 doi:10.20944/preprints202012.0580.v1

Article

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

Akshai Ramesh,[†] Venkatesh Balavadhani Parthasarathy,[†] Rejwanul Haque,[*‡] and Andy Way[†]

ADAPT Centre;
[†] School of Computing, Dublin City University, Dublin, Ireland
[‡] School of Computing, National College of Ireland, Dublin, Ireland

akshai.ramesh2@mail.dcu.ie, venkatesh.balavadhaniparthasa2@mail.dcu.ie, rejwanul.haque@adaptcentre.ie, andy.way@adapctcentre.ie
[*] Correspondence: rejwanul.haque@adaptcentre.ie

Abstract: Statistical machine translation (SMT) which was the dominant paradigm in machine translation (MT) research for nearly three decades has recently been superseded by the end-to-end deep learning approaches to MT. Although deep neural models produce state-of-the-art results in many translation tasks, they are found to under-perform on resource-poor scenarios. Despite some success, none of the present-day benchmarks that have tried to overcome this problem can be regarded as a universal solution to the problem of translation of many low-resource languages. In this work, we investigate the performance of phrase-based SMT (PB-SMT) and 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 and rankings of our MT systems via a social media platform-based human evaluation scheme.

Keywords: Machine Translation, Statistical Machine Translation, Neural Machine Translation, Terminology Translation, Low-resource Machine Translation, Byte-Pair Encoding

1. Introduction

In recent years, MT researchers have proposed approaches to counter the data sparsity problem and to improve the performance of NMT systems in low-resource scenarios, e.g. augmenting training data from source and/or target monolingual corpora [1,2], unsupervised learning strategies in the absence of labeled data [3,4], exploiting training data involving other languages [5,6], multi-task learning [7], 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 too [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 PB-SMT and NMT on a variety of use-cases. As for low-resource scenarios, as mentioned

 doi:10.20944/preprints202012.0580.v1

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 GT) 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 data sets 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 is 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 on high-resource settings too.

Junczys-Dowmunt *et al.* [23] compare PB-SMT and NMT on a range of translation-pairs and show that for all translation directions NMT is either on par with or surpasses PB-SMT. Bentivogli *et al.* [14] analyse the output of MT systems in an English-to-German translation task by considering different linguistic categories. Toral and Sánchez-Cartagena [24] conduct an evaluation to compare NMT and PB-SMT outputs across broader aspects (e.g. fluency, reordering) for 9 language directions. Castilho *et al.* [15] conduct an extensive qualitative and quantitative comparative evaluation of PB-SMT and NMT using automatic metrics and professional translators. Popović [25] carries out an extensive comparison between NMT and PB-SMT language-related issues for the German–English language pair in both translation directions. These works [14,15,24,25] show that NMT provides better translation quality than the previous state-of-the-art PB-SMT. This trend continues 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 translation of catalogues of technical tools [34]. An opposite picture is also seen in the case of translation of the domain text; Nunez *et al.* [35] showed PB-SMT outperforms NMT when translating user-generated content.

The MT researchers have tested and compared PB-SMT and NMT in the resource-poor settings too. 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]. Hence, the findings of this line of MT research have yielded indeed a mixed bag of results, where way ahead unclear. This work investigates 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-gram language model trained with modified Kneser-Ney smoothing [37]. Our PB-SMT log-linear features include:

 doi:10.20944/preprints202012.0580.v1

(a) 4 translational features (forward and backward phrase and lexical probabilities), (b) 8 lexicalised reordering probabilities (*wbe-mslr-bidirectional-fe-allff*), (c) 5-gram LM probabilities, (d) 5 OSM features [38], and (e) word-count and distortion penalties. The weights of the parameters are optimized using the margin-infused relaxed algorithm [39] on the development set. For decoding, the cube-pruning algorithm [40] is applied, with a distortion limit of 12.

To build our NMT systems, we used the OpenNMT toolkit [41]. The NMT systems are Transformer models [42]. The tokens of the training, evaluation and validation sets are segmented into sub-word units using Byte-Pair Encoding (BPE) [43]. Recently, Sennrich and Zhang [8] demonstrated that commonly used hyper-parameters configuration 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 played with some of the hyperparameters, and found that the following configuration lead to the best results in our low-resource translation settings: (i) the BPE vocabulary size: 8,000, (ii) the sizes of encoder and decoder layers: 4 and 6, respectively, (iii) learning-rate: 0.0005, (iv) batch size (token): 4,000, and (v) Transformer head size: 4. As for the remaining hyperparameters, we followed the recommended best set-up from Vaswani *et al.* [42]. The validation on development set is performed using three cost functions: cross-entropy, perplexity and BLEU [44]. The early stopping criteria is based on cross-entropy; however, the final NMT system is selected as per highest BLEU score on the validation set. The beam size for search is set to 12.

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 to each other. English is a less inflected language, whereas Hindi and Tamil are morphologically rich and highly inflected languages. Our first investigation is from a less inflected language to a highly inflected language (i.e. English-to-Tamil), and the second one is between two morphologically complex and inflected languages (i.e. Hindi-to-Tamil). Thus, we compare 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 experimentation we used data from three different sources: OPUS[1] [45], WikiMatrix[2] [46] and PMIndia[3] [47]. As mentioned above, we carried out experiments on two translation-pairs, English-to-Tamil and Hindi-to-Tamil, and study translation of a specialised domain data, i.e. software localisation. Corpus statistics are shown in Table 1. We carried out experiments using two different setups: (i) in the first setup, the MT systems were built on a training set compiled from all data domains listed above; we call this setup MIXED, and (ii) in the second setup, the MT systems were built on a training set compiled only from different software localisation data from OPUS, *viz.* GNOME, KDE4 and Ubuntu; we call this setup IT. The development and test set sentences were randomly drawn from these localisation corpora. 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.

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

[1] http://opus.nlpl.eu/
[2] https://ai.facebook.com/blog/wikimatrix/
[3] http://data.statmt.org/pmindia

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 December 2020 doi:10.20944/preprints202012.0580.v1

Table 1. Data Statistics

Hindi-to-Tamil				
		Sentences.	Words [Hi]	Words [Ta]
Train sets	MIXED	1,00,047	1,705,034	1,196,008
	vocab		104,564	284,921
	avg. sent		17	14
	IT	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
English-to-Tamil				
		Sentences	Words [En]	Words [Ta]
Train sets	MIXED	222,367	5,355,103	4,066,449
	vocab		424,701	423,599
	avg. sent		25	19
	IT	68,352	448,966	407,832
	vocab		31,216	77,323
	avg. sent		7	6
devset		1,500	17,903	13,879
testset		1,500	16,020	12,925

we use a language identifier[4] 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 in the Moses toolkit. For tokenising and normalising Hindi and Tamil sentences, we used the Indic NLP library.[5] 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. 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% is assumed to be statistically significant [48]. 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 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.[6] The PB-SMT and

[4] cld2: https://github.com/CLD2Owners/cld2
[5] https://github.com/anoopkunchukuttan/indic_nlp_library
[6] For both translation tasks we carried out a number of experiments by augmenting the training data from source and/or target monolingual corpora via forward- and back-translation [1,49,50]. We found that adding synthetic data

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 December 2020 doi:10.20944/preprints202012.0580.v1

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

Table 2. The Mixed Setup.

	English-Tamil	Hindi-Tamil
PB-SMT	9.56	5.48
NMT	4.35	2.10

English-to-Tamil and Hindi-to-Tamil translation tasks, and found that the differences in the BLEU scores are statistically significant.

4.1.2. The IT Setup

This section presents the results obtained on the IT setup. The BLEU scores of the MT systems are reported in Table 3. When we compare the BLEU scores of this table with those of Table 2, we see a huge rise in terms of the BLEU scores for PB-SMT and NMT 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 MIXED setup (2.12 BLEU points versus 2.10 BLEU points).

Table 3. The IT Setup.

	English-to-Tamil	Hindi-to-Tamil
PB-SMT	15.47	4.12
NMT	9.14	2.12

As far as the English-to-Tamil translation and the IT setup are concerned, the PB-SMT system outperforms the NMT system statistically significantly, and we see an improvement of an absolute of 6.33 points (corresponding to 69.3% relative) in terms of BLEU on the test set. The same trend is seen in the Hindi-to-Tamil translation task too.

We have 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 is composed of a variety of domains, i.e. religious, IT, political news. 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 small-sized 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 training data setups (MIXED and IT) and (ii) the translation-directions (English-to-Tamil and Hindi-to-Tamil).

via the forward-translation strategy hurts the MT system's performance, and the back-translation strategy brings about roughly similar BLEU scores.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 December 2020 doi:10.20944/preprints202012.0580.v1

4.2. Reasons for very low BLEU Scores

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

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

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

4.3. 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.

4.3.1. Terminology Translation

Terminology translation is arguably viewed as one of the most challenging problems in MT [51–53]. Since this work focuses on studying 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 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 the OOV terms. Nonetheless, the NMT systems are able to correctly translate a handful of OOV terms; this phenomenon is also corroborated by Haque *et al.* [52] while investigating translation of the judicial domain terms.

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. The SMT system translates the segment as ‘most ipod, iphone’. In the second example, we see that a part (‘Open’) of a multiword term (‘Open script’) is correctly translated into Tamil, and the NMT system omits its remaining part (‘script’) in translation. As for the SMT system, the source text is translated as ‘opened script’. In the third example, we show another multiword English term (‘Color set’) and its Tamil translation (i.e. English equivalent ‘set

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 December 2020 doi:10.20944/preprints202012.0580.v1

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	தொற [*tira*]
SMT	தொறக்கப்பட்டது 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*]

Table 5. Term omission.

the color') by the NMT system, which is wrong. As for the SMT system, the source text is translated as 'set color'. Here, we see that both the MT systems made correct lexical choices for each word of the source term, although the meaning 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 Hindi-to-Tamil NMT system incorrectly translates 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 is its transliteration.

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

Table 6. Incorrect lexical selection in translation.

4.3.2. 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 NMT 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 three meanings in Hindi ('condition', 'recent' and 'hall') and their Tamil translations are different too. The Hindi-to-Tamil NMT system chooses the Tamil translation for the Hindi word हाल which is incorrect in the context of the source sentence. As for the SMT system, it translates the source text as "names of games played **recently**". It makes a correct lexical selection for the word in question.

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*]

Table 7. Reordering error in translation.

4.3.3. 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 by the NMT system. The English

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 December 2020 doi:10.20944/preprints202012.0580.v1

equivalent of the Tamil translation is '*This billion people 1.25*'. As we can see, this error makes the translation less fluent. The SMT system overtranslates the English source sentence, i.e. "It has a population of 1.25 billion in one country".

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

Table 8. Word drop in translation.

4.3.4. Word Omission

Haque *et al.* [52] observed that NMT tends to omit more terms in translation than PB-SMT. We found that this is true in our case with non-term entities too as we observed that the NMT systems often omit words in the translations. As an example, in Table 8, 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 translates only the first word of the English sentence and drops the remainder of the sentence during translation, and the SMT system translates the first two words of the English sentence and drops the remainder of the sentence for translation.

Hindi	खड़ा ऊपर से अंदर [*khada oopar se andar*]
NMT	நில் [*Nil*]
SMT	உள்ளே நிற்கிறது [*uḷḷē nirkiratu*]
Hindi	रपट [*rapat*]
NMT	நாள் [*Nāḷ*]
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*]

Table 9. Miscellaneous errors in translation.

4.3.5. Miscellaneous Errors

We report a few more erroneous translations by the Hindi-to-Tamil NMT system in Table 9. The errors in these translations occur for a variety of reasons. The translations of the source sentences sometimes contain strange words that have no relation to the meaning of the source sentence. The top two example translations belong to this category. The translation of the first sentence by the SMT system is partially correct. As for the second example, the SMT system translates it as 'report' which is incorrect too. We also see that the translations occasionally contain repetitions of other translated words. This repetition of words is seen only for the NMT system. The bottom two translation examples of Table 9 belong to this category. These findings are corroborated by some of the studies that pursued this line of research (e.g. Farajian *et al.* [54]). Unsurprisingly, such erroneous translations are seen more with the Hindi-to-Tamil translation direction. As for SMT, the MT system translates the third and fourth sentences incorrectly and correctly, respectively. In both cases, unlike NMT, the translations do not contain any repetition of other translated words.

We sometimes found the appearance of one or more unexpected words in the translation, which completely changes the meaning of the translation, as shown in Table 10. However, the SMT system correctly translates the first two source sentences shown in Table 10. In the case of the third sentence, it translates the source sentence as 'move to trash'.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 December 2020 doi:10.20944/preprints202012.0580.v1

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

English	move all to trash
NMT	அனைத்தை செய்திகளும் கூப்பைக்கு நகர்த்தை [*anaittu ceytikaḷum kuppaikku nakarttu*]
SMT	அனைத்தையும் கூப்பைக்கு நகர்த்தவும் [*anaittaiyum kuppaikku nakarttavum*]
English	data
NMT	தரவு தகவல் [*Taravu takaval*]
SMT	தகவல்கள் [*takavalkaḷ*]
English	waste
NMT	கூப்பையில் இருந்து சீட்டை நகற்று [*kuppaiyil iruntu cīṭṭai nakarru*]
SMT	கூப்பையில் நகற்று [*kuppaiyil nakarru*]

Table 10. Spurious Words in the translation.

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 produces 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 generates such translations.

4.4. 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*], expected correct Tamil translation is நேசித்தவர்கள் [*Nesithavargal*], and 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 'colors' is the translation of the Tamil word வண்ணங்கள் [*vaṇṇankaḷ*]. However, when the BPE segmenter is applied to the target-side word வண்ணங்கள் [*vaṇṇankaḷ*], it is split into three sub-words வ ண்ண ங்கள் [*va ṇṇa nkaḷ*] whose English equivalent is 'do not forget' which has no relation to வண்ணங்கள் [*vaṇṇankaḷ*] (English equivalent: 'colors').

Unlike 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 properties (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. We saw that the BPE-based segmentation can completely change the underlying semantic agreements of the source and target sentences, which, in turn, may provide the learner with wrong (reasoning) knowledge about the sentence-pairs. 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 Banerjee and Bhattacharyya [55] 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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 December 2020 doi:10.20944/preprints202012.0580.v1

4.5. The MT System Ranking

4.5.1. Evaluation Plan

We further assess the quality of our MT systems (the English-to-Tamil PB-SMT and NMT systems) via a manual evaluation scheme. For this, we select our PB-SMT and NMT systems from the MIXED and IT setups. Additionally, we considered Google Translate (GT)[7] in this ranking task in order to compare it with PB-SMT and NMT. We randomly sampled a set of 100 source sentences from the test set (cf. Table 1), and their translations by the MT systems including GT. In order to conduct this evaluation, we developed a webpage that was made available online and accessible to the evaluators who ranked the MT systems according to their translation quality.

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

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

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

---

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 December 2020 doi:10.20944/preprints202012.0580.v1

the automatic evaluation scores. Therefore, the automatic evaluation results contradict the human ranking results above.

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

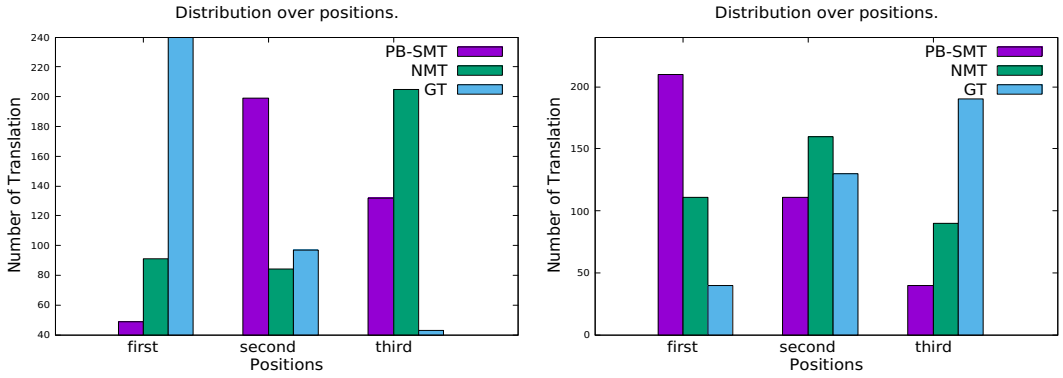

Figure 1. Distributions of translations over three positions (Mixed (top) and IT (bottom) setups).

	Mixed setup			IT setup		
	NMT	PB	GT	NMT	PB	GT
set1 (nw<=3)	1st	3rd	2nd	1st	2nd	3rd
set2 (3<nw<=9)	3rd	2nd	1st	2nd	1st	3rd
set3 (nw>9)	3rd	2nd	1st	2nd	1st	3rd
test set	3rd	2nd	1st	2nd	1st	3rd

Table 11. Ranks of the MT Systems.

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

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

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

 doi:10.20944/preprints202012.0580.v1

5. Conclusion

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) 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 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 conjecture 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 underestimate the translation quality mainly due to relatively free word order in Tamil. In this context, Shterionov *et al.* [56] computed the degree of underestimation in quality of three most-widely used automatic MT evaluation metrics: BLEU, METEOR [57] and TER [58], showing that for NMT, this may be up to 50%. Way [59] reminds the MT community how important subjective evaluation is in MT and there is no easy replacement of that in MT evaluation. We refer the interested readers to Way [60] who also drew attention to this phenomenon.

Our error analysis on the translations by the English-to-Tamil and Hindi-to-Tamil MT systems reveals many positive and negative sides of the two paradigms: PB-SMT and NMT: (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. As for SMT, unlike NMT, the MT systems usually do not make such mistakes, i.e. repetitions, strange, spurious or unexpected words in translation.

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.

From our human ranking task we found that sentence-length could be a crucial factor for the performance of the NMT systems in low-resource scenarios, i.e. NMT turns out to be best-performing for very short sentences (number of words $<= 3$). This finding indeed does not correlate with the findings of our automatic evaluation process, where PB-SMT is found to be the best-performing, and GT and NMT are comparable. This finding could be interest to translation service providers who use MT in their production for low-resource languages and may exploit the MT models based on the length of the source sentences to be translated.

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

We believe that the findings of this work provide significant contributions to this line of MT research. In future, we intend to consider more languages from different language families. We also plan to judge errors in translations using the multidimensional quality metrics error annotation framework [61] 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 [62] which operates at the character level and COMET [63] which achieved new state-of-the-art performance on the WMT 2019 Metrics Shared Task [64] 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.* [53] who examined terminology translation in NMT in an industrial setup while using the terminology integration approaches presented in Dinu *et al.* [51], we intend to investigate terminology translation in NMT using the MT models of Dinu *et al.* [51] on English-to-Tamil and Hindi-to-Tamil.

Acknowledgments

The ADAPT Centre for Digital Content Technology is funded under the Science Foundation Ireland (SFI) Research Centres Programme (Grant No. 13/RC/2106) and is co-funded under the European Regional Development Fund. The publication has emanated from research supported in part by a research grant from SFI under Grant Number 13/RC/2077.

References

1. Sennrich, R.; Haddow, B.; Birch, A. Improving Neural Machine Translation Models with Monolingual Data. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Association for Computational Linguistics: Berlin, Germany, 2016; pp. 86–96. doi:10.18653/v1/P16-1009.
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. Proceedings of the 6th Workshop on Asian Translation; , 2019; pp. 112–122.
3. Artetxe, M.; Labaka, G.; Agirre, E. Unsupervised Statistical Machine Translation. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing; , 2018; pp. 3632–3642.
4. Lample, G.; Ott, M.; Conneau, A.; Denoyer, L.; Ranzato, M. Phrase-Based & Neural Unsupervised Machine Translation. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2018.
5. Firat, O.; Cho, K.; Sankaran, B.; Vural, F.T.Y.; Bengio, Y. Multi-way, multilingual neural machine translation. *Computer Speech & Language* **2017**, *45*, 236–252.
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. *Transactions of the Association for Computational Linguistics* **2017**, *5*, 339–351.
7. Niehues, J.; Cho, E. Exploiting Linguistic Resources for Neural Machine Translation Using Multi-task Learning. Proceedings of the Second Conference on Machine Translation; , 2017; pp. 80–89.
8. Sennrich, R.; Zhang, B. Revisiting Low-Resource Neural Machine Translation: A Case Study. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; Association for Computational Linguistics: Florence, Italy, 2019; pp. 211–221. doi:10.18653/v1/P19-1021.
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 preprint arXiv:2001.08210* **2020**.
10. Currey, A.; Miceli Barone, A.V.; Heafield, K. Copied Monolingual Data Improves Low-Resource Neural Machine Translation. Proceedings of the Second Conference on Machine Translation; Association for Computational Linguistics: Copenhagen, Denmark, 2017; pp. 148–156. doi:10.18653/v1/W17-4715.
11. Marie, B.; Fujita, A. Unsupervised neural machine translation initialized by unsupervised statistical machine translation. *arXiv preprint arXiv:1810.12703* **2018**.
12. Søgaard, A.; Ruder, S.; Vulić, I. On the Limitations of Unsupervised Bilingual Dictionary Induction. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1:

Long Papers); Association for Computational Linguistics: Melbourne, Australia, 2018; pp. 778–788. doi:10.18653/v1/P18-1072.

13. Montoya, H.E.G.; Rojas, K.D.R.; Oncevay, A. A Continuous Improvement Framework of Machine Translation for Shipibo-Konibo. Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages; 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. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing; , 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. Proceedings of MT Summit XVI, the 16th Machine Translation Summit; , 2017; pp. 116–131.
16. Koehn, P.; Knowles, R. Six Challenges for Neural Machine Translation. Proceedings of the First Workshop on Neural Machine Translation; , 2017; pp. 28–39.
17. Östling, R.; Tiedemann, J. Neural machine translation for low-resource languages. *CoRR* **2017**, *abs/1708.05729*, [1708.05729].
18. Dowling, M.; Lynn, T.; Poncelas, A.; Way, A. SMT versus NMT: Preliminary comparisons for Irish. Proceedings of the AMTA 2018 Workshop on Technologies for MT of Low Resource Languages (LoResMT 2018); , 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. Proceedings of the Fourth Conference on Machine Translation; , 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. Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1); Association for Computational Linguistics: Florence, Italy, 2019; pp. 407–411. doi:10.18653/v1/W19-5346.
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. Proceedings of the 7th Workshop on Asian Translation (WAT2020); , 2020.
22. Ramesh, A.; Parthasarathy, V.B.; Haque, R.; Way, A. Investigating Low-resource Machine Translation for English-to-Tamil. Proceedings of Proceedings of the AACL-IJCNLP 2020 Workshop on Technologies for MT of Low Resource Languages (LoResMT 2020); , 2020.
23. Junczys-Dowmunt, M.; Dwojak, T.; Hoang, H. Is neural machine translation ready for deployment? A case study on 30 translation directions. *arXiv preprint arXiv:1610.01108* **2016**.
24. Toral, A.; Sánchez-Cartagena, V.M. A Multifaceted Evaluation of Neural versus Phrase-Based Machine Translation for 9 Language Directions. *CoRR* **2017**, *abs/1701.02901*, [1701.02901].
25. Popović, M. Comparing Language Related Issues for NMT and PBMT between German and English. *The Prague Bulletin of Mathematical Linguistics* **2017**, *108*, 209–220.
26. Toral, A.; Way, A. What level of quality can Neural Machine Translation attain on literary text? In *Translation Quality Assessment*; Springer, 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. Proceedings of MT Summit XVI, the 16th Machine Translation Summit; , 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. User track of the 20th Annual Conference of the European Association for Machine Translation (EAMT); , 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. Proceedings of the 3rd Workshop on Asian Translation (WAT2016); , 2016; pp. 47–57.
30. Kinoshita, S.; Oshio, T.; Mitsuhashi, T. Comparison of SMT and NMT trained with large Patent Corpora: Japio at WAT2017. Proceedings of the 4th Workshop on Asian Translation (WAT2017). Asian Federation of Natural Language Processing, 2017, pp. 140–145.

31. Klubička, F.; Toral, A.; Sánchez-Cartagena, V.M. Fine-grained human evaluation of neural versus phrase-based machine translation. *CoRR* **2017**, *abs/1706.04389*, [1706.04389].
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. *CoRR* **2018**, *abs/1802.01451*, [1802.01451].
33. Isabelle, P.; Cherry, C.; Foster, G.F. A Challenge Set Approach to Evaluating Machine Translation. *CoRR* **2017**, *abs/1704.07431*, [1704.07431].
34. Beyer, A.M.; Macketanz, V.; Burchardt, A.; Williams, P. Can out-of-the-box NMT Beat a Domain-trained Moses on Technical Data? Proceedings of EAMT User Studies and Project/Product Descriptions; , 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. NEAL Proceedings of the 22nd Nordic Conference on Computional Linguistics (NoDaLiDa), September 30-October 2, Turku, Finland. Linköping University Electronic Press, 2019, number 167, 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.; Dyer, C.; Bojar, O.; Constantin, A.; College, W.; Herbst, E. Moses: Open Source Toolkit for Statistical Machine Translation. ACL 2007, Proceedings of the Interactive Poster and Demonstration Sessions; , 2007; pp. 177–180.
37. Kneser, R.; Ney, H. Improved backing-off for M-gram language modeling. 1995 International Conference on Acoustics, Speech, and Signal Processing, 1995, Vol. 1, pp. 181–184 vol.1. doi:10.1109/ICASSP.1995.479394.
38. Durrani, N.; Schmid, H.; Fraser, A. A Joint Sequence Translation Model with Integrated Reordering. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies; , 2011; pp. 1045–1054.
39. Cherry, C.; Foster, G. Batch tuning strategies for statistical machine translation. Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; , 2012; pp. 427–436.
40. Huang, L.; Chiang, D. Forest Rescoring: Faster Decoding with Integrated Language Models. Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics; , 2007; pp. 144–151.
41. Klein, G.; Kim, Y.; Deng, Y.; Senellart, J.; Rush, A. OpenNMT: Open-Source Toolkit for Neural Machine Translation. Proceedings of ACL 2017, System Demonstrations; Association for Computational Linguistics: Vancouver, 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. *CoRR* **2017**, *abs/1706.03762*, [1706.03762].
43. Sennrich, R.; Haddow, B.; Birch, A. Neural Machine Translation of Rare Words with Subword Units. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); , 2016; pp. 1715–1725.
44. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. BLEU: a Method for Automatic Evaluation of Machine Translation. ACL-2002: 40th Annual Meeting of the Association for Computational Linguistics; ACL: Philadelphia, PA, 2002; pp. 311–318.
45. Tiedemann, J. Parallel Data, Tools and Interfaces in OPUS. Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012); , 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 preprint 1907.05791* **2019**.
47. Haddow, B.; Kirefu, F. PMIndia–A Collection of Parallel Corpora of Languages of India. *arXiv preprint 2001.09907* **2020**.
48. Koehn, P. Statistical Significance Tests for Machine Translation Evaluation. Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP); Lin, D.; Wu, D., Eds.; , 2004; pp. 388–395.
49. Burlot, F.; Yvon, F. Using Monolingual Data in Neural Machine Translation: a Systematic Study. Proceedings of the Third Conference on Machine Translation: Research Papers; Association for Computational Linguistics: Belgium, Brussels, 2018; pp. 144–155. doi:10.18653/v1/W18-6315.

50. Bogoychev, N.; Sennrich, R. Domain, Translationese and Noise in Synthetic Data for Neural Machine Translation. *arXiv preprint arXiv:1911.03362* **2019**.
51. Dinu, G.; Mathur, P.; Federico, M.; Al-Onaizan, Y. Training Neural Machine Translation to Apply Terminology Constraints. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; Association for Computational Linguistics: Florence, Italy, 2019; pp. 3063–3068. doi:10.18653/v1/P19-1294.
52. 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. Proceedings of the International Conference on Recent Advances in Natural Language Processing; , 2019; pp. 437–446.
53. Exel, M.; Buschbeck, B.; Brandt, L.; Doneva, S. Terminology-Constrained Neural Machine Translation at SAP. Proceedings of the 22nd Annual Conference of the European Association for Machine Translation; European Association for Machine Translation: Lisboa, Portugal, 2020; pp. 271–280.
54. Farajian, M.A.; Turchi, M.; Negri, M.; Bertoldi, N.; Federico, M. Neural vs. Phrase-Based Machine Translation in a Multi-Domain Scenario. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers; , 2017; pp. 280–284.
55. Banerjee, T.; Bhattacharyya, P. Meaningless yet meaningful: Morphology grounded subword-level NMT. Proceedings of the Second Workshop on Subword/Character Level Models; , 2018; pp. 55–60.
56. Shterionov, D.; Superbo, R.; Nagle, P.; Casanellas, L.; O'dowd, T.; Way, A. Human versus automatic quality evaluation of NMT and PBSMT. *Machine Translation* **2018**, *32*, 217–235.
57. Banerjee, S.; Lavie, A. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization; Association for Computational Linguistics: Ann Arbor, Michigan, 2005; pp. 65–72.
58. Snover, M.; Dorr, B.; Schwartz, R.; Micciulla, L.; Makhoul, J. A study of translation edit rate with targeted human annotation. Proceedings of association for machine translation in the Americas. Cambridge, MA, 2006, Vol. 200.
59. Way, A. Quality expectations of machine translation. In *Translation quality assessment*; Castilho, S.; Moorkens, J.; Gaspari, F.; Doherty, S., Eds.; Springer, 2018; pp. 159–178.
60. 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, 2019.
61. Lommel, A.R.; Uszkoreit, H.; Burchardt, A. Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. *Tradumática: tecnologies de la traducció* **2014**, pp. 455–463.
62. Popović, M. chrF: character n-gram F-score for automatic MT evaluation. Proceedings of the Tenth Workshop on Statistical Machine Translation, 2015, pp. 392–395.
63. Rei, R.; Stewart, C.; Farinha, A.C.; Lavie, A. COMET: A Neural Framework for MT Evaluation. *arXiv preprint arXiv:2009.09025* **2020**.
64. Ma, Q.; Wei, J.; Bojar, O.; Graham, Y. Results of the WMT19 Metrics Shared Task: Segment-Level and Strong MT Systems Pose Big Challenges. Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1); Association for Computational Linguistics: Florence, Italy, 2019; pp. 62–90.