

4 Results and Discussion

4.1 Automatic Evaluation

We present the comparative performance of the PB-SMT and NMT systems in terms of the widely used automatic evaluation metric BLEU [25]. Sections 4.1.1 and 4.1.2 present the performance of the MT systems on the IT-1 and IT2 tasks, respectively.

4.1.1 The IT-1 MT Task

We show the BLEU scores on the test set in Table 3. The first and second rows of the table represent the English-to-Tamil and Hindi-to-Tamil translation tasks, respectively. 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 that we have just a single reference translation set for evaluation.

Additionally, we evaluate Google Translate (GT)[4] on our test sets in order to compare it with PB-SMT and NMT in this setting. The BLEU scores of the GT MT systems are shown in the last column of Table 3. We see from Table 3 that PB-SMT surpassed NMT by a large margin in terms of BLEU in both the English-to-Tamil and Hindi-to-Tamil translation tasks. We found that the differences in the BLEU scores are statistically significant. The same is also true in the case of PB-SMT and GT in the English-to-Tamil translation. However, we see that the BLEU scores of the NMT systems and the GT MT system are comparable in this translation task. We also found that in the Hindi-to-Tamil translation task, both the PB-SMT and GT MT systems significantly outperform the NMT systems. However, PB-SMT and GT are comparable as far as the BLEU scores are concerned.

Table 3: The IT-1 task.

	PB-SMT	NMT		GT
		(w/o) BPE	(w) BPE	
En-to-Ta	9.56	4.01	4.35	4.37
Hi-to-Ta	5.48	2.10	1.23	5.66

Interestingly, when we compare the numbers of the third (w/o BPE: word-level) and fourth (with BPE: sub-word level) columns, we see that word-level NMT is better than sub-word level NMT for Hindi-to-Tamil translation. We also found that the NMT system built on the word-level training data significantly outperformed the one built on sub-word level training data. For this reason, we chose this setup for the IT-2 task in the case of Hindi-to-Tamil translation.

4.1.2 The IT-2 MT Task

This section presents the results of the IT-2 translation task. The BLEU scores of the MT systems are reported in Table 4. When

[4] `https://translate.google.com/`

we compare the BLEU scores of this table with those of Table 4, we see a huge rise in terms of the BLEU scores for PB-SMT and NMT as far as English-to-Tamil translation is concerned, and the improvements are found to be statistically significant. As for the Hindi-to-Tamil translation, we see a substantial deterioration in BLEU (an absolute difference of 1.36 points, a 24.9% relative loss in terms of BLEU) for PB-SMT. We found that this loss is statistically significant too. We also see that in this task the BLEU score of the NMT system is nearly identical to the one in the IT-1 task (2.12 BLEU points versus 2.10 BLEU points).

Table 4: The IT-2 task.

English-to-Tamil		Hindi-to-Tamil	
PB-SMT	NMT	PB-SMT	NMT
35.81	32.8	4.12	2.12

As far as the English-to-Tamil translation and the IT-2 task are concerned, the PB-SMT system outperforms the NMT system statistically significantly, and we see an improvement of an absolute of 3.01 (corresponding to 9.17% relative) in terms of BLEU on the test set. The same trend is seen in the Hindi-to-Tamil translation task too.

We have a number of observations from the results of the IT-1 and IT-2 MT tasks. The benchmark BPE technique of [24] does not seem to work well for translation between the morphologically rich and complex languages, i.e. Hindi-to-Tamil. As discussed in table 1, in the IT-2 task, the MT systems were built exclusively on in-domain training data, and in the IT-1 task, the training data is composed of a variety of domains, i.e. religious, IT, political news. We found that BPE has a positive impact on the performance of the MT systems (PB-SMT and NMT) for English-to-Tamil translation. Use of in-domain data only in training does not have any positive impact on the Hindi-to-Tamil translation, and we even saw a significant deterioration in performance on BLEU for PB-SMT. We conjecture that the morphological complexity of the languages (Hindi and Tamil) involved in this translation could be one of the reasons why the NMT and PB-SMT systems performed so poorly when trained exclusively on specialised domain data. When we compare PB-SMT and NMT, we see that PB-SMT is always the leading system in both the following cases: (i) across the MT tasks (IT-1 and IT-2) and (ii) the translation-directions (English-to-Tamil and Hindi-to-Tamil). When we added GT for comparison, we saw that it stays in between both paradigms (i.e. the GT MT systems are worse than our PB-SMT systems and better than our NMT systems in terms of the BLEU score) in the IT-1 task, where the training data is a mixture of domains. However, if we look at the IT-2 task, we see that both our PB-SMT and NMT systems surpass GT by huge margins except the Hindi-to-Tamil task.

4.2 Error Analysis

We conducted a thorough error analysis on the English-to-Tamil and Hindi-to-Tamil NMT systems only built for the IT-1 task. For this, we randomly sampled 100 sentences from the respective test sets (English-to-Tamil and Hindi-to-Tamil). This analysis aimed to trace translational errors (e.g. translation of out-of-vocabulary (OOV) terms), and the outcome of this analysis is presented in the following sections.

4.2.1 Terminology Translation

Since this work focuses on studying translation of a specialised domain data, we looked at this area of translation with a special focus. We first looked at the translations of the OOV terms in order to see how they are translated into the target. We found that both the NMT systems (English-to-Tamil and Hindi-to-Tamil) either incorrectly translate the software terms or drop them during translation. This happened for almost all OOV terms. Nonetheless, the NMT systems are able to correctly translate a handful of OOV terms; this phenomenon is also corroborated by [26] while investigating translation of the judicial domain terms.

Table 5: Term omission.

English	Support for most ipod / iphone / ipad devices
Tamil	பெரும்பாலும் . / சாதனங்களும் ஆதரவு [*perumpālum. / cāṭaṉaṅkaḷum ātarava*]
English	Open script
Tamil	திற [*tiṟa*]
English	Color set
Tamil	வண்ணத்தை அமைத்திடு [*vaṇṇattai amaittiṭu*]
Hindi	फ्रीसेल [*freecell*]
Tamil	இலவசகளம் [*ilavacakaḷam*]

We show four examples in Table 5. In the first example, we show a source English sentence and its Tamil translation. We see from the translation that the NMT system drops the source-side terms ‘ipod’, ‘iphone’ and ‘ipad’

in the target translation. In second example, we see that a part ('Open') of a multiword term ('Open script') is correctly translated into Tamil, and the MT system omits its remaining part ('script') in translation. In the third example, we show another multiword English term ('Color set') and its Tamil translation which is wrong (i.e. English equivalent 'set the color'). Here, we see the MT system made correct lexical choices for each word of the source term, although the meaning of the target-equivalent of the respective translation is different to that of the source term. This can be viewed as a cross-lingual disambiguation problem. In the fourth example, we show a single word source Hindi sentence ('Freecell') which is a term and name of a computer game. The English-to-Tamil NMT system incorrectly translates this term into Tamil, and the English equivalent of the Tamil translation is in fact 'freebugs'.

Table 6: ILS in translation.

Hindi	**हाल** में खेले गए खेल के नाम [*haal mein khele gae khel ka nam*]
Tamil	விளையாட்டு பெயர்கள் நிபந்தனையின் கீழ் விளையாடப்படுகின்றன [*Viḷaiyāṭṭu peyarkaḷ nipantaṉaiyiṉ kīḻ viḷaiyāṭappaṭukiṉṟa*]

4.2.2 Lexical Selection

We observed that both NMT systems (English-to-Tamil and Hindi-to-Tamil) often make incorrect lexical selection for polysemous words, i.e. the MT systems often produce a target translation of a word that has no connection with the underlying context of the source sentence in which the word appears. As an example, we show a Hindi sentence and its Tamil translation in Table 6. The ambiguous word हाल ('*haal*') has two meanings in Hindi ('condition' and 'hall') and their Tamil translations are different too. The Hindi-to-Tamil NMT system chooses the Tamil translation for the Hindi word हाल that is incorrect in the context of the source sentence.

Table 7: Reordering error in translation.

English	It is a country of 1.25 billion people
Tamil	இது பில்லியன் மக்களுக்கு 1.25 [*Itu pilliyaṉ makkaḷukku 1.25*]

4.2.3 Wrong Word Order

We observed that the NMT systems occasionally commit reordering errors in translation. In Table 7, we show an English source sentence and its Tamil translation. The English equivalent of the Tamil translation is '*This billion people 1.25*'. As we can see, this error makes the translation less fluent.

Table 8: Word drop in translation.

English	statistics of games played
Tamil	- புள்ளிவிவரம் [*puḷḷivivaram*]
En Equiv	[statistics]

4.2.4 Word Omission

[26] observed that NMT tends to omit more terms in translation than PB-SMT. We conjecture that this can be true in our case with non-term entities too as we observed that the NMT systems often omit words in the translations although we do not compare this with PB-SMT as in [26]. As an example, in Table 8, we show an English sentence, its Tamil translation and the English equivalent of the Tamil translation. We see from the table that the NMT system translates only the first word of the English sentence and drops the remainder of the sentence during translation.

Table 9: Miscellaneous errors in translation.

	Strange Translation
Hindi	खड़ा ऊपर से अंदर [*khada oopar se andar*]
En Equiv	Standing up inside
Tamil	நில் [*Nil*]
En Equiv	*Stop*
Hindi	रपट [*rapat*]
En Eqiv	report
Tamil	நாள் [*Nāḷ*]
En Equiv	day
	Repetition of words
Hindi	नही [nahee]
En Equiv	nothing
Tamil	இல்லை இல்லை இல்லை இல்லை இல்லை இல்லை
En Equiv	nothing nothing nothing nothing nothing
Hindi	गलत [galat]
En Equiv	wrong
Tamil	தவறு தவறு தவறு தவறு தவறு
En Equiv	wrong wrong wrong wrong wrong

4.3 The BPE segmentation on the Hindi-to-Tamil translation

We saw in Section 4.1 that the BPE-based segmentation negatively impacts the translation between the two morphologically rich and complex languages, i.e. Hindi-to-Tamil. Since this segmentation process does not follow any linguistic rules and can abruptly segment a word at any character position, this may result in syntactic and morphological disagreements between the source–target sentence-pair

and aligned words, respectively. We also observed that this may violate the underlying semantic agreement between the source–target sentence-pairs. As an example, we found that the BPE segmentation breaks the Hindi word अपनों [*Aapnon*] into two morphemes अप [*Aap*] and नों [*non*], whose Tamil translation is நேசித்தவர்கள் [*Nesithavargal*], and English equivalent is 'ours'. Here, अप [Aap] is a prefix whose meaning is 'away' which no longer encodes the original meaning of 'ours' and does not correlate with the Tamil translation நேசித்தவர்கள் [*Nesithavargal*].

We show here another similar example, where the Hindi word रंगों [*rangon*] whose English equivalent is 'colors' is the translation of the Tamil word வண்ணங்கள் [*vaṇṇaṅkaḷ*]. However, when the BPE segmenter is applied to the target-side word வண்ணங்கள் [*vaṇṇaṅkaḷ*], it is split into three sub-words வ ண்ண ங்கள் [*va ṇṇa ṅkaḷ*] whose English equivalent is 'do not forget' which has no relation to வண்ணங்கள் [*vaṇṇaṅkaḷ*] (English equivalent: 'colors').

Unlike the European languages, the Indian languages are usually fully-phonetic with compulsory encoding of vowels. In our case, Hindi and Tamil differ a lot in terms of orthographic property (e.g. different phonology, no schwa deletion in Tamil). The grammatical structures of Hindi and Tamil are different too, and they are morphologically divergent and from different language families. This could be one of the reasons why the BPE-based NMT model is found to be underperforming in this translation task. This finding is corroborated by [27] who in their work found that the Morfessor-based segmentation can yield better translation quality than the BPE-based segmentation for linguistically distant language-pairs, and other way round for the close language-pairs.

5 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 the rarely tested morphologically divergent low-resource language-pairs, Hindi-to-Tamil and English-to-Tamil. We studied translations on two tasks, i.e. training data compiled from (i) freely available reasonable quality variety of data domains (e.g. political news, wikipedia), and (ii) exclusively different software localisation data domains.

We observed that the BPE-based segmentation can completely change the underlying semantic agreements of the source and target sentences of the languages with greater morphological complexity. This could be one of the reasons why the Hindi-to-Tamil NMT system's translation quality is poor when the system is trained on the sub-word-level training data in comparison to one that was trained on the word-level training data.

Use of in-domain data only at training has a positive impact on translation from a less inflected language to a highly inflected language, i.e. English-to-Tamil. However, it does not impact the Hindi-to-Tamil translation. We again conjecture that the morphological complexity of the languages (Hindi and Tamil) involved in translation could be one of the reasons why the MT systems performed reasonably poorly even when they were exclusively trained on specialised domain data.

Our error analysis on the translations by the English-to-Tamil and Hindi-to-Tamil NMT systems reveals that (i) NMT makes many mistakes when translating domain terms, and fails poorly when translating OOV terms, (ii) NMT often makes incorrect lexical selections for polysemous words and omits words and domain terms in translation, and occasionally commit reordering errors, and (iii) translations produced by the NMT systems occasionally contain repetitions of other translated words, strange translations and one or more unexpected words that have no connection with the source sentence. We observed that whenever the NMT system encounters a source sentence containing OOVs, it tends to produce one or more unexpected words or repetitions of other translated words.

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 judge errors in translations using the multidimensional quality metrics error annotation framework [28] which is a widely-used standard translation quality assessment toolkit in the translation industry and in MT research.

References

[1] Venkatesh B P Akshai Ramesh. Investigating low resource machine translation. `https://github.com/akshairamesh/Investigating_Low_Res_NMT`.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on*

Learning Representations (ICLR 2015), San Diego, CA, 2015.

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

[4] Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *HLT-NAACL 2003: conference combining Human Language Technology conference series and the North American Chapter of the Association for Computational Linguistics conference series*, pages 48–54, Edmonton, AB, 2003.

[5] Philipp Koehn and Rebecca Knowles. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, 2017.

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

[7] Meghan Dowling, Teresa Lynn, Alberto Poncelas, and Andy Way. 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, 2018.

[8] Rico Sennrich, Barry Haddow, and Alexandra Birch. 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, 2016.

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

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

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

[12] Jan Niehues and Eunah Cho. 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, 2017.

[13] Rico Sennrich and Biao Zhang. 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, 2019.

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

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

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

[17] Héctor Erasmo Gómez Montoya, Kervy Dante Rivas Rojas, and Arturo Oncevay. 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, 2019.

[18] Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 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, 2016.

[19] Sheila Castilho, Joss Moorkens, Federico Gaspari, Rico Sennrich, Vilelmini Sosoni, Panayota Georgakopoulou, Pintu Lohar, Andy Way, Antonio Valerio, Miceli Barone, and Maria Gialama. 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, 2017.

[20] Noe Casas, José AR Fonollosa, Carlos Escolano, Christine Basta, and Marta R Costa-jussà. 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, 2019.

[21] Sukanta Sen, Kamal Kumar Gupta, Asif Ekbal, and Pushpak Bhattacharyya. IITP-MT system for Gujarati-English news translation task at WMT 2019. In *Proceedings of the Fourth Conference on Machine Translation*, pages 407–411, Florence, Italy, 2019.

[22] Colin Cherry and George Foster. 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, 2012.

[23] Liang Huang and David Chiang. 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, 2007.

[24] Rico Sennrich, Barry Haddow, and Alexandra Birch. 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, August 2016.

[25] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, 2002. ACL.

[26] Rejwanul Haque, Mohammed Hasanuzzaman, and Andy Way. Investigating terminology translation in statistical and neural machine translation: A case study on English-to-Hindi and Hindi-to-English. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pages 437–446, Varna, Bulgaria, 2019.

[27] Tamali Banerjee and Pushpak Bhattacharyya. Meaningless yet meaningful: Morphology grounded subword-level NMT. In *Proceedings of the Second Workshop on Subword/Character Level Models*, pages 55–60, New Orleans, 2018.

[28] Arle Richard Lommel, Hans Uszkoreit, and Aljoscha Burchardt. Multidimensional Quality Metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumática: tecnologies de la traducció*, (12):455–463, 2014.