digital

MDPI

Article

Comparing Statistical and Neural Machine Translation Performance on Hindi-To-Tamil and English-To-Tamil

Akshai Ramesh [1], Venkatesh Balavadhani Parthasarathy [1], Rejwanul Haque [1,2,*] and Andy Way [1]

[1] ADAPT Centre, School of Computing, Dublin City University, Dublin 9, Ireland; akshai.ramesh2@mail.dcu.ie (A.R.); venkatesh.balavadhaniparthasa2@mail.dcu.ie (V.B.P.); andy.way@adaptcentre.ie (A.W.)
[2] School of Computing, National College of Ireland, Dublin 1, Ireland
* Correspondence: rejwanul.haque@adaptcentre.ie

Abstract: Phrase-based statistical machine translation (PB-SMT) has been the dominant paradigm in machine translation (MT) research for more than two decades. Deep neural MT models have been producing state-of-the-art performance across many translation tasks for four to five years. To put it another way, neural MT (NMT) took the place of PB-SMT a few years back and currently represents the state-of-the-art in MT research. Translation to or from under-resourced languages has been historically seen as a challenging task. Despite producing state-of-the-art results in many translation tasks, NMT still poses many problems such as performing poorly for many low-resource language pairs mainly because of its learning task's data-demanding nature. MT researchers have been trying to address this problem via various techniques, e.g., exploiting source- and/or target-side monolingual data for training, augmenting bilingual training data, and transfer learning. Despite some success, none of the present-day benchmarks have entirely overcome the problem of translation in low-resource scenarios for many languages. In this work, we investigate the performance of PB-SMT and NMT on two rarely tested under-resourced language pairs, English-To-Tamil and Hindi-To-Tamil, taking a specialised data domain into consideration. This paper demonstrates our findings and presents results showing the rankings of our MT systems produced via a social media-based human evaluation scheme.

Keywords: machine translation; statistical machine translation; neural machine translation; terminology translation; low-resource machine translation; byte pair encoding

Citation: Ramesh, A.; Parthasarathy, V.B.; Haque, R.; Way, A. Comparing Statistical and Neural Machine Translation Performance on Hindi-To-Tamil and English-To-Tamil. *Digital* **2021**, *1*, 86–102. https://doi.org/10.3390/digital1020007

Academic Editor: Phivos Mylonas

Received: 12 January 2021
Accepted: 20 March 2021
Published: 2 April 2021

1. Introduction

In recent years, machine translation (MT) researchers have proposed approaches to counter the data sparsity problem and to improve the performance of neural MT (NMT) systems in low-resource scenarios, e.g., augmenting training data from source and/or target monolingual corpora [1,2], unsupervised learning strategies in the absence of labelled data [3,4], exploiting training data involving other languages [5,6], multi-task learning [7], the selection of hyperparameters [8], and pre-trained language model fine-tuning [9]. 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. [1] is less effective in low-resource settings where it is hard to train a good back-translation model [10]; unsupervised MT does not work well for distant languages [11] due to the difficulty of training unsupervised cross-lingual word embeddings for such languages [12], and the same is applicable in the case of transfer learning [13]. 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 [14,15] that compared phrase-based statistical machine translation (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 [16], Östling and Tiedemann [17], Dowling et al. [18]) found that PB-SMT

can provide better translations than NMT, and many found the opposite results [8,19,20]. Hence, the findings of this line of MT research have indeed yielded a mixed bag of results, leaving the way ahead unclear.

To this end, 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 [21]. We also produced rankings of the MT systems (PB-SMT, NMT, and a commercial MT system (Google Translate (GT))) (https://translate.google.com/, (accessed on 5 March 2020) on English-To-Tamil via a social media platform-based human evaluation scheme and demonstrate our findings in this low-resource domain-specific text translation task [22]. The next section talks about some of the papers that compared PB-SMT and NMT on a variety of use-cases.

The remainder of the paper is organized as follows. In Section 2, we discuss related work. Section 3 explains the experimental setup including the descriptions of our MT systems and details of the datasets used. Section 4 presents the results with discussions and analysis, while Section 5 concludes our work with avenues for future work.

2. Related Work

The advent of NMT in MT research has led researchers to investigate how NMT is better (or worse) than PB-SMT. This section presents some of the papers that compared PB-SMT and NMT on a variety of use-cases. Although our primary objective of this work was to study translations of the MT systems (PB-SMT and NMT) in under-resourced conditions, we provide a brief overview on some of the papers that compared PB-SMT and NMT in high-resource settings as well.

Junczys-Dowmunt et al. [23] compared PB-SMT and NMT on a range of translation pairs and showed that for all translation directions, NMT is either on par with or surpasses PB-SMT. Bentivogli et al. [14] analysed the output of MT systems in an English-to-German translation task by considering different linguistic categories. Toral and Sánchez-Cartagena [24] conducted an evaluation to compare NMT and PB-SMT outputs across broader aspects (e.g., fluency, reordering) for nine language directions. Castilho et al. [15] conducted an extensive qualitative and quantitative comparative evaluation of PB-SMT and NMT using automatic metrics and professional translators. Popović [25] carried out an extensive comparison between NMT and PB-SMT language-related issues for the German–English language pair in both translation directions. The works [14,15,24,25] showed that NMT provides better translation quality than the previous state-of-the-art PB-SMT. This trend continued in other studies and use-cases: translation of literary text [26], MT post-editing setups [27], industrial setups [28], translation of patent documents [29,30], less-explored language pairs [31,32], highly investigated "easy" translation pairs [33], and the translation of catalogues of technical tools [34]. An opposite picture is also seen in the case of the translation of text pertaining to a specific domain; Nunez et al. [35] showed that PB-SMT outperforms NMT when translating user-generated content.

The MT researchers have tested and compared PB-SMT and NMT in resource-poor settings as well. Koehn and Knowles [16], Östling and Tiedemann [17] and Dowling et al. [18] found that PB-SMT can provide better translations than NMT in low-resource scenarios. In contrast to these findings, however, many studies have demonstrated that NMT is better than PB-SMT in low-resource situations [8,19]. This work investigated translations of a software localisation text with two low-resource translation pairs, Hindi-To-Tamil and English-To-Tamil, taking two MT paradigms, PB-SMT and NMT, into account.

3. Experimental Setups

3.1. The MT Systems

To build our PB-SMT systems, we used the Moses toolkit [36]. We used a 5-glanguage model trained with modified Kneser–Ney smoothing [37]. Our PB-SMT log-linear features included: (a) 4 translational features (forward and backward phrase and lexical probabil-

ities), (b) 8 lexicalised reordering probabilities (*wbe-mslr-bidirectional-fe-allff*), (c) 5-g LM-probabilities, (d) 5 OSM features [38], and (e) word count and distortion penalties. The weights of the parameters were optimized using the margin-infused relaxed algorithm [39] on the development set. For decoding, the cube-pruning algorithm [40] was applied, with a distortion limit of 12.

To build our NMT systems, we used the OpenNMT toolkit [41]. The NMT systems are Transformermodels [42]. The tokens of the training, evaluation, and validation sets were segmented into sub-word units using Byte-Pair Encoding (BPE) [43]. Recently, Sennrich and Zhang [8] 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 configurations for Transformer in our low-resource settings. In particular, we found that the following configuration led to the best results in our low-resource translation settings: (i) BPE vocabulary size: 8000, (ii) the sizes of the encoder and decoder layers: 4 and 6, respectively, (iii) the learning rate: 0.0005, (iv) the batch size (token): 4000, and (v) Transformer head size: 4. As for the remaining hyperparameters, we followed the recommended best setup from Vaswani et al. [42]. The validation on the development set was performed using three cost functions: cross-entropy, perplexity, and BLEU [44]. The early stopping criteria were based on cross-entropy; however, the final NMT system was selected as per the highest BLEU score on the validation set. The beam size for search was set to 12.

3.2. Choice of Languages

In order to test MT on low-resource scenarios, we chose English and two Indian languages: Hindi and Tamil. English, Hindi, and Tamil are Germanic, Indo-Aryan, and Dravidian languages, respectively, so the languages we selected for investigation are from different language families and morphologically divergent from each other. English is a less inflected language, whereas Hindi and Tamil are morphologically rich and highly inflected languages. Our first investigation was from a less inflected language to a highly inflected language (i.e., English-To-Tamil), and the second one was between two morphologically complex and inflected languages (i.e., Hindi-To-Tamil). Thus, we compared translation in PB-SMT and NMT with two difficult translation pairs involving three morphologically divergent languages.

3.3. Data Used

This section presents our datasets. For the experiment, we used data from three different sources: OPUS(http://opus.nlpl.eu/, (accessed on 21 January 2020) (Tanzil v1, (https://opus.nlpl.eu/Tanzil-v1.php, (accessed on 21 January 2020)) TED2020v1, (https://opus.nlpl.eu/TED2020-v1.php, (accessed on 21 January 2020) OpenSubtitles-v2018, (https://opus.nlpl.eu/OpenSubtitles-v2018.php (accessed on 21 January 2020), and IT(see below)) [45], WikiMatrix (https://ai.facebook.com/blog/wikimatrix/ (accessed on 21 January 2020)) [46], and PMIndia(http://data.statmt.org/pmindia (accessed on 21 January 2020)) [47]. As mentioned above, we carried out experiments on two translation pairs, English-To-Tamil and Hindi-To-Tamil, and studied the translation of specialised domain data, i.e., software localisation. The 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 called 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 called this setup IT. The development and test set sentences were randomly drawn from these localisation corpora. As can be seen from Table 1, the number of training set sentences of the Hindi-To-Tamil task is less than half of that of the training set size of the English-To-Tamil task.

Table 1. Data statistics.

Hindi-To-Tamil				
		Sentences.	Words [Hi]	Words [Ta]
Training sets	MIXED vocab avg. sent	100,047	1,705,034 104,564 17	1,196,008 284,921 14
	IT vocab avg. sent	48,461	3,54,426 31,258 8	2,76,514 67,069 7
devset		1500	10,903	7879
testset		1500	9362	6748
English-To-Tamil				
		Sentences	Words [En]	Words [Ta]
Training sets	MIXED vocab avg. sent	222,367	5,355,103 424,701 25	4,066,449 423,599 19
	IT vocab avg. sent	68,352	448,966 31,216 7	407,832 77,323 6
devset		1500	17,903	13,879
testset		1500	16,020	12,925

In order to remove noise from the datasets, 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 used a language identifier (cld2: https://github.com/CLD2Owners/cld2 (accessed on 21 January 2020)) 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 violate certain sentence-length ratios. In order to perform tokenisation for English, we used the standard tool in the Moses toolkit. For tokenising and normalising Hindi and Tamil sentences, we used the Indic NLP library. (https://github.com/anoopkunchukuttan/indic_nlp_library (accessed on 21 January 2020)) Without a doubt, BPE is seen as the benchmark strategy for reducing data sparsity for NMT. We built our NMT engines on both word- and subword-level training corpora in order to test BPE's effectiveness on low-resource translation tasks.

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. Additionally, we used a character-based *n*-gram precision metric chrF [48]. The confidence level (%) of the improvement obtained by one MT system with respect to another MT system is reported. An improvement in system performance at a confidence level above 95% was assumed to be statistically significant [49]. Sections 4.1.1 and 4.1.2 present the performance of the MT systems on the MIXED and IT setups, respectively.

4.1.1. The MIXED Setup

We show the BLEU and chrF scores on the test set in Table 2. 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 produced relatively low BLEU scores on the test set given the difficulty of the translation pairs. However, these BLEU scores underestimated the translation quality, given the relatively free word order in Tamil and the fact that we had only a single reference translation set for evaluation. When we compared the chrF scores with the BLEU scores, we saw that the chrF scores were quite high.

Popović [50] pointed out that the character *n*-gram F-score (i.e., chrF) is shown to correlate very well with human relative rankings of different MT outputs, especially for morphologically rich target languages. Therefore, in our case, the disparity in BLEU and chrF scores is not surprising as Tamil is a morphologically rich and complex language, and we observed that Tamil translations were penalised heavily by the BLEU metric (we discuss this issue in Section 4.3). In this regard, we quote an important observation from [51], which might be relevant here, "*n*-gram-based metrics such as BLEU significantly underplay the real benefit to be seen when NMT output is evaluated".

Table 2. The MIXED setup. PB-SMT, phrase-based statistical machine translation; NMT, neural machine translation.

	English-To-Tamil		**Hindi-To-Tamil**	
	BLEU	**chrF**	**BLEU**	**chrF**
PB-SMT	9.56	78.92	5.48	75.70
NMT	4.35	73.90	2.10	69.10

We see from Table 2 that PB-SMT surpassed NMT by a large margin in terms of BLEU and chrF in both the English-To-Tamil and Hindi-To-Tamil translation tasks, and we found that the differences in the scores were statistically significant.

4.1.2. The IT Setup

This section presents the results obtained on the IT setup. The BLEU and chrF scores of the MT systems are reported in Table 3. When we compared the BLEU scores of this table with those of Table 2, we saw 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 were found to be statistically significant. As for the Hindi-To-Tamil translation, we saw 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 was statistically significant as well. We also saw that in this task, the BLEU score of the NMT system was nearly identical to the one in the MIXED setup (2.12 BLEU points versus 2.10 BLEU points).

Table 3. The IT setup.

	English-To-Tamil		**Hindi-To-Tamil**	
	BLEU	**chrF**	**BLEU**	**chrF**
PB-SMT	15.47	83.33	4.12	73.80
NMT	9.14	79.02	2.12	69.10

As far as the English-To-Tamil translation and the IT setup are concerned, the PB-SMT system outperformed the NMT system statistically significantly, and we saw an improvement of an absolute of 6.33 points (corresponding to 69.3% relative) in terms of BLEU on the test set. The same trend was seen in the Hindi-To-Tamil translation task as well. The relative improvements of chrF scores across the different MT systems were comparable to those found with the BLEU metric.

We had a number of observations from the results of the MIXED and IT setups. As discussed in Section 3.3, in the IT task, the MT systems were built exclusively on in-domain training data, and in the MIXED setup, the training data were composed of a variety of domains, i.e., religious, IT, political news. Use of in-domain data only in training did 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 conjectured 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 small-sized specialised domain data. When we compared PB-SMT and NMT, we

saw that PB-SMT was always the leading system in both of the following cases: (i) across the training data setups (MIXED and IT) and (ii) the translation-directions (English-To-Tamil and Hindi-To-Tamil).

4.2. Data Augmentation

We carried out additional experiments by augmenting the training data from source and/or target monolingual corpora via forward- and back-translation [1,52,53]. This set of experiments was carried out for the IT translation task only. The first system was built on training data consisting of (i) authentic training data and (ii) target-original synthetic data (TOSD). The second system was built on training data consisting of (i) authentic training data, (ii) source-original synthetic data (SOSD), and (iii) TOSD. The BLEU scores of the MT systems on the test set are presented in Table 4. As can be seen from Table 4, adding synthetic data via the forward-translation strategy hurt the MT system's performance, and the back-translation strategy brought about roughly similar BLEU scores. The Tamil and English monolingual sentences were taken from the Indic corpus (https://github.com/AI4Bharat/indicnlp_corpus (accessed on 21 January 2020))[54] and the Europarl Parallel Corpus (https://www.statmt.org/europarl/ (accessed on 21 January 2020)) [55].

Table 4. The IT translation task (NMT systems built on augmented training data). TOSD, target-original synthetic data; SOSD, source-original synthetic data.

	English-To-Tamil		**Hindi-To-Tamil**	
	BLEU	**chrF**	**BLEU**	**chrF**
NMT Baseline	9.14	79.02	2.12	69.10
NMT Baseline + 1M TOSD	9.11	78.80	2.10	69.10
NMT Baseline + 1M TOSD+ 1MSOSD	8.32	77.02	1.76	68.72

4.3. Reasons for Very Low BLEU Scores

The BLEU scores reported in the sections above were 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 was indeed responsible for the increase in non-overlapping *n*-gram counts. We also found that translations contained lexical variations of Tamil words of the reference translation, again resulting in the increase of the non-overlapping *n*-gram counts. We show such translations from the Hindi-To-Tamil task in Table 5.

Table 5. Translations that are good in quality were unfairly penalised by the BLEU metric.

(1)	src: hyp: ref:	छवि आयात करें பிம்ப இறக்குமதி செய் பிம்பம் உள்வாங்கு
(2)	src: hyp: ref:	कोई गलती नहीं எந்த தவறு இல்லை பிழை இல்லை
(3)	src: hyp: ref:	information தகவல் அறிமுகம்
(4)	src: hyp: ref:	file கோப்பு file
(5)	src: hyp: ref:	authentication is required to change your own user data பயனர் தரவை மாற்ற அனுமதி தேவை உங்களுடைய சொந்த பயனர் தரவை மாற்ற அனுமதி தேவை

4.4. Error Analysis

We conducted a thorough error analysis of the English-To-Tamil and Hindi-To-Tamil NMT and PB-SMT systems built on the in-domain training data. For this, we randomly sampled 100 sentences from the respective test sets (English-To-Tamil and Hindi-To-Tamil). The outcome of this analysis is presented in the following sections.

4.4.1. Terminology Translation

Terminology translation is arguably viewed as one of the most challenging problems in MT [56–58]. Since this work focuses on studying the translation of data from a specialised domain, we looked at this area of translation with a special focus. We first looked at the translations of OOVterms 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 translated the software terms or dropped them during translation. This happened for almost all the OOV terms. Nonetheless, the NMT systems were able to correctly translate a handful of OOV terms; this phenomenon was also corroborated by Haque et al. [57] while investigating the translation of the judicial domain terms.

We show four examples in Table 6. In the first example, we show a source English sentence and its Tamil translation. We saw from the translation that the NMT system dropped the source-side terms "ipod", "iphone", and "ipad" in the target translation. The SMT system translated the segment as "most ipod, iphone". In the second example, we saw that a part ("Open") of a multiword term ("Open script") was correctly translated into Tamil, and the NMT system omitted its remaining part ("script") in the translation. As for the SMT system, the source text was translated as "opened script". In the third example, we show another multiword English term ("colour set") and its Tamil translation (i.e., English equivalent "set the colour") by the NMT system, which is wrong. As for the SMT system, the source text was translated as "set colour". Here, we saw that both the MT systems made correct lexical choices for each word of the source term, although the meaning of the respective translation was 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 the name of a computer game. The Hindi-To-Tamil NMT system incorrectly translated this term into Tamil, and the English equivalent of the Tamil translation is in fact "freebugs". The translation of the fourth segment by the SMT system was its transliteration.

Table 6. Term omission.

English	Support for most ipod / iphone / ipad devices
NMT	பெரும்பாலும் . / சாதனங்களும் ஆதரவு [perumpālum. / cātanankaḷum ātaravu]
SMT	பெரும்பாலான ipod / iphone / [perumpālāna ipod / iphone /]
English	Open Script
NMT	திற [*tira*]
SMT	திறக்கப்பட்டது தாள் [*tirakkappaṭṭatu tāḷ*]
English	Color Set
NMT	வண்ணத்தை அமைத்திடு [*vaṇṇattai amaittiṭu*]
SMT	வண்ணத்தை அமை [*vaṇṇattai amai*]
Hindi	फ्रीसेल [*Freecell*]
NMT	இலவசகளம் [*ilavacakaḷam*]
SMT	ஃப்ரீசெல் [*ilavacakaḷam*]

4.4.2. Lexical Selection

We observed that both NMT systems (English-To-Tamil and Hindi-To-Tamil) often made the incorrect lexical selection of polysemous words, i.e., the NMT systems often produced a target translation of a word that had no connection with the underlying context of the source sentence in which the word appeared. As an example, we show a Hindi

sentence and its Tamil translation in Table 7. The ambiguous words हाल (‘*haal*’) has three meanings in Hindi (‘condition’, ‘recent’ and ‘hall’) and their Tamil translations were different as well. The Hindi-To-Tamil NMT system chose the Tamil translation for the Hindi word, हाल which is incorrect in the context of the source sentence. As for the SMT system, it translated the source text as “names of games played recently”. It made the correct lexical selection for the word in question.

Table 7. Incorrect lexical selection in translation.

Hindi	हाल में खेले गए खेल के नाम [*haal mein khele gae khel ka nam*]
NMT	விளையாட்டு பெயர்கள் நிபந்தனையின் கீழ் விளையாடப்படுகின்றன [*Viḷaiyāṭṭu peyarkaḷ nipantanaiyin kīl viḷaiyāṭappaṭukina*]
SMT	சமீபத்தில் விளையாடிய விளையாட்டு பெயர்கள் [*camīpattil viḷaiyāṭiya viḷaiyāṭṭu peyarkaḷ*]

4.4.3. Wrong Word Order

We observed that the NMT systems occasionally committed reordering errors in translation. In Table 8, we show an English source sentence and its Tamil translation by the NMT system. The English equivalent of the Tamil translation is “This billion people 1.25”. As we can see, this error made the translation less fluent. The SMT system over-translated the English source sentence, i.e., “It has a population of 1.25 billion in one country”.

Table 8. Reordering error in translation.

English	It is a country of 1.25 billion people
NMT	இது பில்லியன் மக்களுக்கு 1.25 [*Itu billion makkaḷukku 1.25*]
SMT	இது ஒரு நாட்டில் 1.25 பில்லியன் மக்கள் . [*itu oru nāṭṭil 1.25 pilliyan makkaḷ*]

4.4.4. Word Omission

Haque et al. [57] observed that NMT tends to omit more terms in translation than PB-SMT. We found that this was true in our case with non-term entities as well, as we observed that the NMT systems often omitted words in the translations. As an example, in Table 9, we show an English sentence, its Tamil translations and the English equivalents of the Tamil translations. We see from the table that the NMT system translated only the first word of the English sentence and dropped the remainder of the sentence during translation, and the SMT system translated the first two words of the English sentence and dropped the remainder of the sentence for translation.

Table 9. Word drop in translation.

English	Statistics of games played
NMT	புள்ளிவிவரம் [*puḷḷivivaram*]
SMT	புள்ளிவிவரம் விளையாட்டுகளின் [*puḷḷivivaram viḷaiyāṭṭukaḷi*]

4.4.5. Miscellaneous Errors

We report a few more erroneous translations by the Hindi-To-Tamil NMT system in Table 10. The errors in these translations occurred for a variety of reasons. The translations of the source sentences sometimes contained strange words that had no relation to the meaning of the source sentence. The top two example translations belonged to this category. The translation of the first sentence by the SMT system was partially correct. As for the second example, the SMT system translated it as “report”, which is incorrect as well. We also saw that the translations occasionally contained repetitions of other translated words. This repetition of words was seen only for the NMT system. The bottom two translation examples of Table 10 belonged to this category. These findings were corroborated

by some of the studies that pursued this line of research (e.g., Farajian et al. [59]). Unsurprisingly, such erroneous translations were seen more with the Hindi-To-Tamil translation direction. As for SMT, the MT system translated the third and fourth sentences incorrectly and correctly, respectively. In both cases, unlike NMT, the translations did not contain any repetition of other translated words.

Table 10. Miscellaneous errors in translation.

Hindi	खड़ा ऊपर से अंदर [*khada oopar se andar*]
NMT	நில்[*Nil*]
SMT	உள்ளே நிற்கிறது [*ullē nirkiratu*]
Hindi	रपट [*rapat*]
NMT	நாள் [*Nāl]*
SMT	செய்தி [*ceyti*]
Hindi	नही [*nahee*]
NMT	இல்லை இல்லை இல்லை இல்லை [*llai illai illai illai illai*]
SMT	இல்லை [*llai*]
Hindi	गलत [*galat*]
NMT	தவறு தவறு தவறு தவறு [*thavaru thavaru thavaru*]
SMT	தவறு [*thavaru*]

We sometimes found the appearance of one or more unexpected words in the translation, which completely changed the meaning of the translation, as shown in Table 11. However, the SMT system correctly translated the first two source sentences shown in Table 11. In the case of the third sentence, it translated the source sentence as "move to trash".

We also observed that the translation-equivalents of some words were in fact the transliterations of the words themselves.

Table 11. Spurious Words in the translation.

English	move all to trash
NMT	அனைத்து செய்திகளும் குப்பைக்கு நகர்த்து [*anaittu ceytikaḷum kuppaikku nakarttu*]
SMT	அனைத்தையும் குப்பைக்கு நகர்த்தவும் [*anaittaiyum kuppaikku nakarttavum*]
English	data
NMT	தரவு தகவல் [*Taravu takaval*]
SMT	தகவல்கள் [*takavalkal*]
English	waste
NMT	குப்பையில் இருந்து சீட்டை நகற்று [*kuppaiyil iruntu cīṭṭai nakarru*]
SMT	குப்பையில் நகற்று [*kuppaiyil nakarru*]

We observed this happening only for the English-To-Tamil direction. For example, the English word "pixel" has a specific Tamil translation (i.e., படத்துணுக்கு [*paṭattuṇukku*]). However, the NMT system produced a transliterated form of that word in the target translation. In practice, many English words, especially terms or product names, are often directly used in Tamil text. Accordingly, we found the presence of transliterated forms of some words in the Tamil text of the training data. This could be the reason why the NMT systems generated such translations.

4.5. The BPE Segmentation on the Hindi-To-Tamil Translation

We saw in Section 4.1 that the BPE-based segmentation negatively impacted the translation between the two morphologically rich and complex languages, i.e., Hindi-To-Tamil. Since this segmentation process did not follow any linguistic rules and could 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 broke the Hindi word अपनों [*Aapnon*] into two morphemes अप [*Aap*] and नों [*non*]; the expected correct Tamil translation is நேசித்தவர்கள் [*Nesithavargal*], and the English equivalent is "ours". Here, अप [Aap] is a prefix whose meaning is "you", 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 "colours" is the translation of the Tamil word வண்ணங்கள் [*vaṇṇankal*]. However, when the BPE segmenter was applied to the target-side word வண்ணங்கள் [*vaṇnankaḷ*], it was split into three sub-words வ ண்ண ங்கள் [*va ṇṇa nkaḷ*], whose English equivalent is "do not forget", which has no relation to வண்ணங்கள் [*vaṇṇankal*] (English equivalent: "colours").

Unlike European languages, the Indian languages are usually fully phonetic with compulsory encoding of vowels. In our case, Hindi and Tamil differ greatly in terms of orthographic properties (e.g., different phonology, no schwa deletion in Tamil). The grammatical structures of Hindi and Tamil are different as well, and they are morphologically divergent and from different language families. We saw that the BPE-based segmentation could completely change the underlying semantic agreements of the source and target sentences, which, in turn, may provide the learner with the wrong (reasoning) knowledge about the sentence pairs. This could be one of the reasons why the BPE-based NMT model was found to be underperforming in this translation task. This finding was corroborated by Banerjee and Bhattacharyya [60], who in their work found that the Morfessor-based segmentation could yield better translation quality than the BPE-based segmentation for linguistically distant language pairs, and the other way round for close language pairs.

4.6. The MT System Ranking

4.6.1. Evaluation Plan

We further assessed the quality of our MT systems (the English-To-Tamil PB-SMT and NMT systems) via a manual evaluation scheme. For this, we selected our PB-SMT and NMT systems from the MIXED and IT setups. Additionally, we considered GT 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 web page 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, $3 < \text{nw} \leq 9$, and $\text{nw} > 9$. We called these sets sentence-length sets. We recall Table 1 where the average sentence length of the English IT corpus is seven. This was the justification for our choice of sentence length range. We sampled 100 sentences from the test set in such a way that the sentences were equally distributed over the sentence-length sets. Thus, the first, second and third sentence-length sets contained 34, 33, and 33 sentences, respectively. The web page displayed 10 sentences together with the translations by the MT systems, which were taken from the sentence-length sets, with a minimum of three sentences from each set. The evaluators, who were 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 (https://www.linkedin.com/ (accessed on 15 March 2020) and Facebook. (https://www.facebook.com/ (accessed on 15 March 2020)). If we were to ask the evaluators to rank a large number of sentences, it would be 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 web page 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 (https://en.wikipedia.org/wiki/Fleiss%27_kappa (accessed on 15 March 2020)). The next section presents the ranking results.

4.6.2. Ranking Results

We adopted the idea of bilingual group pairwise judgements as in Papineni et al. [44] in order to rank the MT systems. We took the pairwise scores of three MT systems and linearly normalised them across the three systems. We show our ranking results for the MIXED setup in the left half of Table 12. We see from the table that NMT was found to be the winner for first sentence-length set (nw $\leq$ 3) followed by GT and PB-SMT. As for the other sentence-length-based sets, GT became the winner followed by PB-SMT and NMT. The same trend was observed when the systems were 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 saw that PB-SMT was to the best choice, and GT and NMT were comparable if the MT systems were ranked according to the automatic evaluation scores. Therefore, the automatic evaluation results contradicted the human ranking results above.

Using the submissions from the ranking task, we also obtained 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 saw that the majority of the translations were from the NMT systems followed by the PB-SMT and GT MT systems. When we looked at the second position, we saw that PB-SMT was the winner, and NMT and GT were nearly neck-and-neck.

Table 12. Ranks of the MT systems.

	MIXED Setup			**IT Setup**		
	NMT	**PB**	**GT**	**NMT**	**PB**	**GT**
set1(nw $\leq$ 3)	1st	3rd	2nd	1st	2nd	3rd
set2 (3 < nw $\leq$ 9)	3rd	2nd	1st	2nd	1st	3rd
set3 (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 in the right half of Table 12. This time, we saw that NMT was the winner for first the sentence-length set (nw $\leq$ 3) followed by PB-SMT and GT. As for the other sentence-length-based sets and whole test set (100 sentences), PB-SMT became 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 saw that the majority of the translations that were tagged as "bes" were from PB-SMT followed by NMT and GT. In the case of the "worst" position, we saw that the majority of the translations were from the GT system followed by the NMT and PB-SMT systems. When we looked at the second position, we saw that NMT was the winner and that PB-SMT was not far behind, and the same was true for PB-SMT and GT.

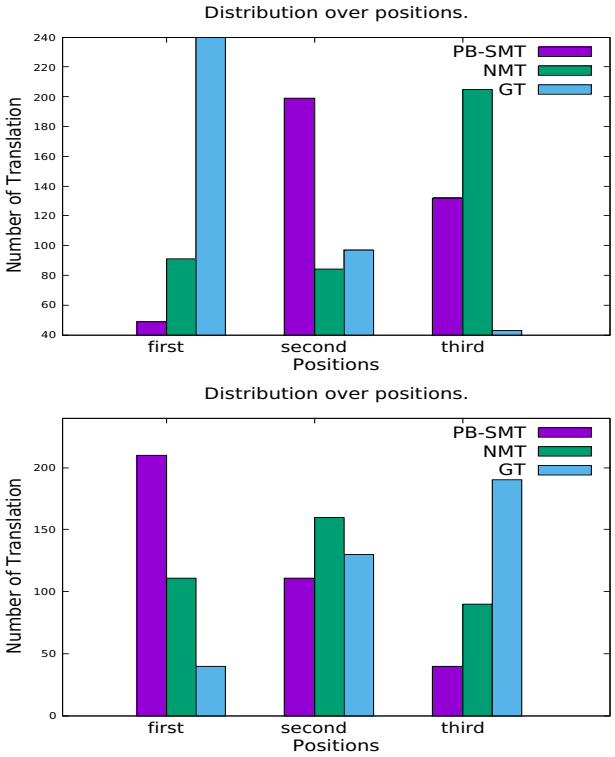

Figure 1. Distributions of translations over three positions (MIXED (**top**) and IT (**bottom**) setups). GT, Google Translate. nw, word count.

As for the first set of sentences (i.e., short sentences (nw $\leq$ 3)), we observed that the translations by the NMT systems were found to be more meaningful compared to those by the other MT systems. This was true for both the MIXED and IT setups. As an example, the English sentence "Nothing" was translated as எதுவும் இல்லை ("nothing")in Tamil by the NMT system, which, however, was 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 were 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 and 0.8 represents substantial agreement. In this sense, there was substantial agreement among the evaluators when they selected positions for the MT systems.

5. Conclusions

In this paper, we investigated NMT and PB-SMT in resource-poor scenarios, choosing a specialised data domain (software localisation) for translation and two rarely tested morphologically divergent language pairs, Hindi-To-Tamil and English-To-Tamil. We studied translations on two setups, i.e., training data compiled from (i) a 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 carried out a manual error analysis on the translations produced by our MT systems. 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.

Use of in-domain data only at training had a positive impact on translation from a less inflected language to a highly inflected language, i.e., English-To-Tamil. However, it did not impact the Hindi-To-Tamil translation. We conjectured that the morphological complexity of the source and target 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.

We looked at the translations produced by our MT systems and found that in many cases, the BLEU scores underestimated the translation quality mainly due to the relatively free word order in Tamil. In this context, Shterionov et al. [61] computed the degree of underestimation in the quality of three most widely used automatic MT evaluation metrics: BLEU, METEOR [62], and TER [63], showing that for NMT, this may be up to 50%. Way [64] reminded 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 [51] who also drew attention to this phenomenon.

Our error analysis on the translations by the English-To-Tamil and Hindi-To-Tamil MT systems revealed many positive and negative sides of the two paradigms: PB-SMT and NMT: (i) NMT made many mistakes when translating domain terms and failed poorly when translating OOV terms; (ii) NMT often made incorrect lexical selections for polysemous words and omitted words and domain terms in translation, while occasionally committing reordering errors; and (iii) translations produced by the NMT systems occasionally contained repetitions of other translated words, strange translations, and one or more unexpected words that had no connection with the source sentence. We observed that whenever the NMT system encountered a source sentence containing OOVs, it tended to produce one or more unexpected words or repetitions of other translated words. As for SMT, unlike NMT, the MT systems usually did not make such mistakes, i.e., repetitions, strange, spurious, or unexpected words in translation.

We observed that the BPE-based segmentation could 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 was poor when the system was trained on the sub-word-level training data in comparison to the one that was trained on the word-level training data.

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 turned out to be the best performing for very short sentences (number of words ≤ 3). This finding indeed did not correlate with the findings of our automatic evaluation process, where PB-SMT was found to be the best performing, while GT and NMT were comparable. This finding could be of 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 became 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 became 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 of the MT systems in the MIXED setup. As for the IT setup, PB-SMT was the first choice, NMT the second choice, and GT the third choice of the MT systems. Although a manual evaluation process is an expensive task, in the future, we want to conduct a ranking evaluation process with five MT systems, i.e., with the NMT and PB-SMT systems from MIXED and IT setups and GT.

We believe that the findings of this work provide significant contributions to this line of MT research. In the 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 [65], which is a widely used standard translation quality assessment toolkit in the translation industry and in MT research. The MT evaluation metrics such as chrF, which operates at the character level, and COMET[66], which achieved new state-of-the-art performance on the WMT2019 Metrics Shared Task [67], obtained high levels of correlation with human judgements. We intend to consider these metrics (chrF and COMET) in our future investigation. As in Exel et al. [58], who examined terminology translation in NMT in an industrial setup while using the terminology

integration approaches presented in Dinu et al. [56], we intend to investigate terminology translation in NMT using the MT models of Dinu et al. [56] on English-To-Tamil and Hindi-To-Tamil. In the future, we aim to carry out experiments with different configurations for BPE and NMT architectures including an ablation study to better understand the effects of various components and settings. We also would like to carry out experiments to see if our PB-SMT system can be improved with using monolingual training data. We aim to investigate the possibility of building BPE-based SMT models and word-based NMT models as well. Thus, we can compare word-based NMT with BPE-based NMT. Since BPE model training depends on the training data, in the future, we aim to see how effective it would be if we train the BPE models on additional monolingual data. As for building the NMT systems, we plan to perform two-stage training process where we will first train a model on the MIXED data and then "fine-tune" it on the IT data.

Author Contributions: Conceptualization, A.R., V.B.P., R.H. and A.W.; methodology, A.R., V.B.P., R.H. and A.W.; software, A.R. and V.B.P.; validation, A.R. and V.B.P.; formal analysis, A.R. and V.B.P.; investigation, A.R., V.B.P., R.H. and A.W.; resources, A.R. and V.B.P.; data curation, A.R. and V.B.P.; writing—original draft preparation, R.H. and A.W.; writing—review and editing, R.H. and A.W.; visualization, A.R. and V.B.P.; supervision, R.H. and A.W.; project administration, R.H. and A.W.; funding acquisition, R.H. and A.W. All authors have read and agreed to the published version of the manuscript.

Funding: 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. The publication has emanated from research supported in part by a research grant from SFI under Grant Number 13/RC/2077.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data used in this work is freely available for research. We have provided url link for each of the data sets used in our experiments in the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sennrich, R.; Haddow, B.; Birch, A. 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), Berlin, Germany, 7–12 August 2016; Association for Computational Linguistics: Berlin, Germany, 2016; pp. 86–96. [CrossRef]
2. Chen, P.J.; Shen, J.; Le, M.; Chaudhary, V.; El-Kishky, A.; Wenzek, G.; Ott, M.; Ranzato, M. Facebook AI's WAT19 Myanmar-English Translation Task Submission. In Proceedings of the 6th Workshop on Asian Translation, Hong Kong, China, 4 November 2019; pp. 112–122.
3. Artetxe, M.; Labaka, G.; Agirre, E. Unsupervised Statistical Machine Translation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing , Brussels, Belgium, 31 October–4 November 2018; pp. 3632–3642.
4. Lample, G.; Ott, M.; Conneau, A.; Denoyer, L.; Ranzato, M. Phrase-Based & Neural Unsupervised Machine Translation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), Brussels, Belgium, 31 October–4 November 2018.
5. Firat, O.; Cho, K.; Sankaran, B.; Vural, F.T.Y.; Bengio, Y. Multi-way, multilingual neural machine translation. *Comput. Speech Lang.* **2017**, *45*, 236–252. [CrossRef]
6. Johnson, M.; Schuster, M.; Le, Q.V.; Krikun, M.; Wu, Y.; Chen, Z.; Thorat, N.; Viégas, F.; Wattenberg, M.; Corrado, G.; Hughes, M.; Dean, J. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Trans. Assoc. Comput. Linguist.* **2017**, *5*, 339–351. [CrossRef]
7. Niehues, J.; Cho, E. Exploiting Linguistic Resources for Neural Machine Translation Using Multi-task Learning. In Proceedings of the Second Conference on Machine Translation, Copenhagen, Denmark, 7–8 September 2017; pp. 80–89.
8. Sennrich, R.; Zhang, B. Revisiting Low-Resource Neural Machine Translation: A Case Study. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; Association for Computational Linguistics: Florence, Italy, 2019; pp. 211–221. [CrossRef]
9. Liu, Y.; Gu, J.; Goyal, N.; Li, X.; Edunov, S.; Ghazvininejad, M.; Lewis, M.; Zettlemoyer, L. Multilingual Denoising Pre-training for Neural Machine Translation. *arXiv* **2020**, arXiv:2001.08210.

10. Currey, A.; Miceli Barone, A.V.; Heafield, K. Copied Monolingual Data Improves Low-Resource Neural Machine Translation. In Proceedings of the Second Conference on Machine Translation; Association for Computational Linguistics, Copenhagen, Denmark, 7–8 September 2017; pp. 148–156. [CrossRef]
11. Marie, B.; Fujita, A. Unsupervised neural machine translation initialized by unsupervised statistical machine translation. *arXiv* **2018**, arXiv:1810.12703.
12. Søgaard, A.; Ruder, S.; Vulić, I. 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), Melbourne, Australia, 15–20 July 2018; Association for Computational Linguistics: Melbourne, Australia, 2018; pp. 778–788. [CrossRef]
13. Montoya, H.E.G.; Rojas, K.D.R.; Oncevay, A. A Continuous Improvement Framework of Machine Translation for Shipibo-Konibo. In Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages, Dublin, Ireland, 19–23 August 2019; European Association for Machine Translation: Dublin, Ireland, 2019; pp. 17–23.
14. Bentivogli, L.; Bisazza, A.; Cettolo, M.; Federico, M. Neural versus Phrase-Based Machine Translation Quality: A Case Study. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, Texas, USA, 1–4 November 2016; pp. 257–267.
15. Castilho, S.; Moorkens, J.; Gaspari, F.; Sennrich, R.; Sosoni, V.; Georgakopoulou, P.; Lohar, P.; Way, A.; Valerio, A.; Barone, M.; Gialama, M. A Comparative Quality Evaluation of PBSMT and NMT using Professional Translators. In Proceedings of the MT Summit XVI, the 16th Machine Translation Summit, Nagoya, Japan, 18–22 September 2017; pp. 116–131.
16. Koehn, P.; Knowles, R. Six Challenges for Neural Machine Translation. In Proceedings of the First Workshop on Neural Machine Translation, Vancouver, BC, Canada, 30 July–4 August 2017; pp. 28–39.
17. Östling, R.; Tiedemann, J. Neural machine translation for low-resource languages. *arXiv* **2017**, arXiv:1708.05729.
18. Dowling, M.; Lynn, T.; Poncelas, A.; Way, A. SMT versus NMT: Preliminary comparisons for Irish. In Proceedings of the AMTA 2018 Workshop on Technologies for MT of Low Resource Languages (LoResMT 2018), Boston, MA, USA, 21 March 2018; pp. 12–20.
19. Casas, N.; Fonollosa, J.A.; Escolano, C.; Basta, C.; Costa-jussà, M.R. 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, Florence, Italy, 1–2 August 2019; pp. 155–162.
20. Sen, S.; Gupta, K.K.; Ekbal, A.; Bhattacharyya, P. 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), Florence, Italy, 1–2 August 2019; Association for Computational Linguistics: Florence, Italy, 2019; pp. 407–411. [CrossRef]
21. Ramesh, A.; Parthasarathy, V.B.; Haque, R.; Way, A. 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, 4–7 December 2020.
22. Ramesh, A.; Parthasarathy, V.B.; Haque, R.; Way, A. Investigating Low-resource Machine Translation for English-to-Tamil. In Proceedings of AACL-IJCNLP 2020 Workshop on Technologies for MT of Low Resource Languages (LoResMT 2020), Suzhou, China, 4–7 December 2020.
23. Junczys-Dowmunt, M.; Dwojak, T.; Hoang, H. Is neural machine translation ready for deployment? A case study on 30 translation directions. *arXiv* **2016**, arXiv:1610.01108.
24. Toral, A.; Sánchez-Cartagena, V.M. A Multifaceted Evaluation of Neural versus Phrase-Based Machine Translation for 9 Language Directions. *arXiv* **2017**, arXiv:1701.02901.
25. Popović, M. Comparing Language Related Issues for NMT and PBMT between German and English. *Prague Bull. Math. Linguist.* **2017**, *108*, 209–220. [CrossRef]
26. Toral, A.; Way, A. What level of quality can Neural Machine Translation attain on literary text? In *Translation Quality Assessment*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 263–287.
27. Specia, L.; Harris, K.; Blain, F.; Burchardt, A.; Macketanz, V.; Skadiņa, I.; Negri, M.; Turchi, M. Translation Quality and Productivity: A Study on Rich Morphology Languages. In Proceedings of the MT Summit XVI, the 16th Machine Translation Summit, Nagoya, Japan, 18–22 September 2017; pp. 55–71.
28. Shterionov, D.; Nagle, P.; Casanellas, L.; Superbo, R.; O'Dowd, T. Empirical evaluation of NMT and PBSMT quality for large-scale translation production. In Proceedings of the User Track of the 20th Annual Conference of the European Association for Machine Translation (EAMT), Prague, Czech Republic, 29–31 May 2017; pp. 74–79.
29. Long, Z.; Utsuro, T.; Mitsuhashi, T.; Yamamoto, M. Translation of Patent Sentences with a Large Vocabulary of Technical Terms Using Neural Machine Translation. In Proceedings of the 3rd Workshop on Asian Translation (WAT2016), Osaka, Japan, 11–16 December 2016; pp. 47–57.
30. Kinoshita, S.; Oshio, T.; Mitsuhashi, T. Comparison of SMT and NMT trained with large Patent Corpora: Japio at WAT2017. In Proceedings of the 4th Workshop on Asian Translation (WAT2017), Asian Federation of Natural Language Processing, Taipei, Taiwan, 27 November–1 December 2017; pp. 140–145.
31. Klubička, F.; Toral, A.; Sánchez-Cartagena, V.M. Fine-grained human evaluation of neural versus phrase-based machine translation. *arXiv* **2017**, arXiv1706.04389.
32. Klubička, F.; Toral, A.; Sánchez-Cartagena, V.M. Quantitative Fine-Grained Human Evaluation of Machine Translation Systems: A Case Study on English to Croatian. *arXiv* **2018**, arXiv1802.01451.

33. Isabelle, P.; Cherry, C.; Foster, G.F. A Challenge Set Approach to Evaluating Machine Translation. *arXiv* **2017**, arXiv1704.07431.
34. Beyer, A.M.; Macketanz, V.; Burchardt, A.; Williams, P. Can out-of-the-box NMT Beat a Domain-trained Moses on Technical Data? In Proceedings of the EAMT User Studies and Project/Product Descriptions, Prague, Czech Republic, 29–31 May 2017; pp. 41–46.
35. Nunez, J.C.R.; Seddah, D.; Wisniewski, G. Comparison between NMT and PBSMT Performance for Translating Noisy User-Generated Content. In Proceedings of the NEAL 22nd Nordic Conference on Computional Linguistics (NoDaLiDa), Turku, Finland, 30 September–2 October 2019; pp. 2–14.
36. Koehn, P.; Hoang, H.; Birch, A.; Callison-Burch, C.; Federico, M.; Bertoldi, N.; Cowan, B.; Shen, W.; Moran, C.; Zens, R.; et al. Moses: Open Source Toolkit for Statistical Machine Translation. In Proceedings of the ACL 2007, Proceedings of the Interactive Poster and Demonstration Sessions, Prague, Czech Republic, 25–27 June 2007; pp. 177–180.
37. Kneser, R.; Ney, H. Improved backing-off for M-gram language modeling. In Proceedings of the 1995 International Conference on Acoustics, Speech, and Signal Processing, Detroit, MI, USA, 9–12 May 1995; Volume 1, pp. 181–184. [CrossRef]
38. Durrani, N.; Schmid, H.; Fraser, A. A Joint Sequence Translation Model with Integrated Reordering. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, OR, USA, 21 June 2011; pp. 1045–1054.
39. Cherry, C.; Foster, G. 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, Montreal, QC, Canada, 3–8 June 2012; pp. 427–436.
40. Huang, L.; Chiang, D. Forest Rescoring: Faster Decoding with Integrated Language Models. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, 23–30 June 2007; pp. 144–151.
41. Klein, G.; Kim, Y.; Deng, Y.; Senellart, J.; Rush, A. OpenNMT: Open-Source Toolkit for Neural Machine Translation. In Proceedings of the ACL 2017, System Demonstrations, Vancouver, BC, Canada, July 30 – August 4 2017; Association for Computational Linguistics: Vancouver, BC, Canada, 2017; pp. 67–72.
42. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv/1706.03762.
43. Sennrich, R.; Haddow, B.; Birch, A. 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), Berlin, Germany, 7–12 August 2016; pp. 1715–1725.
44. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. BLEU: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the ACL-2002: 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 2002; pp. 311–318.
45. Tiedemann, J. Parallel Data, Tools and Interfaces in OPUS. In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012), Istanbul, Turkey, 23–25 May 2012; pp. 2214–2218.
46. Schwenk, H.; Chaudhary, V.; Sun, S.; Gong, H.; Guzmán, F. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *arXiv* **2019**, arXiv:1907.05791.
47. Haddow, B.; Kirefu, F. PMIndia—A Collection of Parallel Corpora of Languages of India. *arXiv* **2020**, arXiv:2001.09907.
48. Popović, M. chrF: Character n-gram F-score for automatic MT evaluation. In Proceedings of the Tenth Workshop on Statistical Machine Translation, Lisbon, Portugal, 17–18 September 2015; Association for Computational Linguistics: Lisbon, Portugal, 2015; pp. 392–395. [CrossRef]
49. Koehn, P. Statistical Significance Tests for Machine Translation Evaluation. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP), Barcelona, Spain, 25–26 July, 2004; pp. 388–395.
50. Popović, M. chrF++: Words helping character n-grams. In Proceedings of the Second Conference on Machine Translation, Copenhagen, Denmark, 7–8 September 2017; Association for Computational Linguistics: Copenhagen, Denmark, 2017; pp. 612–618. [CrossRef]
51. Way, A. Machine Translation: Where are we at today? In *The Bloomsbury Companion to Language Industry Studies*; Angelone, E., Ehrensberger-Dow, M., Massey, G., Eds.; Bloomsbury Academic Publishing: London, UK, 2019.
52. Burlot, F.; Yvon, F. Using Monolingual Data in Neural Machine Translation: A Systematic Study. In Proceedings of the Third Conference on Machine Translation: Research Papers, Belgium, Brussels, 31 October–1 November 2018; Association for Computational Linguistics: Belgium, Brussels, 2018; pp. 144–155. [CrossRef]
53. Bogoychev, N.; Sennrich, R. Domain, Translationese and Noise in Synthetic Data for Neural Machine Translation. *arXiv* **2019**, arXiv:1911.03362.
54. Kunchukuttan, A.; Kakwani, D.; Golla, S.; N.C., G.; Bhattacharyya, A.; Khapra, M.M.; Kumar, P. AI4Bharat-IndicNLP Corpus: Monolingual Corpora and Word Embeddings for Indic Languages. *arXiv* **2020**, arXiv:2005.00085.
55. Koehn, P. Europarl: A parallel corpus for statistical machine translation. In *MT Summit*; Citeseer: Phuket, Thailand, 2005; pp. 79–86.
56. Dinu, G.; Mathur, P.; Federico, M.; Al-Onaizan, Y. Training Neural Machine Translation to Apply Terminology Constraints. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; Association for Computational Linguistics: Florence, Italy, 2019; pp. 3063–3068. [CrossRef]

57. Haque, R.; Hasanuzzaman, M.; Way, A. 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, Varna, Bulgaria, 2–4 September 2019; pp. 437–446.
58. Exel, M.; Buschbeck, B.; Brandt, L.; Doneva, S. Terminology-Constrained Neural Machine Translation at SAP. In Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, Lisboa, Portugal, 3–5 November 2020; European Association for Machine Translation: Lisboa, Portugal, 2020; pp. 271–280.
59. Farajian, M.A.; Turchi, M.; Negri, M.; Bertoldi, N.; Federico, M. Neural vs. Phrase-Based Machine Translation in a Multi-Domain Scenario. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Valencia, Spain, 3–7 April 2017; Volume 2, Short Papers, pp. 280–284.
60. Banerjee, T.; Bhattacharyya, P. Meaningless yet meaningful: Morphology grounded subword-level NMT. In Proceedings of the Second Workshop on Subword/Character Level Models, New Orleans, LA, USA, 6 June 2018; pp. 55–60.
61. Shterionov, D.; Superbo, R.; Nagle, P.; Casanellas, L.; O'dowd, T.; Way, A. Human versus automatic quality evaluation of NMT and PBSMT. *Mach. Transl.* **2018**, *32*, 217–235. [CrossRef]
62. Banerjee, S.; Lavie, A. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Ann Arbor, MI, USA, 29 June 2005; Association for Computational Linguistics: Ann Arbor, MI, USA, 2005; pp. 65–72.
63. Snover, M.; Dorr, B.; Schwartz, R.; Micciulla, L.; Makhoul, J. A study of translation edit rate with targeted human annotation. In Proceedings of the Association for Machine Translation in the Americas, Cambridge, MA, USA, 8–12 August 2006.
64. Way, A. Quality expectations of machine translation. In *Translation Quality Assessment*; Castilho, S., Moorkens, J., Gaspari, F., Doherty, S., Eds.; Springer: Berlin/Heidelberg, Germany, 2018; pp. 159–178.
65. Lommel, A.R.; Uszkoreit, H.; Burchardt, A. Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. *TradumÁTica Tecnol. TraducciÓ* **2014**, 455–463. [CrossRef]
66. Rei, R.; Stewart, C.; Farinha, A.C.; Lavie, A. COMET: A Neural Framework for MT Evaluation. *arXiv* **2020**, arXiv:2009.09025.
67. Ma, Q.; Wei, J.; Bojar, O.; Graham, Y. Results of the WMT19 Metrics Shared Task: Segment-Level and Strong MT Systems Pose Big Challenges. In Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), Florence, Italy, 1–2 August 2019; Association for Computational Linguistics: Florence, Italy, 2019; pp. 62–90.