Machine Translation (2021) 35:71–99
https://doi.org/10.1007/s10590-021-09263-3

Experience of neural machine translation between Indian languages

Shubham Dewangan[1] · **Shreya Alva**[1] · **Nitish Joshi**[1] · **Pushpak Bhattacharyya**[1]

Received: 25 February 2020 / Accepted: 29 March 2021 / Published online: 4 May 2021

Abstract

In this paper we explore neural machine translation (NMT) for Indian languages. Reported work on Indian language Statistical Machine Translation (SMT) demonstrated good performance within the Indo-Aryan family, but relatively poor performance within the Dravidian family as well as between the two families. Interestingly, by common observation NMT generates more fluent output than SMT. This led us to investigate NMT's potential for translation involving Indian languages. The current practice in NMT is to train the models with subword units. Among subwording methods, byte pair encoding (BPE) is a popular choice. We conduct extensive experiments with BPE-based NMT models for Indian languages. An interesting outcome of our study is the finding that the optimal value for *BPE merge* for Indian language pairs seems to be falling in the range of 0–5000 which is fairly low compared to that observed for European Languages. Additionally, we apply other techniques such as phrase table injection and linguistic feature based enhancements on corpora, plus BERT augmented NMT to boost performance. To the best of our knowledge, this is the first comprehensive study on Indian language NMT (ILNMT) covering major languages in India. As an empirical paper, we expect this work could serve as a benchmark for ILNMT research.

Keywords Neural machine translation · Indian languages · BiLSTM encoder decoder model · Byte pair encoding (BPE) · Phrase table injection · Morpheme and word features

✉ Shubham Dewangan
shubhamd@cse.iitb.ac.in

Shreya Alva
shreya@cse.iitb.ac.in

Nitish Joshi
nitishjoshi@cse.iitb.ac.in

Pushpak Bhattacharyya
pb@cse.iitb.ac.in

[1] IIT Bombay, Mumbai, India

1 Introduction

India is a diverse nation in many respects, culturally, geographically, and certainly linguistically. It is home to 780 languages, 22 amongst them being 'scheduled' (as per the Constitution of India). 21 out of these 22 scheduled languages as well as 10 out of 99 non-scheduled languages have over a million speakers[1]. Despite, or perhaps due to this linguistic diversity, most of these languages can be categorized as of relatively low resource. These languages belong to five families—Indo-Aryan (popular members being Hindi, Punjabi, Gujarati, Marathi, Bengali), Dravidian (notably Tamil, Telugu, Malayalam, Kannada), Austro-Asiatic (khasi and Munda) and Sino-Tibetan (e.g., Manipuri and Bodo). This diversity makes translation solutions essential, to ease the communication between the myriad Indian states, a concern that the government has ever been keen to address. Since the advent of neural techniques in Machine Translation (MT), a lot of work has been done for European languages (Sennrich et al. 2016; Ding et al. 2019b). We aim to extend research in the field to incorporate Indian languages as well-with our focus being Hindi, Punjabi, Gujarati, Marathi, Bengali, Tamil, Telugu and Malayalam. To the best of our knowledge, this is a first of its kind effort.

Bahdanau et al. (2015) was a seminal work in neural machine translation (NMT). It introduced the bi-directional encoder–decoder model which translates input sentences by encoding the one-hot representation of its constituent words using a forward and a backward RNN, mapping them to produce annotations. These annotations are used to generate context for the word. The output sentence is generated word by word, in a sequential manner. The decoder takes the context, the previously generated output word and the hidden state of the decoder as input to predict the next output word. Such systems mandated that the size of the vocabulary be fixed due to memory and computational constraints (selecting 50,000 most frequent words as the vocabulary size was a popular choice). However, this rendered these systems unable to deal with the OOV (out of vocabulary) problem. They could not translate unseen words in the test set correctly. Such OOV words are replaced by a special token, <UNK>. OOV words impact adversely the output fluency and adequacy.

1.1 Byte pair encoding

Like in all cases of NMT, our work also makes heavy use of byte pair encoding (BPE). Instead of using words as input and output tokens during translation, Sennrich et al. (2016) found that subword models not only make the translation process simpler but also address the OOV problem[2]. Currently, the most common subword methods are BPE, WordPiece (Wu et al. 2016) and subword based procesing like orthographic syllables (Kudo 2018a). BPE is a technique that iteratively merges the most frequent pair of characters or character sequences into a sequence with a

[1] https://censusindia.gov.in/2011Census/C-16_25062018_NEW.pdf.
[2] https://github.com/rsennrich/subword-nmt.

single, unused character sequence. The number of iterations for this algorithm is a hyperparameter called merge operations. WordPiece (Schuster and Nakajima 2012) is similar to BPE, except that it merges by likelihood instead of frequency. Subword regularization (Kudo 2018b) harnesses Bayesian sampling to account for different segmentation possibilities and assigns probabilities to those, enabling the system to tackle segmentation ambiguity. In this work, we will focus on BPE. While BPE is widely used as data processing step in almost all NMT experiments, in the interest of time-complexity and/or computational power, the number of merge operations is often not tuned. Finding the optimal number of merge operations can lead to markedly better performance, all other parameters of the system being the same. In prior research, we find researchers use a relatively large number of merge operations (from 30k to 90k). One question we sought to answer in this work is the optimal number of merge operations required for translating between Indian languages using BiLSTM models in a low-resource setting. These BPE based NMT systems also serve as our baseline.

We also decided to enhance the NMT model with some tweaks. Recognizing the power of data in training NMT systems, we have created models enhanced with additional data sources and linguistic features, establishing state-of-the-art BLEU scores for some Indian language pairs. To the best of our knowledge, no prior work compiles findings for NMT systems among Indian languages.

2 A brief look at Indian language characteristics and properties

As per Census 2011/ Ethnologue[3], "Languages spoken in India belong to five language families, the major ones being the Indo-Aryan languages spoken by 78.05% of Indians and the Dravidian languages spoken by 19.64% of Indians. Languages spoken by the remaining 2.31% of the population belong to the Austroasiatic, Sino-Tibetan, Tai-Kadai and a few other minor language families and isolates".[4] Indian languages (henceforth IL) have their own characteristics and properties which have bearing on their translatability within themselves and from-to languages outside. For example, almost all Indian languages are SOV (subject-object-verb) ordered, while English is SVO (subject–verb–object). Hence a translator—human or automatic—has to apply the transformation VO → OV, while translating. In this section we give an account of typical IL characteristics. In what follows, we describe a few well-known and well-cited language phenomena common across ILs. For the first of these, viz., "Reduplication", we describe computational aspects too in detail from the point of view of machine translation. For these discussions, we use Roman alphabets for all languages, since there is large a variety of scripts in ILs. Thus आम in Hindi (written in Devnagari script)

[3] https://censusindia.gov.in/2011Census/C-16_25062018_NEW.pdf.

[4] Figure 1 has been constructed by drawing inspiration from the following images: https://images.app.goo.gl/CYukRDcQTsytwpQ67, https://qphs.fs.quoracdn.net/main-qimg-f6e580591e48cc0829fdffcc8d4f1ae3, https://en.wikipedia.org/wiki/File:AustroAsiatic_tree_Peiros2004.png.

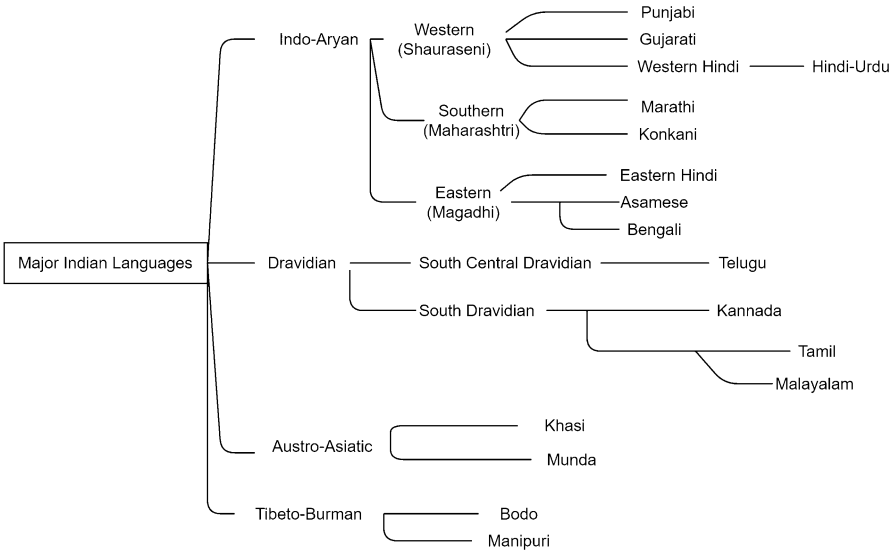


Fig. 1 Tree diagram to illustrate the language closeness of major Indian languages

meaning 'mango' in English will be written throughout this section as 'aam' (notice the doubling of 'a' to represent the long vowel आ).

2.1 Common IL characteristics and properties

Reduplication This is the language phenomenon of repetition of a word to express many speech acts like intensity, plurality, emphasis and so on. All parts of speech can be reduplicated. Below is an example of noun reduplication:

H-redup: *ghar ghar meM* (Hindi)
HG-redup: *home home in* (Hindi Gloss)
E-redup: *in all homes* (English)

Here the reduplication of *'ghar'* indicates plurality-*homes*. Since reduplication is a pan-Indian phenomenon, its translation is not a major challenge—simply replicate the reduplication of the source language in the target language. Notice that the English sentence E-redup has no repeated word *'home'*. Thus transfer of reduplicate across ILs is easier to manage than for to-and-from languages outside. Sometimes, a bit of additional work may be necessary. For example, Bengali equivalent of *"ghar ghar meM"* is:

B-redup: *ghare ghare* (Bengali)
BG: *home-in home-in* (Bengali Gloss)

Table 1 Different behaviours of the phrases in various languages corresponding to the Hindi Phrase "*ghar ghar meM*"

Languages	Has two constituents?	Constituents joined	Suffix/postposition/null for 1st constituent	Suffix/postposition/null for 2nd constituent
Bengali	Yes	No	Suffix	Suffix
Gujarati	Yes	No	Null	Null
Marathi	Yes	Yes	Suffix	Suffix
Punjabi	Yes	No	Null	Null
Tamil	Yes	No	Null	Suffix
Telugu	Yes	Yes	Null	Suffix
English	No	Yes	Not applicable (NA)	NA

Notice the *'e'* morpheme meaning *'in'* which is attached to both the constituent words of the reduplicate. In Hindi, the *'meM'* string came in only for the second *'ghar'*. Also *'meM'* is postposition for Hindi, while *'e'* is suffix for Bengali. There are more varieties across languages, though the same fundamental process applies:

Gujarati: *Ter Ter* (*'T'* is retroflexive *'t'*; constituents do not get attached; both get neither suffix nor postposition)
Marathi: *gharogharii* (constituents are joined, with the first constituent getting the *'o'* suffix and the second the *'ii'* suffix)
Nepali: *ghar ghar* (neither suffix, nor postposition)
Punjabi: *ghar ghar* (like Nepali)
Telugu: *intiintiki* (constituents joined, first constituent gets 'null' suffix, while the second gets *'ki'*)
Tamil: *viidu viidaak* (first constituent null suffix, second constituent *'aak'* suffix)

Thus for Hindi *"ghar ghar meM"*, we can create Table 1 by way of documenting the translation requirement:

This discussion makes it apparent that with some adjustments translation of reduplicates is easier for within family languages, with the amount of adjustments increasing with "language distance". It will now be interesting to see how MT will be able to handle reduplicates. In RBMT (rule based MT), we will have to give explicit rules for transfer. Thus Hindi → Punjabi will work with a rule like:

$$X_{hindi}\, X_{hindi}\, meM \rightarrow X_{punjabi}\, X_{punjabi}$$

Such transfer is within the ambit of regular expressions.

In Statistical MT (SMT), the phrase table captures reduplication correspondences. Since reduplication maintains the fixity of the structure strongly, i.e., does not allow the two constituents to be distant from each other, the phrase correspondences are accurately recorded in the phrase table and are used in the

decoding stage. In NMT, the decoder state and the attention input to the decoder largely captures reduplication. With abundant training data, all forms of reduplication get encoded in the encoder.

Now, we mention a few other common IL characteristics, without discussing the machinery to translate them.

Dative subjects The experiencer subject takes dative case marking in ILs:

H-dative-subject: *mujhe sar dard hai* (Hindi)
HG-dative-subject: *to me head ache exists* (Hindi Gloss)
E-dative-subject: *I have a head ache* (English)

Notice that English uses nominative case. The translation between English and Hindi needs to arrange for Nominative ↔ Dative transfer, but not so much across Indian languages, except for morphological adjustments (suffix vs. postpositions).

Conjunctive particles these constructs indicate sequentiality of actions.

H-conjunctive-particle: *ghar jaa kar khaanaa khaayaa* (Hindi)
HG-conjunctive-particle: *home go do food ate* (Hindi Gloss)
E-conjunctive-particle: *(somebody) ate after going home* (English)

The 'kar' particle in Hindi is called the conjunctive particle and is seen in almost all Indian languages either as postposition or as a suffix (e.g., in Bengali).

Conjunct verbs ILs use "verbalizers" on nouns and adjectives to form what are called conjunct verbs. Conjunct Verbs are forms of complex predicates and are a pan-IL phenomenon. Examples are:

H-conjunct-verb-on-noun: *salaha denaa* (Hindi)
HG-conjunct-verb-on-noun: *advice give* (gloss)
E-conjunct-verb-on-noun: *advise* (English)
H-conjunct-verb-on-adj: *saaf karnaa* (Hindi)
HG-conjunct-verb-on-adj: *clean do* (Hindi Gloss)
E-conjunct-verb-on-adj: *clean* (English verb)

Compound verbs Compound Verbs are also forms of complex predicates. Two verbs form this structure. The first verb is called the polar and the second the vector. The polar verb carries the semantic load and the vector the syntactic, speech act and aspectual load. The vector verb is typically from a small set of verbs and in the compound verb formation, the verb is 'bleached' of its normal meaning.

H-compound-verb: *bol uthnaa* (Hindi)
HG-compound-verb: *speak rise* (Hindi Gloss)
E-compound-verb: *start speaking abruptly* (English)
B-compound-verb: *heshe phelaa* (Hindi)
BG-compound-verb: *laugh drop* (Bengali Gloss)
E-compound-verb: *laugh suddenly* (English, often means without control)

Understanding and accounting for IL phenomena such as delineated above help (a) choose the training data for machine translation more effectively, (b) do error analysis more insightfully and (c) demarcate the scope of our system more clearly. An authentic and exhaustive treatment of IL properties and characteristics is Subbārāo (2012).

3 Related work

This work contains systematic experimentation with different BPE settings and techniques to boost NMT performance such as phrase table injection and supplying morpheme and word level features in the context of Indian languages MT.

As background to our work we discuss relevant prior work. Sennrich et al. (2016) proposed the BPE method and compared the system performance when using 59,500 BPE and 89,500 joint BPE operations for English-German and English-Russian language pairs respectively. They found 90k merge operations to work well and used this figure for their winning submission for WMT 2017 new translation shared task (Sennrich et al. 2017).

Wu et al. (2016) experimented extensively with WMT data for English-German and English-French language pairs and recommended 8000 to 32,000 merge operations to achieve optimal BLEU score performance for the WordPiece method. Denkowski and Neubig (2017) explored several hyperparameter settings, including the number of BPE merge operations, to establish a strong baseline for NMT on LSTM-based architectures. While Denkowski and Neubig (2017) demonstrated that BPE models are clearly better than word-level models, their experiments on 16k and 32k BPE configuration did not show much difference. They, therefore, recommended 32K as generally effective vocabulary size and 16K as a contrasting condition when building systems on less than 1 million parallel sentences. However, while studying deep character-based LSTM-based translation models, Cherry et al. (2018) also ran experiments for BPE configurations between 0 and 32k merge operations and found that the system performance deteriorates with the increasing number of BPE merge operations.

Recently, Renduchintala et al. (2018) also showed that it is important to tune the number of BPE merge operations and found no typical optimal BPE configuration for their LSTM-based architecture, while doing experiments over several language pairs in the low-resource setting. There appears to be no consensus on what the best practice for BPE application should be in terms of the hyperparameter of number of merge operations. In a recent work, Ding et al. (2019a) conducted experiments with all the data from IWSLT 2016 shared task, covering translation of English from and to Arabic, Czech, French and German with LSTM and Transformer architectures. They report that for LSTMs, there is no typical optimal BPE configuration, whereas for Transformer architectures, generally, a smaller number of BPE merge operations is an optimal choice. As for Indian languages, Kunchukuttan and Bhattacharyya (2016) compare BPE and orthographic syllable as translation units for Statistical Machine Translation across multiple language families and find that BPE emerges as a superior choice.

Tang et al. (2016) proposed an end-to-end learning algorithm with an external phrase memory that maintained reliable phrase translations (not multiple mappings as are present in the phrase table used by SMT systems). They utilize their phrase table, which is essentially a list of rules, to preprocess data to split the words in the source and target sentences into two groups, the phrases and the words not-in-phrases. They proposed a modified decoder that could function in word mode (sequential word generation) and phrase mode (generate multiple words). Their encoder used RNNSearch (Bahdanau et al. 2015), and could be set to choose between word or phrase mode for operation. Their approach yielded, on average a 3.45 BLEU point improvement over generic models for Chinese-English translation, using NIST datasets as their test set.

Zhao et al. (2018) proposed a method to incorporate the phrase table as recommendation memory into an NMT system. They present a novel approach to find the target words worthy of recommendation from the phrase table, calculate their recommendation scores and harness them so that NMT systems make better predictions. Given a source sentence and a phrase translation table, they first construct a word recommendation set at each decoding step by using a matching method. Then they calculate a bonus value for each recommendable word which is integrated into the NMT process. They have demonstrate substantial increase in performances of Chinese-English and English-Japanese translation tasks.

Sen et al. (2019) injected parallel phrase pairs in order to translate texts from old to modern English. The authors worked in a low resource setting with a very small corpus of around 2700 parallel sentences, achieving remarkable gains. Their technique of harvesting phrase pairs was relatively simple—they only considered the probability of a target phrase given a source phrase. Sennrich and Haddow (2016) implemented a generalized version of the encoder-decoder model (Bahdanau et al. 2015) and found that supplying linguistic features improves the performance of NMT systems in the case of English-German and English-Romanian NMT systems. They added morphological features, part-of-speech tags, and syntactic dependency labels as input features and observed an improvement of 1.5 BLEU for German → English, 0.6 BLEU for English → German, and 1.0 BLEU for English → Romanian.

3.1 Related literature on Indian language NMT (ILNMT)

A recent monograph (Kunchukuttan and Bhattachatryya 2021) is a compendium of ILSMT and ILNMT experiences for both translation and transliteration tasks. For exmaple, “Shata-Anuvadak”-like (Kunchukuttan et al. 2014a) experience for ILNMT is reported in the monograph which is a precursor to our work described in this article.

Banerjee and Bhattacharyya (2018) present results of their experimentation with morpheme based and BPE based segmentation and find that the morfessor based segmentation works well for distant language pairs. They also propose M-BPE, i.e., using morfessor to perform morpheme segmentation and then apply BPE on the morphemes as text units. The authors try this technique on three language pairs: English-Hindi, Bengali-Hindi and English-Bengali. They find this combination of morpheme

Table 2 ISO 639-1 language codes

Language	Hindi	Punjabi	Bengali	Gujarati	Marathi	Tamil	Telugu	Malayalam
Code	hi	pa	bn	gu	mr	ta	te	ml

Table 3 Results of the baseline SMT system

	hi	pa	bn	gu	mr	ta	te	ml
hi	–	70.06	36.31	53.29	33.78	11.36	21.59	10.95
pa	71.26	–	30.27	46.24	25.54	8.96	17.92	7.49
bn	36.16	31.84	–	31.24	19.79	8.88	13.18	8.62
gu	53.09	47.6	29.35	–	26.99	9.95	16.57	7.97
mr	41.66	34.75	23.68	33.84	–	8.34	12.02	7.25
ta	21.79	19.32	14.77	17.28	11.1	–	9.3	6.41
te	27.2	25.14	16.87	22.22	13.47	7.29	–	6.58
ml	14.5	25.14	10.01	10.99	7.01	4.67	6.25	–

The language codes are as follows: Hindi (hi), Punjabi (pa), Bengali (bn), Gujarati (gu), Marathi (mr), Tamil (ta), Telugu (te) and Malayalam (ml)

segmentation and BPE segmentation on the morphemes to surpass performance of both forms of segmentation, individually.

Dabre et al. (2020) is a survey of multilingual NMT (MNMT). In their discussions on low resource NMT, they touch upon ILNMT too. However, this discussion forms a small part of the general perspective on MNMT.

Murthy et al. (2019) show in the context of ILNMT involving 5 Indian languages—Bengali, Gujarati, Marathi, Malayalam and Tamil—that divergent word order adversely limits the benefits from transfer learning when little to no parallel corpus between the source and target language is available. To bridge this divergence, the authors pre-order the assisting language sentence to match the word order of the source language and train the parent model. Their experiments establish the efficacy of this method.

Revanuru et al. (2017) show impressive performance on Urdu-Hindi, Punjabi-Hindi and Gujarati-Hindi pairs even though the neural net is relatively shallow. Similar kind of pair wise ILNMT effort is seen in Akella et al. (2020) which shows that the performance of translation models can be significantly improved by using back-translation through a filtered back-translation process and subsequent fine-tuning on the limited pair-wise language corpora. The languages considered are Hindi, Urdu, Gujarati, Marathi, Punjabi, Odia, Tamil and Malayalam.

4 SMT baseline

Kunchukuttan et al. (2014a) presents extensive work in SMT for Indian languages, building benchmark Phrase Based SMT systems and systems with post-editing for transliteration between 110 language pairs (English and 10 Indian languages).

Phrase based SMT systems perform demonstrably well for low resource language pairs, and are often used as a benchmark for comparison with NMT systems. We will be following suit, and from this point onward, we will treat the results presented in Kunchukuttan et al. (2014a) as the baseline SMT system (Tables 2, 3).

5 Investigating the effect of BPE merge operations: our baseline NMT systems

This section details our experiments with varying BPE merge operations in an endeavor to find the trend for our languages of interest. We designate BPE merge operation ranges as low (0 to 5000 merges), mid (5000 to 20,000 merges) and high (above 20,000). We chose 7 values to represent the low [0, 2500 (2.5k), 5000 (5k)] and mid [7500 (7.5k), 10000 (10k), 15000 (15k), 20,000 (20k)] ranges of merge operations.

5.1 Experimental setup

In this section we present the dataset that was used in our experiments, the architecture of our systems and the metric used to evaluate our models.

5.1.1 Dataset

For experiments between Indian languages we have utilized Indian languages corpora initiative (ILCI) Phase 1 corpus (Jha 2010), which is parallel across 11 languages (English and 10 Indian languages) and contains sentences from health and tourism domains. The corpus was pre-processed to solve issues related to incorrect characters, redundant Unicode representation of some Indic characters using tools from Indic NLP library[5] (Kunchukuttan et al. 2014b). For every language pair, the corpus was split up as follows: training set of 46,277 sentences, test set of 2000 sentences and tuning set of 500 sentences. The train, test and tune splits were completely parallel across all languages involved. The languages considered are: Hindi (hi), Punjabi (pa), Bengali (bn), Gujarati (gu), Marathi (mr), Tamil (ta), Telugu (te) and Malayalam (ml). The former 4 languages belong to the Indo-Aryan family, whereas the latter 3 belong to the Dravidian family.

5.1.2 Evaluation metric

We evaluate our models using the standard BLEU score metric (Papineni et al. 2002). We report the tokenized BLEU score as computed by the multi-bleu.pl script, downloaded from the public implementation of Moses[6].

[5] https://github.com/anoopkunchukuttan/indic_nlp_library.
[6] https://github.com/moses-smt/mosesdecoder.

5.1.3 Training details

Our NMT systems were constructed using OpenNMT-py with the following configuration for the Indian language pairs: The model architecture used was a Bidirectional RNN Encoder-Decoder model with attention. The choice of gated unit was LSTM (Hochreiter and Schmidhuber 1997). The number of layers in the encoder and decoder were 3. The size of the RNN was 500 units. This configuration was chosen after constructing models with 2, 3 and 4 layers between 3 language pairs, namely Hindi-Gujarati, Hindi-Telugu and Telugu-Tamil. These pairs were chosen to represent the cases of translation among and between Indo-Aryan and Dravidian families. This choice was motivated by the analysis presented in Kunchukuttan et al. (2014a); they demonstrated that when morphologically richer languages (such as Dravidian languages) are involved, the translation model entropy is higher. We noted the change in validation perplexity and the final BLEU score over 1,50,000 training steps. 3 layers gave the best performance for both validation perplexity and validation accuracy for our dataset.

The optimizer used was Stochastic Gradient Descent, with an initial learning rate of 1, and batch-size of 1024. During training the initial 8000 steps were for warm-up followed by 1,50,000 training steps.

5.2 Results

We present the comprehensive list of results in this section. Improv./Degrad. over SMT denotes difference between best NMT system and baseline SMT system.

Tables 4, 5, 6 and 7 show the BLEU scores for BiLSTM models with models that translate text at word level and BPE segmented text with merge operations ranging from 0 (character level) to 20,000. Empirically, we observe that plot of translation quality against the number of merge operations behaves similarly across the language pairs considered. The best performance emerges at levels of low BPE merge operations. Word-level translation performs most poorly. This can be attributed to data sparsity. On comparing these BLEU scores with those presented in Table 2 of Kunchukuttan et al. (2014a), we observe that NMT systems (trained on low BPE merge operations) surpass their SMT systems when we translate between Dravidian languages.

5.3 Discussion

In this section, we examine our observations both quantitatively and qualitatively. Figures 2a–d represent the optimal number of merge operations across pairs of Indian languages for Baseline NMT systems. For example, in Fig. 2a, 2500 merge operations is optimal for 13 Indian language pairs. Overall, 2500 emerges as the optimal choice for 23 language pairs and 5000 for 1 language pair. We observe that when translating between Indo-Aryan languages, 2500 often emerges as the best choice. On the other hand, when translating between Indo-Aryan and Dravidian languages (in either direction), character level (that means 0 merge operations) appears

Table 4 Effect of varying BPE merges on BLEU scores for BPE based inter Indo-Aryan NMT systems

	0k	2.5k	5k	7.5k	10k	15k	20k	Word level	I/D Over SMT
hi-pa	62.79	60.77	59.95	59.64	59.17	57.68	57.24	51.34	7.27 ↓
hi-bn	28.51	28.75	28.16	27.59	26.48	24.63	23.6	23.1	7.56 ↓
hi-gu	49.47	52.17	50.90	50.25	49.61	46.94	44.68	39.44	1.12 ↓
hi-mr	31.21	31.66	31.33	29.74	29.77	28.14	25.4	23.41	2.12 ↓
pa-hi	67.76	64.67	70.9	70.59	69.91	68.9	67.95	61.2	0.36 ↓
pa-bn	25.44	25.32	24.22	23.67	22.79	21.31	21.07	18.38	4.83 ↓
pa-gu	43.66	44.74	44.69	43.97	41.94	40.42	39.01	35.1	1.50 ↓
pa-mr	26.78	27.78	26.03	26.13	25.4	23.29	22.75	18.91	2.24 ↑
bn-hi	32.07	31.79	30.97	30.86	29.49	28.54	26.17	23.47	4.09 ↓
bn-pa	27.61	26.96	26.15	25.47	24.72	23.97	22.49	21.41	4.23 ↓
bn-gu	25.82	24.82	24.33	23.77	23.19	22.37	21.07	19.43	5.42 ↓
bn-mr	16.61	16.61	15.83	15.41	14.58	13.76	13.37	11.12	3.18 ↓
gu-hi	52.96	55.02	54.52	53.53	52.98	50.88	49.79	42.27	1.93 ↑
gu-pa	45.22	46.48	44.41	44.37	43.8	40.87	41.52	41.31	1.12 ↓
gu-bn	25.12	25.33	24.17	23.24	22.69	21.27	19.59	16.32	4.02 ↓
gu-mr	25.38	25.62	25.47	24.53	23.24	22.75	21.11	19.87	1.37 ↓
mr-hi	42.23	42.97	42.8	40.71	40.16	38.25	36.12	26.58	1.31 ↑
mr-pa	36.46	37.08	35.29	34.89	34.45	32.31	30.31	23.21	2.33 ↑
mr-bn	21.98	21.82	21.12	20.48	19.87	17.42	16.59	13.24	1.70 ↓
mr-gu	33.19	33.29	31.69	30.56	29.46	28.54	26.67	21.19	0.55 ↓

↑ represents improvement over SMT and ↓ represents degradation with respect to SMT. I/D indicates improvement/degradation

Table 5 Effect of varying BPE merges on BLEU scores for BPE based Indo-Aryan to Dravidian NMT systems

	0k	2.5k	5k	7.5k	10k	15k	20k	Word level	I/D Over SMT
hi-ta	12.86	13.78	13.01	12.33	11.64	10.95	9.68	8.8	2.42 ↑
hi-te	19.18	19.03	18.83	18.49	17.87	17.29	16.58	13.62	2.41 ↓
hi-ml	10.4	10.25	9.9	9.35	9.37	8.3	7.56	6.11	0.55 ↓
pa-ta	11.86	12.4	11.36	9.95	9.31	9.22	8.06	6.51	3.44 ↑
pa-te	16.99	16.83	16.79	16.33	15.75	14.88	13.9	11.79	0.93 ↓
pa-ml	9.87	8.04	7.84	7.06	7.22	5.94	5.75	4.56	2.38 ↑
bn-ta	9.22	9.52	9.48	8.81	8.18	7.2	7.12	5.92	0.64 ↑
bn-te	11.71	11.63	11.08	10.71	10.4	9.06	8.57	7.42	1.47 ↓
bn-ml	8.06	8.12	7.74	7.02	6.89	5.85	5.59	4.3	0.50 ↓
gu-ta	11.59	11.66	11.46	10.8	9.95	7.95	8.03	6.08	1.64 ↑
gu-te	16.5	16.62	15.72	15.82	15.03	14.4	12.69	10.12	0.05 ↑
gu-ml	9.45	8.71	8.17	8.2	7.79	6.03	5.57	4.16	1.48 ↑
mr-ta	10.01	9.93	9.59	9.17	8.49	7.22	6.76	5.38	1.67 ↑
mr-te	13.98	13.89	13.15	12.86	12.41	11.36	10.07	7.49	1.96 ↑
mr-ml	8.73	8.73	6.67	7.11	5.52	4.22	3.29	3.65	1.48 ↑

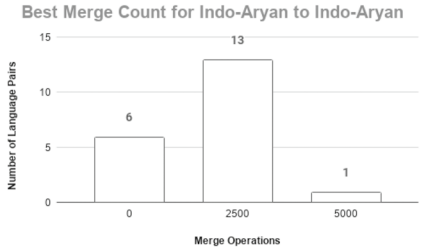
Table 6 Effect of varying BPE merges on BLEU scores for BPE based Dravidian to Indo-Aryan NMT systems

	0k	2.5k	5k	7.5k	10k	15k	20k	Word level	I/D Over SMT
ta-hi	21.75	20.6	20.05	18.99	17.72	15.44	14.78	10.64	0.04 ↓
ta-pa	20.11	18.18	16.99	15.58	14.37	12.96	12.57	10.79	0.79 ↑
ta-bn	12.77	12.62	11.73	10.41	10.08	8.78	8.16	7.3	2.00 ↓
ta-gu	17.22	16.08	15.16	13.45	12.64	11.62	11	8.92	0.06 ↓
ta-mr	10.97	9.33	9.37	7.9	7.72	6.26	5.39	4.92	0.13 ↓
te-hi	30.46	31.01	29.81	28.43	28.1	26.27	24.82	19.67	3.81 ↑
te-pa	26.93	26.38	25.69	25.08	24.03	22.59	20.97	15.97	1.79 ↑
te-bn	16.9	16.98	15.82	14.91	15.12	13.35	12.07	9.82	0.11 ↑
te-gu	24.1	22.76	22.65	21.33	20.59	19.82	18.04	14.26	1.88 ↑
te-mr	15.25	14.78	14.67	13.89	13.29	11.81	10.63	7.67	1.78 ↑
ml-hi	19.42	17.55	16.2	14.91	14.13	11.81	11.34	7.27	4.92 ↑
ml-pa	16.89	15.38	12.42	11.82	10.42	9.08	7.86	6.68	8.25 ↓
ml-bn	12.02	11.19	10.38	8.78	8.2	7.01	6.16	5.07	2.01 ↑
ml-gu	15.05	12.39	12.38	11.4	8.6	8.18	7.25	5.87	4.06 ↑
ml-mr	9.96	8.09	6.06	6.96	5.99	4.19	4.03	3.53	2.95 ↑

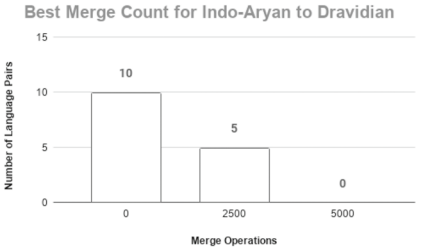
Table 7 Effect of varying BPE merges on BLEU scores for BPE based Dravidian to Dravidian NMT systems

	0k	2.5k	5k	7.5k	10k	15k	20k	Word level	I/D Over SMT
ta-te	9.65	9.21	8.32	7.85	7.51	7	6	4.9	0.35 ↑
ta-ml	6.55	7.26	6.52	5.93	5.79	5.03	4.76	4.61	0.85 ↑
te-ta	10.33	10.26	9.74	9.05	8.96	7.69	7.03	5.37	3.04 ↑
te-ml	7.96	8.48	7.87	7.11	6.72	5.9	5.02	3.7	1.90 ↑
ml-ta	7.43	7.54	6.93	6.08	5.86	4.93	4.66	4.22	2.87 ↑
ml-te	7.95	7.09	7.04	6.34	5.65	5.17	4.36	3.49	1.70 ↑

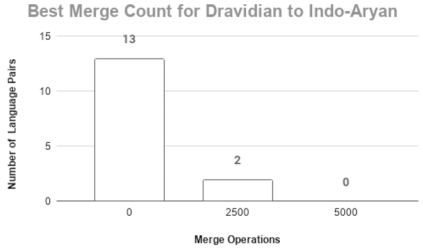
to be a better choice (indicated by 32 language pairs). This leads us to believe language relatedness and morphology play a role in influencing our hyperparameter of interest. It is accepted that subwording helps NMT systems combat data sparsity, a problem which is compounded in the case of morphologically richer languages (Kunchukuttan et al. 2014a). When Dravidian languages are involved, their morphological richness calls for segmentation. Thus we see NMT systems involving Dravidian languages benefiting from subwording—in 26 out of 36 models, we see an improvement in BLEU score (ranging from +0.35 to +4.92). In systems that translate from Indo-Aryan to Dravidian languages, the improvement over SMT ranges from +0.05 to +3.44. In 5 out the 10 cases where the NMT system scores lesser than the SMT system, its performance comes close (the difference is less than one BLEU point). We also note that score drops are particularly drastic when Punjabi and Bengali are involved.



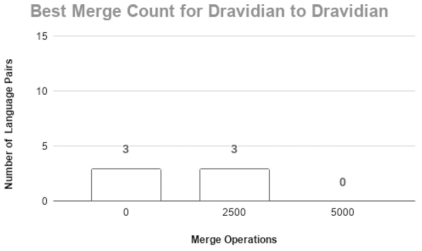
(a) Distribution of optimal merge operations for inter Indo-Aryan language systems



(b) Distribution of optimal merge operations for Indo-Aryan to Dravidian language systems



(c) Distribution of optimal merge operations for Dravidian to Indo-Aryan language systems



(d) Distribution of optimal merge operations for inter Dravidian language language systems

Fig. 2 Distribution of optimal merge operations across Indian language pairs

5.4 Qualitative analysis

In this section, we present some of the actual outputs produced by our systems on a random sample of the test set, considering the language pair Telugu-Hindi (te-hi). Source (Src) sentence refers to the input given to the system and reference (ref) refers to the corresponding test sentence from the parallel test corpus. We compare the outputs of SMT, Baseline NMT System (abbreviated as BS) at 2500 merges (BS-2.5k) and at 7500 merges (BS-7.5k). The objective of this comparison is to demonstrate the effect of different merge operations on the system output. For the convenience of the reader, Indian language output has been represented in different forms, denoted by these suffixes: system's output—the original text (OP), transliterated (TL), word to word translation in English (WW) and English Translation (ET).

Te-Src: ఉండే ఏర్పాటు ప్రభుత్వ విశ్రాంతి క్యాంప్ మరియు కారలా హోటల్లో వుంది .
Hi-Ref: ठहरने की व्यवस्था सरकारी अवकाशीय कैंप व कारलाहोटल में है।
SMT-OP: रहने की व्यवस्था सरकारी वशि्राम कैंप व कारला होटल में स्थति है।
SMT-TL: rahane kee vyavastha sarakaaree vishraam kaimp va kaarala hotal mein sthit hai .
SMT-WW: Living is arrangement government Rest Camp and Carla Hotel located is.

SMT-ET: The accommodation is located in the Government Rest Camp and Carla Hotel.
BS-2.5k-OP: ठहरने की व्यवस्था सरकारी वशि्राम कैंप तथा कारला होटल में है।
BS-2.5k-TL: thaharane kee vyavastha sarakaaree vishraam kaimp tatha kaarala hotal mein hai .
BS-2.5k-WW: Accommodation is arrangement government rest camp and Carla Hotel is in.
BS-2.5k-ET: Accommodation is available at the government rest camp and Karla Hotel.
BS-7.5k-OP: ठहरने का प्रबंध सरकारी वशि्राक और कार होटल में है।
BS-7.5k-TL: thaharane ka prabandh sarakaaree vishraak aur kaar hotal mein hai .
BS-7.5k-WW: Accommodation is arrangement government derelict and car hotel is in.
BS-7.5k-ET: The stay is arranged in government accommodation and car hotel.

6 Phrase table injection: our phrase table augmented NMT systems

This is a technique that combines SMT with NMT. It leverages the phrase table generated during building a phrase-based SMT (PBSMT) system to aid NMT. The phrase table is a key data structure used in translation in an SMT system. It contains the mapping of phrases (groups of words, not necessarily linguistic phrases) between the source and the target languages, along with four phrase translation scores. Given that we are translating from language f to e, these scores are inverse phrase translation probability ($\phi(f|e)$), inverse lexical weighting ($lex(f|e)$), direct phrase translation probability ($\phi(e|f)$) and direct lexical weighting ($lex(e|f)$). The phrase table is used as an additional data source and is constructed from the same parallel corpus that is used for NMT training.

6.1 Dataset and training details

The dataset used is the same as described in Sect. 5.1.1. The PBSMT systems were trained using Moses[7] (Koehn et al. 2007). The grow-diag-final-and heuristic was used for extracting phrases and the msd-bidirectional-fe model was used for lexicalized reordering. Tuning was done by Minimum Error Rate Training (MERT) with default parameters (100 best list, max 25 iterations). 5-gram language models were constructed on the corpus using the Kneser-Ney smoothing algorithm via SRILM.

The following selection criterion was used to extract phrases from the phrase table: the weighted average of translation and lexical probabilities mentioned in the phrase's entry must be higher than mean+std_dev of all the weighted probabilities calculated for all the phrases in the phrase table. The weights for these probabilities can be found in the moses.ini file, at the Translation Model component.

[7] https://github.com/moses-smt/mosesdecoder.

Table 8 BLEU scores for inter Indo-Aryan phrase table injected NMT systems

	0k	2.5k	5k	I/D over NMT	I/D over SMT
hi-pa	62.39	52.82	62.97	0.18 ↑	7.09 ↓
hi-bn	29.4	30.65	30.94	1.19 ↑	5.37 ↓
hi-gu	49.85	52.52	51.76	0.35 ↑	0.77 ↓
hi-mr	31.15	33.36	32.82	1.70 ↑	0.42 ↓
pa-hi	67.48	57.57	65.04	3.41 ↓	3.78 ↓
pa-bn	25.91	27.48	27	2.04 ↑	2.79 ↓
pa-gu	44.39	46.5	45.23	1.76 ↑	0.26 ↑
pa-mr	27.22	28.82	27.93	1.04 ↑	3.28 ↑
bn-hi	32.28	33.82	33.17	1.75 ↑	2.34 ↓
bn-pa	28.35	28.61	28.46	1.00 ↑	3.23 ↓
bn-gu	26.85	27.21	26.25	1.39↑	4.03 ↓
bn-mr	17.02	18.37	17.72	1.76 ↑	1.42 ↓
gu-bn	26.03	26.57	26.02	0.48 ↓	1.45 ↑
gu-pa	44.9	46.25	45.88	0.23 ↓	1.35 ↓
gu-hi	52.43	54.54	54.27	1.24 ↑	2.78 ↓
gu-mr	25.88	27.14	26.6	1.52 ↑	0.15 ↓
mr-hi	42.84	44.65	44.11	1.68 ↑	2.99 ↑
mr-pa	36.62	37.45	36.83	0.37 ↑	2.70 ↑
mr-bn	22.34	23.44	22.63	1.46 ↑	0.24 ↓
mr-gu	33.71	34.6	33.33	1.31 ↑	0.76 ↑

The architecture for the NMT systems is the same as that described in Sect. 5.1.3. While training the systems, both sources of data, namely the parallel sentences of the corpus and the extracted phrase pairs were given equal weightage. Model checkpoints are created at every 15,000 steps. BLEU scores were used for evaluation.

6.2 Results

Building on the insights obtained in Sect. 5.2, these experiments were performed on the low range of merge operations. Systems were built for character level (0k), 2500 (2.5k) and 5000 (5k) merge operations. I/D over NMT represents the difference between the BLEU score of the best phrase augmented system and the best baseline NMT system; similarly I/D over SMT represents the difference between the BLEU score of the best phrase augmented system and the baseline SMT system.

The comprehensive list of results is presented in Tables 8, 9, 10 and 11.

6.3 Discussion

Figure 3a–d represent the optimal number of merge operations across pairs of Indian languages for PTI systems.

Table 9 BLEU scores for Indo-Aryan to Dravidian phrase table injected NMT systems

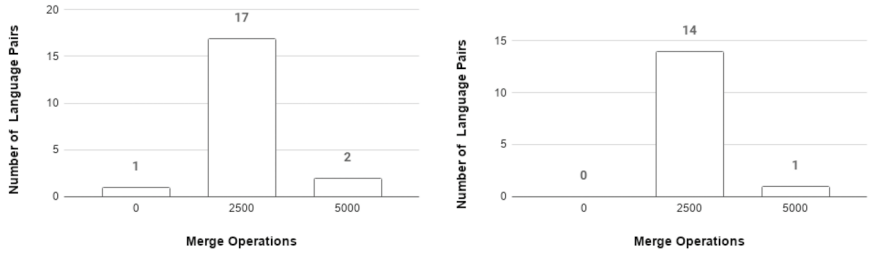
	0k	2.5k	5k	I/D over NMT	I/D over SMT
hi-ta	13.45	15.73	15.09	1.95 ↑	4.37 ↑
hi-te	17.27	17.76	17.58	1.42 ↓	3.83 ↓
hi-ml	10.74	12.19	11.75	1.79 ↑	1.24 ↑
pa-ta	12.28	13.66	13.22	1.26 ↑	4.70 ↑
pa-te	15.23	15.69	16.25	0.74 ↓	1.67 ↓
pa-ml	9.93	10.98	10.48	1.11 ↑	3.49 ↑
bn-ta	9.71	11.4	10.78	1.88 ↑	2.52 ↑
bn-te	12.27	12.27	11.97	1.34 ↓	1.29 ↓
bn-ml	8.52	9.41	8.8	0.25 ↑	0.91 ↓
gu-ta	11.94	13.52	12.95	1.86 ↑	3.57 ↑
gu-te	15.13	15.28	14.84	1.34 ↓	1.29 ↓
gu-ml	10.12	11.17	10.33	1.72 ↑	3.20 ↑
mr-ta	10.39	12.03	11.63	2.02 ↑	3.69 ↑
mr-te	11.99	12.89	12.36	1.09 ↓	0.87 ↑
mr-ml	9.2	9.87	9.52	1.14 ↑	2.62 ↑

Table 10 BLEU scores for Dravidian to Indo-Aryan phrase table injected NMT systems

	0k	2.5k	5k	I/D over NMT	I/D over SMT
ta-hi	22.76	23.65	22.93	1.90 ↑	1.86 ↑
ta-pa	19.97	20.59	19.91	0.48 ↑	1.27 ↑
ta-bn	13.33	14.12	14.55	1.78 ↑	0.22 ↓
ta-gu	17.31	18.26	17.61	1.04 ↑	0.98 ↑
ta-mr	11.08	11.73	11.46	0.76 ↑	0.63 ↑
te-hi	30.66	32.16	31.58	1.15 ↑	4.96 ↑
te-pa	26.77	27.78	27.96	1.03 ↑	2.82 ↑
te-bn	17.68	17.75	17.59	0.77 ↑	0.88 ↑
te-gu	24.61	25.13	24.15	1.03 ↑	2.91 ↑
te-mr	15.42	16.49	15.79	1.24 ↑	3.02 ↑
ml-hi	19.46	20.46	19.35	1.04 ↑	5.96 ↑
ml-pa	17.45	17.41	16.63	0.56 ↑	4.92 ↑
ml-bn	12.23	12.69	11.85	0.67 ↑	2.68 ↑
ml-gu	14.86	15	14.29	0.05 ↓	4.01 ↑
ml-mr	10.34	11.18	10.41	1.22 ↑	4.17 ↑

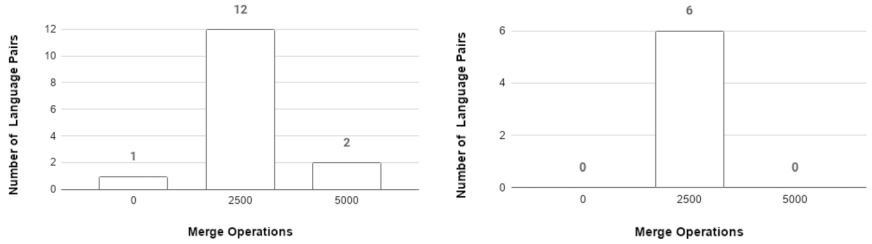
Table 11 BLEU scores for inter Dravidian phrase table injected NMT system

	0k	2.5k	5k	I/D over NMT	I/D over SMT
ta-te	9.88	10.31	9.42	0.66 ↑	1.01 ↑
ta-ml	7.23	8.03	7.91	0.77 ↑	1.62 ↑
te-ta	10.83	11.42	11.33	1.09 ↑	4.13 ↑
te-ml	8.44	9.48	8.86	1.00 ↑	2.90 ↑
ml-ta	7.42	8.23	7.98	0.69 ↑	3.56 ↑
ml-te	7.57	8.31	7.45	0.36 ↑	2.06 ↑



(a) Distribution of optimal merge operations for inter Indo-Aryan language Phrase Augmented systems

(b) Distribution of optimal merge operations for Indo-Aryan to Dravidian language Phrase Augmented systems



(c) Distribution of optimal merge operations for Dravidian to Indo-Aryan language Phrase Augmented systems

(d) Distribution of optimal merge operations for inter Dravidian language language Phrase Augmented systems

Fig. 3 Distribution of optimal merge operations for PTI systems across Indian languages

- This technique yields an improvement over baseline NMT for 41 (out of 50) language pairs where Indo-Aryan languages are involved. For Dravidian languages, it yields an improvement for all 6 pairs. Phrase augmentation shows strong promise when Dravidian languages are at the source.
- The improvement over baseline NMT, in terms of BLEU score, gains were seen for 17 out of 20 language pairs for inter Indo-Aryan systems (ranging from +0.18 to +2.04), Indo-Aryan to Dravidian NMT systems (ranging from +0.25 to +2.02), 14 out of 15 language pairs in Dravidian to Indo-Aryan NMT systems (ranging from +0.48 to +1.9), 6 out of 6 language pairs for inter Dravidian (ranging from +0.36 to +1.09).
- As for improvement over SMT, in terms of BLEU score, gains were seen for 7 out of 20 language pairs for inter Indo-Aryan systems (ranging from +0.15 to +3.28), 10 out of 15 language pairs in Indo-Aryan to Dravidian NMT systems (ranging from +0.87 to +4.37), 14 out of 15 language pairs in Dravidian to Indo-Aryan NMT systems (ranging from +0.88 to +5.96), 6 out of 6 language pairs for inter Dravidian (ranging from +1.01 to +4.13).
- For most systems, 2500 merges yields the best BLEU score, contrary to what was observed in Sect. 5.3. The phrase augmented systems outperform baseline NMT systems for 47 out of 56 language pairs, indicating this is a technique that has good potential of boosting a baseline NMT system.

– We also observe that among the chosen Indo-Aryan languages, systems that have Marathi on the source side benefit the most from phrase injection with 6 out of 7 systems showing an improvement over both SMT and baseline NMT (mr-bn does not improve over SMT and mr-te does not show an improvement over NMT). This can be attributed to Marathi's relative morphological richness compared to other members of the Indo-Aryan family.
– This paradigm consistently performs poorly when the source language is an Indo-Aryan and the target language is Telugu.

7 Morpheme segmented BPE with word features: our enhanced BPE-based NMT models

In this section, we cover NMT models that are enhanced by incorporating features such as words and morphemes.

7.1 Dataset and training details

We utilized the ILCI corpus, mentioned in Sect. 5.1.1 for the following experiments. The architecture of models follows Sect. 5.1.3.

7.2 Morpheme segmentation before BPE

After exploring BPE results with various merge operations on different language pairs, we decided to enhance our model and explore the effect of performing morpheme segmentation on the data before applying BPE. We used unsupervised morpheme segmentation tool from Indic NLP library to segment the data and then used the value of BPE merge operation on which peak (best BLEU score) was observed. We did this experiment for all language pairs as discussed above. This model will be referred to as *morph-seg* from this point onward.

7.2.1 Results

We observed that whenever the peak was at 2.5k merge operations, then applying morph segmentation before BPE, actually lead to higher BLEU score (better results) and similarly, whenever peak was at 0k merge operation, BLEU score was lower.

I/D over NMT represents the difference between BLEU score of the best morph-seg system and the best baseline NMT system; similarly I/D over SMT represents the difference between BLEU score of the best morph-seg system and the baseline SMT system. The comprehensive list of results is presented in Tables 12, 13, 14 and 15.

Table 12 BLEU scores for inter Dravidian morph-seg (Sect. 7.2) NMT systems

	BLEU score	Optimal merges	I/D over NMT	I/D over SMT
ta-te	9.70	0	0.05 ↑	0.40 ↑
ta-ml	7.46	2500	0.20 ↑	1.05 ↑
te-ta	9.44	0	0.89 ↓	2.15 ↑
te-ml	8.33	2.5	0.15 ↓	1.75 ↑
ml-ta	7.64	2500	0.10 ↑	2.97 ↑
ml-te	8.44	0	0.49 ↑	2.19 ↑

Table 13 BLEU scores for inter Indo-Aryan morph-seg (Sect. 7.2) NMT systems

	BLEU score	Optimal merges	I/D over NMT	I/D over SMT
hi-pa	58.72	0	4.07 ↓	11.34 ↓
hi-bn	30.12	2500	1.37 ↑	6.19 ↓
hi-gu	52.51	2500	0.34 ↑	0.78 ↓
hi-mr	33	2500	1.34 ↑	0.78 ↓
pa-hi	71.07	5000	0.17 ↑	0.19 ↓
pa-bn	24.17	0	1.27 ↓	6.10 ↓
pa-gu	45.87	2500	1.06 ↑	0.37 ↓
pa-mr	28.32	2500	0.54 ↑	2.78 ↑
bn-hi	31.07	0	1.00 ↓	4.99 ↓
bn-pa	26.11	0	1.50 ↓	5.73 ↓
bn-gu	24.54	0	1.28 ↓	6.70 ↓
bn-mr	15.95	0	0.66 ↓	3.84 ↓
gu-bn	26.23	2500	0.90 ↑	3.12 ↓
gu-pa	46.65	2500	0.17 ↑	0.95 ↓
gu-hi	55.47	2500	0.45 ↑	2.38 ↑
gu-mr	26.44	2500	0.82 ↑	0.55 ↓
mr-hi	44.07	2500	1.10 ↑	2.41 ↑
mr-pa	36.81	2500	0.27 ↓	2.06 ↑
mr-bn	20.96	0	1.02 ↓	2.72 ↓
mr-gu	34.23	2500	0.94 ↑	0.39 ↑

7.3 Including word feature in morph segmented BPE

As we applied the technique of morpheme segmentation prior to BPE, we suspected that BPE could lead to loss of context due to word segmentation as smaller word segments might have lesser context when compared to whole word.

This lead us to augment our BPE model by adding word features to it. In the reported experiments, this is done by applying BPE on morph segmented data with 2500 (2.5k) merge operations and then merging the whole word embedding to it. In the interest of computational power and time, we chose to proceed with only 2.5k merge operations,

Table 14 BLEU scores for Indo-Aryan and Dravidian morph-seg (Sect. 7.2) NMT systems

	BLEU Score	Optimal merges	I/D over NMT	I/D over SMT
hi-ta	14.25	2500	0.47 ↑	2.89 ↑
hi-te	17.8	0	1.38 ↓	3.79 ↓
hi-ml	9.63	0	0.77 ↓	1.32 ↓
pa-ta	12.58	2500	0.18 ↑	3.62 ↑
pa-te	16.24	0	0.75 ↓	1.68 ↓
pa-ml	8.71	0	1.16 ↓	1.22 ↑
bn-ta	10.19	2500	0.67 ↑	1.31 ↑
bn-te	12.18	0	0.47 ↑	1.00 ↓
bn-ml	8.46	2500	0.34 ↑	0.16 ↓
gu-ta	10.68	0	0.91 ↓	0.73 ↑
gu-te	17.01	2500	0.39 ↑	0.44 ↑
gu-ml	8.86	0	0.59 ↓	0.89 ↑
mr-ta	9.24	0	0.77 ↓	0.90 ↑
mr-te	13.51	0	0.47 ↓	1.49 ↑
mr-ml	7.76	0	0.97 ↓	0.51 ↑

Table 15 BLEU scores for Dravidian to Indo-Aryan morph-seg (Sect. 7.2) NMT systems

	BLEU score	Optimal merges	I/D over NMT	I/D over SMT
ta-hi	21.09	0	0.66 ↓	0.70 ↓
ta-pa	19.42	0	0.69 ↓	0.10 ↑
ta-bn	12.19	0	0.58 ↓	2.58 ↓
ta-gu	15.76	0	1.46 ↓	1.52 ↓
ta-mr	9.92	0	1.05 ↓	1.18 ↓
te-hi	31.13	2500	0.12 ↑	3.93 ↑
te-pa	24.91	0	2.02 ↓	0.23 ↓
te-bn	15.43	2500	1.55 ↓	1.44 ↓
te-gu	22.58	0	1.52 ↓	0.36 ↑
te-mr	14.39	0	0.86 ↓	0.92 ↑
ml-hi	18.38	0	1.04 ↓	3.88 ↑
ml-pa	15.68	0	1.21 ↓	9.46 ↓
ml-bn	10.99	0	1.03 ↓	0.98 ↑
ml-gu	13.68	0	1.37 ↓	2.69 ↑
ml-mr	9.39	0	0.57 ↓	2.38 ↑

since attempts with character level models yielded poor results. As established in previous sections (5, 6) of this paper, 2.5k is a fair choice that has been optimal for several language pairs. For merging BPE embedding with the whole word embedding, we used multi-layer perceptron (MLP) technique which is available in OpenNMT pytorch. The sizes of word embedding vector and BPE embedding vector were set to 300 for

Table 16 BLEU scores for inter Indo-Aryan morph-seg-word (Sect. 7.3) NMT systems

	BLEU score	I/D over NMT*	I/D over SMT
hi-pa	61.02	0.25 ↑	9.04 ↓
hi-bn	31.20	2.45 ↑	5.11 ↓
hi-gu	52.95	0.78 ↑	0.34 ↓
hi-mr	33.11	1.45 ↑	0.67 ↓
pa-hi	68.51	3.84 ↑	2.75 ↓
pa-bn	26.62	1.30 ↑	3.65 ↓
pa-gu	46.52	1.78 ↑	0.28 ↑
pa-mr	28.41	0.63 ↑	2.87 ↑
bn-hi	33.65	1.86 ↑	2.51 ↓
bn-pa	28.58	1.62 ↑	3.26 ↓
bn-gu	26.65	1.83 ↑	4.59 ↓
bn-mr	18.53	1.92 ↑	1.26 ↓
gu-bn	26.97	1.64 ↑	3.20 ↑
gu-pa	46.78	0.30 ↑	0.82 ↓
gu-hi	56.29	1.27 ↑	2.38 ↓
gu-mr	26.98	1.36 ↑	0.01 ↓
mr-hi	44.27	1.30 ↑	2.61 ↑
mr-pa	37.09	0.01 ↑	2.34 ↑
mr-bn	22.34	0.52 ↑	1.34 ↓
mr-gu	34.02	0.73 ↑	0.18 ↑

The model is compared to baseline NMT at 2500 merge operations

consistency and enabling fair comparison with previous models. This model will be referred to as morph-seg-word from this point onward.

7.3.1 Results

We observed that adding whole word as a feature to morph segmented BPE, improves the BLEU score consistently over most language pairs as discussed above with some pairs as exception. We believe that inclusion of a word feature helps to ameliorate the problem of context loss due to BPE, which in turn could be the cause of the improvement in BLEU score.

The comprehensive list of results is presented in Tables 16, 17, 18, and 19.

7.4 Discussion

- This technique yields an improvement over baseline NMT for the language pairs where Indo-Aryan languages are involved (50 out of 56 pairs). For Dravidian languages, it yields an improvement for 3 out of 6 pairs.
- For morph-seg model, as for improvement over SMT, in terms of BLEU scores, gains were seen for for 10 out of 15 language pairs in Indo-Aryan to Dravidian NMT systems (ranging from +0.44 to +3.62), 8 out of 15 language pairs in

Table 17 BLEU scores for Indo-Aryan to Dravidian morph-seg-word (Sect. 7.3) NMT systems

	BLEU score	I/D over NMT*	I/D over SMT
hi-ta	14.32	0.54 ↑	2.96 ↑
hi-te	20.03	1.00 ↑	1.56 ↓
hi-ml	10.45	0.20 ↑	0.50 ↓
pa-ta	12.91	0.51 ↑	3.95 ↑
pa-te	18.05	1.22 ↑	0.13 ↑
pa-ml	10.14	2.10 ↑	2.65 ↑
bn-ta	9.96	0.44 ↑	1.08 ↑
bn-te	13.22	1.59 ↑	0.04 ↑
bn-ml	8.99	0.87 ↑	0.37 ↑
gu-ta	12.22	0.56 ↑	2.27 ↑
gu-te	17.52	0.90 ↑	0.95 ↑
gu-ml	10.27	1.56 ↑	2.30 ↑
mr-ta	10.53	0.60 ↑	2.19 ↑
mr-te	14.60	0.71 ↑	2.58 ↑
mr-ml	9	0.27 ↑	1.75 ↑

The model is compared to baseline NMT at 2500 merge operations

Table 18 BLEU scores for Dravidian and Indo-Aryan morph-seg-word (Sect. 7.3) NMt systems

	BLEU score	I/D over NMT*	I/D over SMT
ta-hi	21.51	0.91 ↑	0.28 ↓
ta-pa	18.99	0.81 ↑	0.33 ↓
ta-bn	12.77	0.15 ↑	2.00 ↓
ta-gu	16.52	0.44 ↑	0.76 ↓
ta-mr	11.19	1.86 ↑	0.09 ↑
te-hi	31.78	0.77 ↑	4.58 ↑
te-pa	27.62	1.24 ↑	2.48 ↑
te-bn	17.1	0.12 ↑	0.23 ↑
te-gu	24.09	1.33 ↑	1.87 ↑
te-mr	16.05	1.27 ↑	2.58 ↑
ml-hi	18.61	1.06 ↑	4.11 ↑
ml-pa	15.96	0.58 ↑	9.18 ↓
ml-bn	11.34	0.15 ↑	1.33 ↑
ml-gu	14.24	1.85 ↑	3.25 ↑
ml-mr	9.58	1.49 ↑	2.57 ↑

The model is compared to baseline NMT at 2500 merge operations

Dravidian to Indo-Aryan NMT systems (ranging from +0.1 to +3.93), 6 out of 6 language pairs for inter Dravidian (ranging from +0.4 to +2.97) and 5 out of 20 language pairs for inter Indo-Aryan systems (ranging from +0.39 to +2.78).

– For morph-seg-word model, as for improvement over SMT, in terms of BLEU scores, gains were seen for for 13 out of 15 language pairs in Indo-Aryan to Dravidian NMT systems (ranging from +0.04 to +3.95), 10 out of 15 language pairs

Table 19 BLEU scores for inter Dravidian morph-seg-word (Sect. 7.3) NMT systems

	BLEU score	I/D over NMT*	I/D over SMT
ta-te	10.85	1.64 ↑	1.55 ↑
ta-ml	7.0	0.26 ↓	0.59 ↑
te-ta	10.51	0.25 ↑	3.22 ↑
te-ml	8.14	0.34 ↓	1.56 ↑
ml-ta	7.37	0.17 ↓	2.70 ↑
ml-te	8.92	1.83 ↑	2.67 ↑

The model is compared to baseline NMT at 2500 merge operations

in Dravidian to Indo-Aryan NMT systems (ranging from +0.09 to +4.58), 6 out of 6 language pairs for inter Dravidian (ranging from +0.59 to +3.22) and 6 out of 20 language pairs for inter Indo-Aryan systems (ranging from +0.18 to +3.2).

8 Qualitative analysis

In this section, we present the actual outputs produced by our systems on a random sample of the test set. We compare the outputs of SMT, Baseline NMT (BS), Phrase Augmented (PAS) and Morph-Seg-Word (MSWS) systems. We have chosen 3 representative language pairs: gu-hi (inter Indo-Aryan), pa-ml (Indo-Aryan to Dravidian) and te-hi (Dravidian to Indo-Aryan) to provide a holistic picture. For the convenience of the reader, Indian language output has been represented in different forms, denoted by these suffixes: system's output—the original text (OP), transliterated (TL), word to word English translation (WW) and proper English Translation (ET).

8.1 Inter Indo-Aryan

For the gu-hi translation system, the morph-seg-word model (Sect. 7.3) gives the best performance. Here we observe a sample output of the system:

Src: કહેવાય છે કે આ વિશાળ ખાડામાં આ ગેસ છેલ્લા ૩૫ વર્ષથી સત ત સળગી રહ્યો છે .
Ref: कहा जाता है कि इस वशिाल खड्डे में ये गसे पछिले 3-5 साल से लगातारजल रही है।
Ref-TL: kaha jaata hai ki is vishaal khadde mein ye gaisen pichhale 3 5 saal se lagaataar jal rahee hai .
Ref-WW: Said it is that this huge pit in these gases last 3 5 years for continuously burning
Ref-ET: It is said that these gases have been burning continuously in this huge pit for the last 3–5 years.
SMT-OP:
कहा जाता है कि इस विशाल गड्ढे में यह गेस पिछले 5,4 साल से लगातार जला रही है।

SMT-TL: kaha jaata hai ki is vishaal gaddhe mein yah gais pichhale 5,4 saal se lagaataar jala rahee hai .
SMT-WW: Said it is that this huge pit in these gases last 5 4 years for continuously burning
SMT-ET: It is said that in this huge pit, it has been continuously burning for the last 5,4 years.
BS-OP: कहा जाता है कि इस वशिाल खाने मे यह गैस पछिले 3 5 साल से नरिंतर जड़ रहे हैं।
BS-TL:kaha jaata hai ki is vishaal khaane mein yah gais pichhale 3 5 saal se nirantar jad rahe hain .
BS-WW: Said it is that this huge food in this gas last 3 5 years for continuously rooted is
BS-ET: It is said that this gas has been continuously rooted in this huge food for the last 3–5 years.
PAS-OP: कहा जाता है कि इस वशिाल खाने मे यह गैस आखिरी 3 5 साल से लगातार जला रहा है।
PAS-TL: kaha jaata hai ki is vishaal khaane mein yah gais aakhiree 3 5 saal se lagaataar jala raha hai .
PAS-WW: Said it is that this huge food in this gas last 3 5 years for continuously burning is
PAS-ET: It is said that in this huge food, this gas has been burning continuously for the last 3–5 years.
MSWS-OP: कहा जाता है कि इस वशिाल गड्ढे मे ये गैस पछिले 3 5 साल से लगातार जला जा रहा है।
MSWS-TL: kaha jaata hai ki is vishaal gaddhe mein ye gais pichhale 3 5 saal se lagaataar jala ja raha hai .
MSWS-WW: Said it is that this huge pit in this gas last 3 5 years for continuously burning is
MSWS-ET: It is said that this gas is being burnt continuously in this huge pit for the last 3–5 years.

The SMT system is unable to fully translate the test sentence. Both BS and PAS are unable to translate 'pit', translating it to 'food' (this can be attributed to the segment mapping that is induced due to BPE segments). The Morph-Seg-Word Model does a perfect translation. Supplying the word feature enables it to avoid mistakes that the baseline and phrase

8.2 Indo-Aryan to Dravidian

For the pa-ml translation systems, NMT performs better than SMT and PAS and MSWS have comparable performance. Here we observe a sample output of the system:

Src: ਗੋਮਤੇਸ਼ਵਰ ਜੀ ਦੀ ੫੭ ਫੁੱਟ ਉੱਚੀ ਮੂਰਤੀ ਵਸ਼ਿਵ ਪ੍ਰਸਾਂਯਿ ਹੈ।
Ref: ഗോമതേശ്വർജിയുടെ 5 7 അടി ഉയരമുള്ള പ്രതിമ വിശ്വപ്രസിദ്ധമാണ്.

Ref-TL: gēāmatēśvarjiyuṭe 5 7 aṭi uyaramuḷḷa pratima viśvaprasid'dhamāṇ .
Ref-WW: Gomateshwrjee's 5 7 foot tall statue is world famous
Ref-ET: The 5 7-foot-tall statue of Gomateshwerjee is famous
SMT: ഗോമതേശ്വര ഭണ്ഡാരിയുടെ 5 7 അടി ഉയരമുള്ള പ്രതിമ ലോക പ്രസിദ്ധമാണ് .
SMT-TL: gōmatēśavara (not translated) bhaṇḍāriyuṭe 5 7 aṭi uyaramuḷḷa pratima lēāka prasid'dhamāṇ .
SMT-WW: gōmatēśavara (not translated) of the treasury's 5 7 feet tall statue is world famous.
SMT-ET: The 5 7 feet tall statue of Gomateshwar of (not translated) the treasury is world famous.
BS: ഗോമതേശ്വർ ജിയുടെ 5 7 അടി ഉയരമുള്ള പ്രതിമ ലോകപ്രസിദ്ധമാണ് .
BS-TL: gēāmatēśvar jiyuṭe 5 7 aṭi uyaramuḷḷa pratima lēākaprasid'dhamāṇ .
BS-WW: Gomateshwar Ji's 5 7 feet tall statue is world famous
BS-ET: The 5 7 feet tall statue of Gomateshwar Ji is world famous.
PAS: ഗോമേതേശ്വർജിയുടെ 5 7 അടി ഉയരമുള്ള പ്രതിമ ലോകപ്രസിദ്ധമാണ് .
PAS-TL: gēāmētēśvarjiyuṭe 5 7 aṭi uyaramuḷḷa pratima lēākaprasid'dhamāṇ .
PAS-WW: Ghometeswarji's 5 7 feet tall statue is world famous
PAS-ET: The 5 7 feet tall statue of Ghometeswarji is world famous.
MSWS: ഗോമതേശ്വർജിയുടെ 5 7 അടി ഉയരമുള്ള പ്രതിമ വിശ്വപ്രസിദ്ധമാണ് .
MSWS-TL: gēāmatēśvarjiyuṭe 5 7 aṭi uyaramuḷḷa pratima viśvaprasid'dhamāṇ .
MSWS-WW: Ghomateswarji's 5 7 feet tall statue is world famous
MSWS-ET: The 5 7-foot-tall statue of Gomateshwerjee is world famous.

The SMT system was, once again, unable to translate a word (Gomateshwar) of the test sentence. All the NMT systems have correct outputs, even though the choice of word for the phrase 'well known' MSWS is different.

8.3 Dravidian to Indo-Aryan

For the **te-hi** translation systems, NMT outperforms SMT and PAS outperforms BS and MSWS. Here we observe a sample output of the system:

Src: హృదయ స్పందన తగ్గిపోవడానికి రోగికి పూర్తి విశ్రాంతిని ఇవ్వాలి .
Ref: हृदय की धड़कन कम करने के लिए रोगी को पूरण आराम करने दें।
Ref-TL: hrday kee dhadakan kam karane ke lie rogee ko poorn aaraam karane den .
Ref-WW: Heart beat reduce to patient the complete rest let
Ref-ET: Allow the patient to take complete rest to reduce the heartbeat.
SMT-OP: हृदय की धड़कनें की परत कम करने के लिये रोगी को पूरण आराम देना चाहिए।
SMT-TL: hrday kee dhadakanen kee parat kam karane ke liye rogee ko poorn aaraam dena chaahie .
SMT-WW: Heart beats of layer reduce to patient the complete rest given should be
SMT-ET: The patient should be given complete rest to reduce the layer of heart-beat.

BS-OP: हृदय की गर्ति कम होने के लिए रोगी को पूरण आराम देना चाहिए।
BS-TL: hrday kee gati kam hone ke lie rogee ko poorn aaraam dena chaahie .
BS-WW: Heart rate reduce to patient the complete rest given should be
BS-ET: The patient should be given complete rest to reduce the heart rate.
PAS-OP: हृदय की धड़कन कम होने के लिए रोगी को पूरण आराम देना चाहिए।
PAS-TL: hrday kee dhadakan kam hone ke lie rogee ko poorn aaraam dena chaahie .
PAS-WW: Heart beat reduce to patient the complete rest given should be
PAS-ET: The patient should be given complete rest to reduce the heartbeat.
MSWS-OP: हृदय की धड़कने कम करने के लिए रोगी को पूरण आराम दें।
MSWS-TL: hrday kee dhadakanen kam karane ke lie rogee ko poorn aaraam den .
MSWS-WW: Heart beats reduce to patients the complete rest given
MSWS-ET: Give the patient complete rest to reduce heartbeat.

BS translates 'heartbeat' to 'speed'. Both PAS and MSWS choose the correct word, but PAS outputs the correct form—singular. SMT output not only uses the incorrect number for the word 'heartbeat', but also introduces a word that did not exist in the source or reference ('layer'), disrupting adequacy and fluency.

9 Conclusion and Future Work

We have presented in this paper a comprehensive study of Indian language NMT (ILNMT), setting a benchmark for ILNMT. We have empirically verified that for low-resource settings, a relatively small number of BPE merge operations delivers, particularly for related languages. We have also proposed a successful training data augmentation technique, that combines SMT with NMT, namely phrase table injection. Though not observed in all cases, it has proved particularly helpful when Dravidian languages are on the source side; improvements in SMT range from 0.88 (te-bn) to 5.96 (ml-hi), while improvements in NMT range from 0.36 (ml-te) to 1.9 (ta-hi) BLEU. When Indo-Aryan languages are at the source, these technique of phrase injection proffers modest improvements, ranging from 0.18 (hi-pa) to 2.04 (pa-bn) BLEU points over our baseline NMT scores. On the ILCI dataset, we have established a new state of the art for over half the language pairs (such as te-hi, hi-mr, ml-bn, pa-mr) in that dataset.

The incisive analysis of language properties and the transfer requirements between different pairs of languages as delineated in section 2 should guide proper selection of training data as well as insightful error analysis. The techniques described in this paper should point to ways of building ILNMT systems skirting around the challenges of low resources. Some possible directions of exploration are better phrase filtering criteria, synthesizing data and incorporating more linguistic features. The 'tricks' of subword based translation augmented with SMT phrases, morpheme segmentation and addition of word features are, in our belief, the only method ushering in large scale NMT for Indian languages. We hope the work described in this paper could serve as a good foundation for research on Indian language neural machine translation.

Acknowledgements We would like to thank the technology development for Indian languages (TDIL) programme and the Department of Electronics and Information Technology, Govt. of India for providing the ILCI corpus. We would also like to thank research scholars, Rudra Murthy, Tamali Banerjee, Jyotsana Khatri, Kevin Patel, and Diptesh Kanojia and members of CFILT for their valuable guidance and support.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Akella K, Himal Allu S, Ragupathi SS, Singhal A, Khan Z, Namboodiri VP, Jawahar CV (2020) Exploring pair-wise NMT for Indian languages. In: Proceedings of int'l conference natural language processing, Patna

Bahdanau D, Cho K, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:14090473

Banerjee T, Bhattacharyya P (2018) Meaningless yet meaningful: morphology grounded subword-level NMT. In: Proceedings of the second workshop on subword/character level models, association for computational linguistics, New Orleans, pp 55–60, https://doi.org/10.18653/v1/W18-1207, https://www.aclweb.org/anthology/W18-1207

Cherry C, Foster G, Bapna A, Firat O, Macherey W (2018) Revisiting character-based neural machine translation with capacity and compression. CoRR abs/1808.09943, arXiv:1808.09943

Dabre R, Chenhu C, Kunchukuttan A (2020) A survey of multilingual neural machine translation. ACM Comput Surv 53, 5, Article 99

Denkowski M, Neubig G (2017) Stronger baselines for trustable results in neural machine translation. In: Proceedings of the first workshop on neural machine translation, association for computational linguistics, Vancouver, pp 18–27, https://doi.org/10.18653/v1/W17-3203, https://www.aclweb.org/anthology/W17-3203

Ding S, Renduchintala A, Duh K (2019a) A call for prudent choice of subword merge operations. CoRR abs/1905.10453, arXiv:1905.10453

Ding S, Renduchintala A, Duh K (2019b) A call for prudent choice of subword merge operations in neural machine translation. arXiv preprint arXiv:190510453

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Jha GN (2010) The TDIL program and the Indian langauge corpora intitiative (ILCI). In: Proceedings of the seventh international conference on language resources and evaluation (LREC'10), European Language Resources Association (ELRA), Valletta, http://www.lrec-conf.org/proceedings/lrec2010/pdf/874_Paper.pdf

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, Herbst E (2007) Moses: open source toolkit for statistical machine translation. In: Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions, association for computational linguistics, Prague, Czech Republic, pp 177–180, https://www.aclweb.org/anthology/P07-2045

Kudo T (2018a) Subword regularization: improving neural network translation models with multiple subword candidates. CoRR abs/1804.10959, arXiv:1804.10959

Kudo T (2018b) Subword regularization: improving neural network translation models with multiple subword candidates. In: Proceedings of the 56th annual meeting of the association for computational linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, pp 66–75, https://doi.org/10.18653/v1/P18-1007, https://www.aclweb.org/anthology/P18-1007

Kunchukuttan A, Bhattacharyya P (2016) Learning variable length units for SMT between related languages via byte pair encoding. CoRR abs/1610.06510, arXiv:1610.06510

Kunchukuttan A, Bhattachatryya P (2021) Low resource machine translation and transliteration. CRC Press, Philadelphia

Kunchukuttan A, Mishra A, Chatterjee R, Shah R, Bhattacharyya P (2014a) Sata-anuvadak: tackling multiway translation of Indian languages. pan 841(54,570):4–135
Kunchukuttan A, Puduppully R, Chatterjee R, Mishra A, Bhattacharyya P (2014b) The IIT Bombay SMT system for icon 2014 tools contest
Murthy R, Kunchukuttan A, Bhattacharyya P (2019) Addressing word-order divergence in multilingual neural machine translation for extremely low resource languages. In: Proceedings of the annual conference of the North American chapter of the association for computational linguistics, Minneapolis
Papineni K, Roukos S, Ward T, Zhu WJ (2002) Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics, Association for Computational Linguistics, pp 311–318
Renduchintala A, Shapiro P, Duh K, Koehn P (2018) Character-aware decoder for neural machine translation. ArXiv abs/1809.02223
Revanuru K, Turlapaty K, Rao S (2017) Neural machine translation of indian languages. In: Proceedings of the 10th annual ACM India Computer Conference, Bhopal
Schuster M, Nakajima K (2012) Japanese and Korean voice search. In: 2012 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, pp 5149–5152
Sen S, Hasanuzzaman M, Ekbal A, Bhattacharyya P, Way A (2019) Take help from elder brother: old to modern English NMT with phrase pair feedback. In: 20th international conference on computational linguistics and intelligent text processing CICLing, La Rochelle
Sennrich R, Haddow B (2016) Linguistic input features improve neural machine translation. In: Proceedings of the first conference on machine translation: volume 1, research papers, association for computational linguistics, Berlin, pp 83–91, https://doi.org/10.18653/v1/W16-2209, https://www.aclweb.org/anthology/W16-2209
Sennrich R, Haddow B, Birch A (2016) 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), association for computational linguistics, Berlin, pp 1715–1725, https://doi.org/10.18653/v1/P16-1162, https://www.aclweb.org/anthology/P16-1162
Sennrich R, Birch A, Currey A, Germann U, Haddow B, Heafield K, Barone AVM, Williams P (2017) The university of Edinburgh's neural MT systems for WMT17. CoRR abs/1708.00726, http://arxiv.org/abs/1708.00726, arXiv:1708.00726
Subbārāo KV (2012) South Asian languages: a syntactic typology. Cambridge University Press
Tang Y, Meng F, Lu Z, Li H, Yu PL (2016) Neural machine translation with external phrase memory. arXiv preprint arXiv:160601792
Wu Y, Schuster M, Chen Z, Le QV, Norouzi M, Macherey W, Krikun M, Cao Y, Gao Q, Macherey K, Klingner J, Shah A, Johnson M, Liu X, Kaiser L, Gouws S, Kato Y, Kudo T, Kazawa H, Stevens K, Kurian G, Patil N, Wang W, Young C, Smith J, Riesa J, Rudnick A, Vinyals O, Corrado G, Hughes M, Dean J (2016) Google's neural machine translation system: Bridging the gap between human and machine translation. CoRR abs/1609.08144, arXiv:1609.08144
Zhao Y, Wang Y, Zhang J, Zong C (2018) Phrase table as recommendation memory for neural machine translation. arXiv preprint arXiv:180509960