An Error-based Investigation of Statistical and Neural Machine Translation Performance on Hindi-to-Tamil and English-to-Tamil

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

Abstract

Statistical machine translation (SMT) was the state-of-the-art in machine translation (MT) research for more than two decades, but has since been superseded by neural MT (NMT). Despite producing state-of-the-art results in many translation tasks, neural models underperform in 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.

1 Introduction

In recent years, MT researchers have proposed approaches to counter the data sparsity problem and to improve the performance of NMT systems in low-resource scenarios, e.g. augmenting training data from source and/or target monolingual corpora (Sennrich et al., 2016a; Chen et al., 2019), unsupervised learning strategies in the absence of labeled data (Artetxe et al., 2018; Lample et al., 2018), exploiting training data involving other languages (Firat et al., 2017; Johnson et al., 2017), multi-task learning (Niehues and Cho, 2017), selection of hyperparameters (Sennrich and Zhang, 2019), and pre-trained language model fine-tuning (Liu et al., 2020). Despite some success, none of the existing benchmarks can be viewed as an overall solution as far as MT for low-resource language-pairs is concerned. For examples, the back-translation strategy of Sennrich et al. (2016a) is less effective in low-resource settings where it is hard to train a good back-translation model (Currey et al., 2017); unsupervised MT does not work well for distant languages (Marie and Fujita, 2018) due to the difficulty of training unsupervised cross-lingual word embeddings for such languages (Søgaard et al., 2018) and the same is applicable in the case of transfer learning too (Montoya et al., 2019).

To this end, we investigate the performance of PB-SMT and NMT systems on two rarely-tested under-resourced language-pairs, English-to-Tamil and Hindi-to-Tamil, taking a specialised data domain (software localisation) into account. In this context, in Ramesh et al. (2020), we investigated the performance of PB-SMT, NMT and a commercial MT system (Google Translate (GT))[1] on English-to-Tamil taking the software localisation data into account, i.e. the same data as the one used in this work. In particular, in Ramesh et al. (2020), we produced rankings of the MT systems (PB-SMT, NMT and GT) via a social media platform-based human evaluation scheme, and demonstrate our findings in this low-resource domain-specific text translation task. 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

[1] https://translate.google.com/

Proceedings of the 7th Workshop on Asian Translation, pages 178–188
December 4, 2020. ©2020 Association for Computational Linguistics

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. (2016) 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. (2016) analyse the output of MT systems in an English-to-German translation task by considering different linguistic categories. Toral and Sánchez-Cartagena (2017) conduct an evaluation to compare NMT and PB-SMT outputs across broader aspects (e.g. fluency, reordering) for 9 language directions. Castilho et al. (2017) conduct an extensive qualitative and quantitative comparative evaluation of PB-SMT and NMT using automatic metrics and professional translators. Popović (2017) 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 (Bentivogli et al., 2016; Castilho et al., 2017; Popović, 2017; Toral and Sánchez-Cartagena, 2017) 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 (Toral and Way, 2018), MT post-editing setups (Specia et al., 2017), industrial setups (Shterionov et al., 2017), translation of patent documents (Long et al., 2016; Kinoshita et al., 2017), less-explored language pairs (Klubička et al., 2017, 2018), highly investigated "easy" translation pairs (Isabelle et al., 2017), and translation of catalogues of technical tools (Beyer et al., 2017). An opposite picture is also seen in the case of translation of the domain text; Nunez et al. (2019) 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 (2017), Östling and Tiedemann (2017), and Dowling et al. (2018) 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 (Casas et al., 2019; Sennrich and Zhang, 2019). 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 (Koehn et al., 2007). We used a 5-gram language model trained with modified Kneser-Ney smoothing (Kneser and Ney, 1995). Our PB-SMT log-linear features include: (a) 4 translational features (forward and backward phrase and lexical probabilities), (b) 8 lexicalised reordering probabilities (*wbe-mslr-bidirectional-fe-allff*), (c) 5-gram LM probabilities, (d) 5 OSM features (Durrani et al., 2011), and (e) word-count and distortion penalties. The weights of the parameters are optimized using the margin-infused relaxed algorithm (Cherry and Foster, 2012) on the development set. For decoding, the cube-pruning algorithm (Huang and Chiang, 2007) is applied, with a distortion limit of 12.

To build our NMT systems, we used the OpenNMT toolkit (Klein et al., 2017). The NMT systems are Transformer models (Vaswani et al., 2017). The tokens of the training, evaluation and validation sets are segmented into sub-word units using Byte-Pair Encoding (BPE) (Sennrich et al., 2016b). Recently, Sennrich and Zhang (2019) demonstrated that commonly used hyper-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. (2017). The validation on development set is performed using three cost functions: cross-entropy, perplexity and BLEU (Papineni et al., 2002). The early stopping criteria is based on cross-entropy; however, the final NMT system is selected as per highest BLEU score on the validation set. The beam size for search is set to 12.

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[2] (Tiedemann, 2012), WikiMatrix[3] (Schwenk et al., 2019) and PMIndia[4] (Haddow and Kirefu, 2020). 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.

Table 1: Data Statistics

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

In order to remove noise from the data sets, we adopted the following measures. We observed that the corpora of one language (say, Hindi) contains sentences of other languages (e.g. English), so we use a language identifier[5] 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.[6] 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 performed statistical significance tests using bootstrap resampling methods (Koehn, 2004). 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

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

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

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

[5] cld2: https://github.com/CLD2Owners/cld2

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

tasks, respectively.[7] The PB-SMT and NMT systems produce relatively low BLEU scores on the test set given the difficulty of the translation pairs. However, these BLEU scores underestimate the translation quality, given the relatively free word order in Tamil, and the fact that we have just a single reference translation set for evaluation. We see from Table 2 that PB-SMT surpassed NMT by a large margin in terms of BLEU in both the English-to-Tamil and Hindi-to-Tamil translation tasks, and found that the differences in the BLEU scores are statistically significant.

Table 2: The Mixed Setup.

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

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).

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. We also reported this phenomenon in Ramesh et al. (2020) and showed such translations from the English-to-Tamil task (cf. Table 3; Section 3.2 of Ramesh et al. (2020)).

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-

[7] 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 (Sennrich et al., 2016a; Burlot and Yvon, 2018; Bogoychev and Sennrich, 2019). We found that adding synthetic data via the forward-translation strategy hurts the MT system's performance, and the back-translation strategy brings about roughly similar BLEU scores.

(1)	src:	छवि आयात करें
	hyp:	பிம்ப இறக்குமதி செய்
	ref:	பிம்பம் உள்வாங்கு
(2)	src:	कोई गलती नहीं
	hyp:	எந்த தவறு இல்லை
	ref:	பிழை இல்லை

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

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 (Dinu et al., 2019; Haque et al., 2019; Exel et al., 2020). 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. (2019) while investigating translation of the judicial domain terms.

Eng	Support for most ipod / iphone / ipad devices
NMT	பெரும்பாலும் . / சாதனங்களும் ஆதரவு [perumpālum. / cātanankaḷum ātaravu]
SMT	பெரும்பாலான ipod / iphone / [perumpālāna ipod / iphone /]
Eng	Open Script
NMT	திற [*tira*]
SMT	திறக்கப்பட்டது தாள் [*tirakkappaṭṭatu tāḷ*]
Eng	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.

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 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 kīl 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.

4.3.3 Wrong Word Order

We observed that the NMT systems occasionally commit reordering errors in translation. In Table 7,

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.

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 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".

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

Table 8: Word drop in translation.

4.3.4 Word Omission

Haque et al. (2019) 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. (2017)). 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'.

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

Eng.	move all to trash
NMT	அனைத்து செய்திகளும் குப்பைக்கு நகர்த்து [*anaittu ceytikaḷum kuppaikku nakarttu*]
SMT	அனைத்தையும் குப்பைக்கு நகர்த்தவும் [*anaittaiyum kuppaikku nakarttavum*]
Eng.	data
NMT	தரவு தகவல் [*Taravu takaval*]
SMT	தகவல்கள் [*takavalkaḷ*]
Eng.	waste
NMT	குப்பையில் இருந்து சீட்டை நகற்று [*kuppaiyil iruntu cīṭṭai nakarru*]
SMT	குப்பையில் 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 [*vaṇnankaḷ*]. However, when the BPE segmenter is applied to the target-side word வண்ணங்கள் [*vaṇnankaḷ*], it is split into three subwords வ ண்ண ங்கள் [*va ṇna nkaḷ*] whose English equivalent is 'do not forget' which has no relation to வண்ணங்கள் [*vaṇnankaḷ*] (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 (2018) who in their work found that the Morfessor-based segmentation can yield better translation quality than the BPE-based segmentation for linguistically distant language-pairs, and other way round for the close language-pairs.

5 Conclusion

In this paper, we investigated NMT and PB-SMT in resource-poor 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.

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. (2018) computed the degree of underestimation in quality of three most-widely used automatic MT evaluation metrics: BLEU, METEOR (Banerjee and Lavie, 2005) and TER (Snover et al., 2006), showing that for NMT, this may be up to 50%. We refer the interested readers to Way (2018, 2019) 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.

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 (Lommel et al., 2014) 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 (Popović, 2015) which operates at the character level and COMET (Rei et al., 2020) which achieved new state-of-the-art performance on the WMT 2019 Metrics Shared Task (Ma et al., 2019) 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. (2020) who examined terminology translation in NMT in an industrial setup while using the terminology integration approaches presented in Dinu et al. (2019), we intend to investigate terminology translation in NMT using the MT models of Dinu et al. (2019) 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. This project has partially received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 713567, and the publication has emanated from research supported in part by a research grant from SFI under Grant Number 13/RC/2077.

References

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

Satanjeev Banerjee and Alon Lavie. 2005. 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*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

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

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

Anne M Beyer, Vivien Macketanz, Aljoscha Burchardt, and Philip Williams. 2017. Can out-of-the-box NMT beat a Domain-trained Moses on Technical Data? In *Proceedings of EAMT User Studies and Project/Product Descriptions*, pages 41–46, Prague, Czech Republic.

Nikolay Bogoychev and Rico Sennrich. 2019. Domain, translationese and noise in synthetic data for neural machine translation. *arXiv preprint arXiv:1911.03362*.

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

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

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

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

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

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

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy. Association for Computational Linguistics.

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

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

Miriam Exel, Bianka Buschbeck, Lauritz Brandt, and Simona Doneva. 2020. Terminology-constrained neural machine translation at SAP. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 271–280, Lisboa, Portugal. European Association for Machine Translation.

M. Amin Farajian, Marco Turchi, Matteo Negri, Nicola Bertoldi, and Marcello Federico. 2017. 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: Volume 2, Short Papers*, pages 280–284, Valencia, Spain.

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

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

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

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

Pierre Isabelle, Colin Cherry, and George F. Foster. 2017. A challenge set approach to evaluating machine translation. *CoRR*, abs/1704.07431.

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

Marcin Junczys-Dowmunt, Tomasz Dwojak, and Hieu Hoang. 2016. Is neural machine translation ready for deployment? a case study on 30 translation directions. *arXiv preprint arXiv:1610.01108*.

Satoshi Kinoshita, Tadaaki Oshio, and Tomoharu Mitsuhashi. 2017. Comparison of smt and nmt trained with large patent corpora: Japio at wat2017. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 140–145. Asian Federation of Natural Language Processing.

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

Filip Klubička, Antonio Toral, and Víctor M. Sánchez-Cartagena. 2017. Fine-grained human evaluation of neural versus phrase-based machine translation. *CoRR*, abs/1706.04389.

Filip Klubička, Antonio Toral, and Víctor M. Sánchez-Cartagena. 2018. Quantitative fine-grained human evaluation of machine translation systems: a case study on English to Croatian. *CoRR*, abs/1802.01451.

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

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

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

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

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

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

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

Zi Long, Takehito Utsuro, Tomoharu Mitsuhashi, and Mikio Yamamoto. 2016. 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)*, pages 47–57, Osaka, Japan.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. 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)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

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

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

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

José Carlos Rosales Nunez, Djamé Seddah, and Guillaume Wisniewski. 2019. Comparison between nmt and pbsmt performance for translating noisy user-generated content. In *NEAL Proceedings of the 22nd Nordic Conference on Computional Linguistics (NoDaLiDa), September 30-October 2, Turku, Finland*, 167, pages 2–14. Linköping University Electronic Press.

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

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

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395.

Maja Popović. 2017. Comparing language related issues for nmt and pbmt between German and English. *The Prague Bulletin of Mathematical Linguistics*, 108(1):209–220.

Akshai Ramesh, Venkatesh Balavadhani Parthasarathy, Rejwanul Haque, and Andy Way. 2020. Investigating low-resource machine translation for English-to-Tamil. In *Proceedings of Proceedings of the AACL-IJCNLP 2020 Workshop on Technologies for MT of Low Resource Languages (LoResMT 2020)*, Suzhou, China.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*.

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany.

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

Dimitar Shterionov, Pat Nagle, Laura Casanellas, Riccardo Superbo, and Tony O'Dowd. 2017. Empirical evaluation of nmt and pbsmt quality for large-scale translation production. In *User track of the 20th Annual Conference of the European Association for Machine Translation (EAMT)*, pages 74–79, Prague, Czech Republic.

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

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, volume 200. Cambridge, MA.

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

Lucia Specia, Kim Harris, Frédéric Blain, Aljoscha Burchardt, Vivien Macketanz, Inguna Skadiņa, Matteo Negri, and Marco Turchi. 2017. Translation quality and productivity: A study on rich morphology languages. In *Proceedings of MT Summit XVI, the 16th Machine Translation Summit*, pages 55–71, Nagoya, Japan.

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

Antonio Toral and Víctor M. Sánchez-Cartagena. 2017. A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. *CoRR*, abs/1701.02901.

Antonio Toral and Andy Way. 2018. What level of quality can neural machine translation attain on literary text? In *Translation Quality Assessment*, pages 263–287. Springer.

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

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

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