Learning variable length units for SMT between related languages via Byte Pair Encoding

Anoop Kunchukuttan, Pushpak Bhattacharyya
Center For Indian Language Technology,
Department of Computer Science & Engineering
Indian Institute of Technology Bombay
{anoopk,pb}@cse.iitb.ac.in

Abstract

We explore the use of segments learnt using Byte Pair Encoding (referred to as *BPE units*) as basic units for statistical machine translation between *related* languages and compare it with *orthographic syllables*, which are currently the best performing basic units for this translation task. BPE identifies the most frequent character sequences as basic units, while orthographic syllables are linguistically motivated pseudo-syllables. We show that BPE units modestly outperform orthographic syllables as units of translation, showing up to 11% increase in BLEU score. While orthographic syllables can be used only for languages whose writing systems use vowel representations, BPE is writing system independent and we show that BPE outperforms other units for non-vowel writing systems too. Our results are supported by extensive experimentation spanning multiple language families and writing systems.

1 Introduction

The term, *related languages*, refers to languages that exhibit lexical and structural similarities on account of sharing a **common ancestry** or being in **contact for a long period of time** (Bhattacharyya et al., 2016). Examples of languages related by common ancestry are Slavic and Indo-Aryan languages. Prolonged contact leads to convergence of linguistic properties even if the languages are not related by ancestry and could lead to the formation of *linguistic areas* (Thomason, 2000). Examples of such linguistic areas are the Indian subcontinent (Emeneau, 1956), Balkan (Trubetzkoy, 1928) and Standard Average European (Haspelmath, 2001) linguistic areas. Genetic as well as contact relationship lead to related languages sharing vocabulary and structural features.

There is substantial government, commercial and cultural communication among people speaking related languages (Europe, India and South-East Asia being prominent examples and linguistic regions in Africa possibly in the future). As these regions integrate more closely and move to a digital society, translation between *related* languages is becoming an important requirement. In addition, translation to/from related languages to a *lingua franca* like English is also very important. However, despite significant communication between people speaking related languages, most of these languages have few parallel corpora resources. It is therefore important to leverage the relatedness of these languages to build good-quality statistical machine translation (SMT) systems given the lack of parallel corpora.

Modelling lexical similarity among related languages is the key to building good-quality SMT systems with limited parallel corpora. **Lexical similarity** implies related languages share many words with similar form (spelling/pronunciation) and meaning *e.g.* `blindness` is `andhapana` in Hindi, `aandhaLepaNaa` in Marathi. These words could be cognates, lateral borrowings or loan words from other languages.

Subword level transformations are an effective way for translation of such shared words. In this work, we propose use of Byte Pair Encoding (BPE) (Gage, 1994; Sennrich et al., 2016), a encoding method inspired from text compression literature, to learn basic translation units for translation between related languages. In previous work, the basic units of translation are either linguistically motivated (word, morpheme, syllable, etc.) or ad-hoc choices (character n-gram). In contrast, BPE is motivated by **statistical properties of text**.

The major contributions of our work are:

- We show that BPE units **modestly outperform orthographic syllable units** (Kunchukuttan and Bhattacharyya, 2016b), the best performing basic unit for translation between related languages, resulting in up to 11% improvement in BLEU score.
- Unlike orthographic syllables, BPE units are **writing system independent**. Orthographic syllables can only be applied to alphabetic and abugida writing systems. We show BPE units improve translation over word and morpheme level models for languages using *abjad* and *logographic* writing systems. Average BLEU score improvements of 18% and 6% over a baseline word-level model for language pairs involving abjad and logographic writing systems respectively were observed.
- Like orthographic syllables, BPE units outperform character, morph and word units when the language pairs show relatively less lexical similarity or belong to different language families (but have sufficient contact relation).
- While orthographic syllables approximate true syllables, we observe that BPE units learnt from the corpus **span various linguistic entities** (syllables, suffixes, morphemes, words, *etc.*). This may enable BPE level models to learn translation mappings at various levels simultaneously.
- We have reported results over a large number of languages (16 language pairs and 17 languages) which span 4 major language families and 10 writing systems of various types. To the best of our knowledge, this is the largest experiment for translation over related languages and the **broad coverage strongly supports our results**.
- We also show BPE units outperform other translation units in a **cross-domain translation** task.

The paper is organized as follows. Section 2 discusses related work. Section 3 discusses why BPE is a promising method for learning subword units and describes how we train BPE unit level translation models. Section 4 describes our experimental set-up. Section 5 reports the results of our experiments and analyses the results. Based on experimental results, we analyse why BPE units outperform other units in Section 6. Section 7 concludes the paper by summarizing our work and discussing further research directions.

2 Related Work

There are two broad set of approaches that have been explored in the literature for translation between related languages that leverage lexical similarity between source and target languages.

The first approach involves **transliteration of source words** into the target languages. This can done by transliterating the untranslated words in a post-processing step (Nakov and Tiedemann, 2012; Kunchukuttan et al., 2014), a technique generally used for handling named entities in SMT. However, transliteration candidates cannot be scored and tuned along with other features used in the SMT system. This limitation can be overcome by integrating the transliteration module into the decoder (Durrani et al., 2010), so both translation and transliteration candidates can be evaluated and scored simultaneously. This also allows transliteration *vs.* translation choices to be made.

Since a high degree of similarity exists at the subword level between related languages, the second approach looks at **translation with subword level basic units**. Character-level SMT has been explored for very closely related languages like *Bulgarian-Macedonian, Indonesian-Malay, Spanish-Catalan* with modest success (Vilar et al., 2007; Tiedemann, 2009a; Tiedemann and Nakov, 2013). Unigram-level learning provides very little context for learning translation models (Tiedemann, 2012). The use of character n-gram units to address this limitation leads to data sparsity for higher order n-grams and provides little benefit (Tiedemann and Nakov, 2013). These results were demonstrated primarily for very close European languages. Kunchukuttan and Bhattacharyya (2016b) proposed *orthographic syllables*, a linguistically-motivated variable-length unit, which approximates a syllable. This unit has outperformed character n-gram, word and morpheme level models as well as transliteration post-editing approaches mentioned earlier. They also showed orthographic syllables can outperform other units even when: (i) the lexical distance between related languages is reasonably large, (ii) the languages do not have a genetic relation, but only a contact relation.

Recently, subword level models have also gen-

erated interest for neural machine translation (NMT) systems. The motivation is the need to limit the **vocabulary of neural MT systems** in encoder-decoder architectures (Sutskever et al., 2014). It is in this context that Byte Pair Encoding, a data compression method (Gage, 1994), was adapted to learn subword units for NMT (Sennrich et al., 2016). Other subword units for NMT have also been proposed: character (Chung et al., 2016), Huffman encoding based units (Chitnis and DeNero, 2015), wordpieces (Schuster and Nakajima, 2012; Wu et al., 2016). Our hypothesis is that such subword units learnt from corpora are particularly suited for translation between related languages. In this paper, we test this hypothesis by using BPE to learn subword units.

3 BPE for related languages

We discuss why BPE is a promising method for learning subword units (subsections 3.1 and 3.2) and describe how we trained our BPE unit level translation models (subsections 3.3 and 3.4).

3.1 Motivation

Byte Pair Encoding is a data compression algorithm which was first adapted for Neural Machine Translation by Sennrich et al. (2016). For a given language, it is used to build a **vocabulary** relevant to translation by *discovering the most frequent character sequences* in the language.

For NMT, BPE enables efficient, high quality, open vocabulary translation by (i) limiting core vocabulary size, (ii) representing the most frequent words as atomic BPE units and rare words as compositions of the atomic BPE units.

For translation between related languages, we see a role for BPE beyond just learning a limited vocabulary for computational considerations. Rather, given the lexical similarity between related languages, we would like to *identify a small, core vocabulary of subwords* from which words in the language can be composed. These subwords represent stable, frequent patterns (possibly linguistic units like syllables, morphemes, affixes) for which mappings exist in other related languages. This alleviates the need for word level translation.

3.2 Comparison with orthographic syllables

We primarily compare BPE units with orthographic syllables (OS) (Kunchukuttan and Bhattacharyya, 2016b), which are good translation units for related languages. The *orthographic syllable* is a sequence of one or more consonants followed by a vowel, *i.e.* a C^+V unit, which approximates a linguistic syllable (*e.g. spacious* would be segmented as *spa ciou s*). Orthographic syllabification is rule based and applies to writing systems which represent vowels (alphabets and abugidas).

Both OS and BPE units are variable length units which provide longer and more relevant context for translation compared to character n-grams. In contrast to orthographic syllables, the BPE units are highly frequent character sequences reflecting the underlying statistical properties of the text. Some of the character sequences discovered by the BPE algorithm may be different linguistic units like syllables, morphemes and affixes. Moreover, BPE can be applied to text in any writing system.

3.3 The BPE Algorithm

We briefly summarize the BPE algorithm (described at length in Sennrich et al. (2016)). The input is a monolingual corpus for a language (one side of the parallel training data, in our case). We start with an *initial vocabulary viz.* the characters in the text corpus. The vocabulary is updated using an iterative greedy algorithm. In every iteration, the most frequent bigram (based on current vocabulary) in the corpus is added to the vocabulary (the *merge* operation). The corpus is again encoded using the updated vocabulary and this process is repeated for a pre-determined number of merge operations. The number of merge operations is the only hyperparameter to the system which needs to be tuned. A new word can be segmented by looking up the learnt vocabulary. For instance, a new word `scion` may be segmented as `sc ion` after looking up the learnt vocabulary, assuming `sc` and `ion` as BPE units learnt during training.

3.4 Training subword level translation model

We train subword level phrase-based SMT models between related languages. Along with BPE level, we also train PBSMT models at morpheme and OS levels for comparison.

For BPE, we learn the vocabulary separately for the source and target languages using the respective part of the training corpus. We segment the data into subwords during pre-processing and indicate word boundaries by a boundary marker (_) as shown in the example below. The boundary marker helps keep track of word boundaries, so

ben	Bengali	kok	Konkani	pan	Punjabi
bul	Bulgarian	kor	Korean	swe	Swedish
dan	Danish	mac	Macedonian	urd	Urdu
hin	Hindi	mar	Marathi	tam	Tamil
ind	Indonesian	mal	Malayalam	tel	Telugu
jpn	Japanese	may	Malay		

(a) List of languages used in experiments along with ISO 639-3 codes. These codes are used in the paper.

Language Family		Type of writing system	
Dravidian	mal,tam,tel	Alphabet	dan[1],swe[1],may[1]
Indo-Aryan	hin,urd,ben		ind[1],buc[2],mac[2]
	kok,mar,pan	Abugida	mal,tam,tel,hin
Slavic	bul,mac		ben,kok,mar,pan
Germanic	dan,swe	Syllabic	kor
Polynesian	may,ind	Logographic	jpn
Altaic	jpn,kor	Abjad	urd

(b) Classification of the languages and writing systems. (i) Indo-Aryan, Slavic and Germanic belong to the larger Indo-European language family. (ii) Alphabetic writing systems used by selected languages: Latin[1] and Cyrillic[2].

Table 1: Languages under experiments: details

the word level representation can be reconstructed after decoding.

word: `Childhood means simplicity .`

subword: `Chi ldhoo d _ mea ns _ si mpli ci ty _ .`

While building phrase-based SMT models at the subword level, we use (a) monotonic decoding since related languages have similar word order, (b) higher order languages models (10-gram) since data sparsity is a lesser concern owing to small vocabulary size (Vilar et al., 2007), and (c) word level tuning (by post-processing the decoder output during tuning) to optimize the correct translation metric (Nakov and Tiedemann, 2012). Following decoding, we used a simple method to regenerate words from subwords (desegmentation): concatenate subwords between consecutive occurrences of boundary marker characters.

4 Experimental Setup

We trained translation systems over the following basic units: character, morpheme, word, orthographic syllable and BPE unit. In this section, we summarize the languages and writing systems chosen for our experiments, the datasets used and the experimental configuration of our translation systems, and the evaluation methodology.

4.1 Languages and writing systems

Our experiments spanned a diverse set of languages: 16 language pairs, 17 languages and 10 writing systems. Table 1 summarizes the key aspects of the languages involved in the experiments.

The chosen languages span 4 major language families (6 major sub-groups: Indo-Aryan, Slavic and Germanic belong to the larger Indo-European language family). The languages exhibit diversity in word order and morphological complexity. Of course, between related languages, word order and morphological properties are similar. The classification of Japanese and Korean into the Altaic family is debated, but various lexical and grammatical similarities are indisputable, either due to genetic or cognate relationship (Robbeets, 2005; Vovin, 2010). However, the source of lexical similarity is immaterial to the current work. For want of a better classification, we use the name *Altaic* to indicate relatedness between Japanese and Korean.

The chosen language pairs also exhibit varying levels of lexical similarity. Table 3 shows an indication of the lexical similarity between them in terms of the Longest Common Subsequence Ratio (LCSR) (Melamed, 1995). The LCSR has been computed over the parallel training sentences at character level (shown only for language pairs where the writing systems are the same or can be easily mapped in order to do the LCSR computation). At one end of the spectrum, Malayalam-India, Urdu-Hindi, Macedonian-Bulgarian are dialects/registers of the same language and exhibit high lexical similarity. At the other end, pairs like Hindi-Malayalam belong to different language families, but show many lexical and grammatical similarities due to contact for a long time (Subbarao, 2012).

The chosen languages cover 5 types of writing systems. Of these, alphabetic and abugida writing systems represent vowels, logographic writing systems do not have vowels. The use of vowels is optional in abjad writing systems and depends on various factors and conventions. For instance, Urdu word segmentation can be very inconsistent (Durrani and Hussain, 2010) and generally short vowels are not denoted. The Korean *Hangul* writing system is syllabic, so the vowels are implicitly represented in the characters.

4.2 Datasets

Table 2a shows train, test and tune splits of the parallel corpora used. The Indo-Aryan and Dravidian language parallel corpora are obtained from

Language Pair	train	tune	test
ben-hin,pan-hin, kok-mar, mal-tam,tel-mal, hin-mal,mal-hin	44,777	1000	2000
urd-hin,ben-urd urd-mal,mal-urd	38,162	843	1707
bul-mac	150k	1000	2000
dan-swe	150k	1000	2000
may-ind	137k	1000	2000
kor-jpn,jpn-kor	69,809	1000	2000

(a) Parallel Corpora Size (no. of sentences)

Language	Size	Language	Size
hin (Bojar et al., 2014)	10M	urd (Jawaid et al., 2014)	5M
tam (Ramasamy et al., 2012)	1M	mar (news websites)	1.8M
mal (Quasthoff et al., 2006)	200K	swe (OpenSubtitles2016)	2.4M
mac (Tiedemann, 2009b)	680K	ind (Tiedemann, 2009b)	640K

(b) Details of additional monolingual corpora for training word-level language models (source and size in number of sentences)

Table 2: Training Corpus Statistics

the multilingual Indian Language Corpora Initiative (ILCI) corpus (Jha, 2012). Parallel corpora for other pairs were obtained from the *OpenSubtitles2016* section of the OPUS corpus collection (Tiedemann, 2009b). Language models for word-level systems were trained on the target side of training corpora plus additional monolingual corpora from various sources (See Table 2b for details). We used just the target language side of the parallel corpora for character, morpheme, OS and BPE-unit level LMs.

4.3 System details

We trained phrase-based SMT systems using the *Moses* system (Koehn et al., 2007), with the *grow-diag-final-and* heuristic for extracting phrases, and Batch MIRA (Cherry and Foster, 2012) for tuning (default parameters). We trained 5-gram LMs with Kneser-Ney smoothing for word and morpheme level models and 10-gram LMs for character, OS and BPE-unit level models. Subword level representation of sentences is long, hence we speed up decoding by using cube pruning with a smaller beam size (pop-limit=1000). This setting has been shown to have minimal impact on translation quality (Kunchukuttan and Bhattacharyya, 2016a).

We used unsupervised morphological-segmenters for generating morpheme representations (trained using *Morfessor* (Smit et al., 2014)). For Indian languages, we used the models distributed as part of the *Indic NLP Library*[1] (Kunchukuttan et al., 2014). We used orthographic syllabification rules from the *Indic NLP Library* for Indian languages, and custom rules for Latin and Slavic scripts. For training BPE models, we used the *subword-nmt*[2] library. We used *Juman*[3] and *Mecab*[4] for Japanese and Korean tokenization respectively.

For mapping characters across Indic scripts, we used the method described by Kunchukuttan et al. (2015) and implemented in the *Indic NLP Library*.

4.4 Evaluation

The primary evaluation metric is *word-level* BLEU (Papineni et al., 2002). We also report LeBLEU (Virpioja and Grönroos, 2015) scores as an alternative evaluation metric. LeBLEU is a variant of BLEU that does an edit-distance based, soft-matching of words and has been shown to be better for morphologically rich languages. We used bootstrap resampling for testing statistical significance (Koehn, 2004).

5 Results and Analysis

This section describes the results of various experiments and analyses them. A comparison of BPE with other units across languages and writing systems, choice of number of merge operations and effect of domain change and training data size are studied. We also report initial results with a joint bilingual BPE model.

5.1 Comparison of BPE with other units

Table 3 shows translation accuracies of all the language pairs under experimentation for different translation units, terms of BLEU as well as LeBLEU scores. The number of BPE merge operations was chosen such that the resultant vocabulary size would be equivalent to the vocabulary size of the orthographic syllable encoded corpus. Since we could not do orthographic syllabification for Urdu, Korean and Japanese, we selected the merge operations as follows: For Urdu, number of merge operations were selected based on Hindi OS vocabulary since Hindi and Urdu are registers of the same language. For Korean and Japanese, the number of BPE merge operations was set to 3000, discovered by tuning on a separate validation set.

[1] http://anoopkunchukuttan.github.io/indic_nlp_library
[2] https://github.com/rsennrich/subword-nmt
[3] http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN
[4] https://bitbucket.org/eunjeon/mecab-ko

Language Pair		BLEU					LeBLEU				
Src-Tgt	**LCSR**	**C**	**W**	**M**	**O**	$\mathbf{B}_{match}$	**C**	**W**	**M**	**O**	$\mathbf{B}_{match}$
ben-hin	52.30	27.95	32.47	32.17	**33.54**	33.22	0.672	0.682	0.708	0.715	**0.716**
pan-hin	67.99	71.26	70.07	71.29	**72.41**	72.22	0.905	0.871	0.899	0.906	**0.907**
kok-mar	54.51	19.83	21.30	22.81	23.43	**23.63**	0.632	0.636	0.659	**0.671**	0.665
mal-tam	39.04	4.50	6.38	7.61	7.84	**8.67**†	0.311	0.314	0.409	0.447	**0.465**
tel-mal	39.18	6.00	6.78	7.86	8.50	**8.79**	0.346	0.314	0.383	0.439	**0.443**
hin-mal	33.24	6.28	8.55	9.23	10.46	**10.73**	0.324	0.393	0.436	**0.477**	0.468
mal-hin	33.24	12.33	15.18	17.08	18.44	**20.54**	0.444	0.460	0.528	0.551	**0.565**
urd-hin	-	52.57	55.12	52.87	NA	**55.55**	0.804	0.795	0.792	NA	**0.823**
ben-urd	-	18.16	27.06	27.31	NA	**28.06**	0.607	0.660	0.671	NA	**0.692**
urd-mal	-	3.13	6.49	7.05	NA	**8.44**	0.247	0.350	0.379	NA	**0.416**
mal-urd	-	8.90	13.22	15.30	NA	**18.48**	0.444	0.454	0.522	NA	**0.568**
bul-mac	62.85	20.61	21.20	-	**21.95**	21.73	0.603	0.606	-	**0.613**	0.599
dan-swe	63.39	35.36	35.13	-	29.66	**35.77**†	0.692	**0.694**	-	0.674	0.682
may-ind	73.54	60.50	**61.33**	-	60.79	59.54†	0.827	**0.832**	-	0.828	0.825
kor-jpn	-	8.51	9.90	-	NA	**10.23**	0.396	0.372	-	NA	**0.408**
jpn-kor	-	8.17	8.44	-	NA	**9.02**	0.372	0.350	-	NA	**0.374**

Table 3: Translation accuracies for various translation units (BLEU and LeBLEU scores reported). The reported scores are:- **W**: word-level, **M**: morpheme, **O**: orthographic syllable, $\mathbf{B}_{match}$: BPE units with number of merge operations selected to match vocabulary size of OS encoding. See discussion related to exceptions for pairs involving Urdu, Korean and Japanese. (a) The values marked in **bold** indicate best score for a language pair (b) **LCSR** indicates lexical similarity (c) *NA: Not Applicable.* (d) † indicates that difference in BLEU scoresbetween $\mathbf{B}_{match}$ and **O** are statistically significant ($p < 0.05$)

Our major observations are described below (based on BLEU scores):

• BPE units are clearly better than the traditional word and morpheme representations. The average BLEU score improvement is 15% over word-based results and 11% over morpheme-based results. The only exception is Malay-Indonesian, which are registers of the same language.

• BPE units also show modest improvement over the recently proposed orthographic syllables over most language pairs (average improvement of 2.6% and maximum improvement of up to 11%). The scores are statistically significant only for some language pairs. The only exceptions are Bengali-Hindi, Punjabi-Hindi and Malay-Indonesian - all these languages pairs have relatively less morphological affixing (*Bengali-Hindi, Punjabi-Hindi*) or are registers of the same language (Malay-Indonesian). For Bengali-Hindi and Punjabi-Hindi, the BPE unit translation accuracies are quite close to OS level accuracies. Since OS level models have been shown to be better than character level models (Kunchukuttan and Bhattacharyya, 2016b), BPE units are better than character level models by transitivity.

• BPE units also outperform other units for translation between language pairs belonging to different language pairs, but having a long contact relationship *viz.* Malayalam-Hindi and Hindi-Malayalam.

• It is worth mentioning that BPE units provide a substantial benefit over OS units when translation involves a morphologically rich language. In translations involving Malayalam, Tamil and Telugu, average accuracy improvement of 6.25% were observed.

The LeBLEU scores also show the same trends as the BLEU scores.

5.2 Applicability to different writing systems

The utility of orthographic syllables as translation units is limited to languages that use writing systems which represent vowels. Alphabetic and abugida writing systems fall into this category. On the other hand, logographic writing systems (Japanese Kanji, Chinese) and abjad writing systems (Arabic, Hebrew, Syriac, etc.) do not represent vowels. To be more precise, abjad writing systems may represent some/all vowels depending on language, pragmatics and conventions. Syllabic writing systems like Korean Hangul do not explicitly represent vowels, since the basic unit (the syllable) implicitly represents the vowels. The major advantage of Byte Pair Encoding is its **writing system independence** and our results show

	O	B_{match}	B_{1k}	B_{2k}	B_{3k}	B_{4k}
ben-hin	**33.54**	33.22	33.16	33.25	33.30	32.99
pan-hin	**72.41**	72.22	72.28	72.19	72.08	71.94
kok-mar	23.43	23.63	**23.84**	23.73	23.79	23.30
mal-tam	7.84	8.67	8.66	8.71	8.63	**8.74**
tel-mal	8.50	8.79	8.99	8.83	**9.12**	8.76
hin-mal	10.46	10.73	**10.96**	10.89	10.61	10.55
mal-hin	18.44	20.54	**21.23**	20.53	20.64	20.19
urd-hin	NA	55.55	**55.69**	55.49	55.57	55.47
ben-urd	NA	28.06	28.12	**28.19**	28.03	27.93
urd-mal	NA	8.44	8.22	8.04	8.02	**8.57**
mal-urd	NA	18.48	18.72	18.47	**18.79**	18.18
bul-mac	21.95	21.73	21.74	**22.27**	21.95	21.94
dan-swe	29.66	35.77	36.38	36.18	**36.61**	36.2
may-ind	**60.79**	59.54	60.63	60.24	60.35	60.15
kor-jpn	NA	NA	10.13	9.8	**10.23**	9.92
jpn-kor	NA	NA	**9.29**	9.23	9.02	8.96

Table 4: Translation accuracies for BPE models trained with different number of merge operations (BLEU). Underlined scores indicate the best BPE configuration when OS is the best-performing for a language pair.

that BPE encoded units are useful for translation involving abjad (Urdu uses an extended Arabic writing system), logographic (Japanese Kanji) and syllabic (Korean Hangul) writing systems. For language pairs involving Urdu, there is an 18% average improvement over word-level and 12% average improvement over morpheme-level translation accuracy. For Japanese-Korean language pairs, an average improvement of 6% in translation accuracy over a word-level baseline is observed.

5.3 Choosing number of BPE merges

The above mentioned results for BPE units do not explore optimal values of the number of merge operations. This is the only hyper-parameter that has to be selected for BPE. We experimented with number of merge operations ranging from 1000 to 4000 and the translation results for these are shown in Table 4. Selecting the optimal value of merge operations lead to a modest, average increase of 1.6% and maximum increase of 3.5% in the translation accuracy over B_{match} across different language pairs .

We also experimented with higher number of merge operations for some language pairs, but there seemed to be no benefit with a higher number of merge operations. Compared to the number of merge operations reported by Sennrich et al. (2016) in a more general setting for NMT (60k), the number of merge operations is far less for translation between related languages with limited parallel corpora. We must bear in mind that their goal was different: available parallel corpus was not an issue, but they wanted to handle as large a vocabulary as possible for open-vocabulary NMT. Yet, the low number of merge operations suggest that BPE encoding captures the core vocabulary required for translation between related tasks.

Pair	C	W	M	O	B_{match}
pan-hin	58.07	58.95	**59.71**	57.95	59.66†
kok-mar	17.97	18.83	18.53	**19.12**	18.42†
mal-tam	4.12	5.49	5.84	5.93	**6.75**†
tel-mal	3.11	3.26	**4.06**	3.83	3.75
hin-mal	3.85	5.18	5.99	6.24	**6.37**
mal-hin	8.42	9.92	11.12	13.36	**14.45**†

(a) BLEU scores

Pair	C	W	M	O	B_{match}
pan-hin	0.869	0.825	0.868	0.863	**0.876**
kok-mar	0.647	0.641	0.643	**0.665**	0.653
mal-tal	0.301	0.261	0.378	0.452	**0.475**
tel-mal	0.246	0.198	0.238	0.297	**0.300**
hin-mal	0.281	0.336	0.354	**0.404**	0.384
mal-hin	0.439	0.371	0.466	0.548	**0.565**

(b) LeBLEU scores

Table 5: Translation accuracies for Agriculture Domain † indicates statistically significant difference in BLEU score between **O** and $\mathbf{B}_{match}$. BLEU score differences between $\mathbf{B}_{match}$ and **W** are also statistically significant (except Konkani-Marathi) ($p < 0.05$)

5.4 Robustness to Domain Change

Since we are concerned with low resource scenarios, a desirable property of subword units is robustness of the translation models to change of translation domain. Kunchukuttan and Bhattacharyya (2016b) have shown that OS level models are robust to domain change. Since BPE units are learnt from a specific corpus, it is not guaranteed that they would also be robust to domain changes. To study the behaviour of BPE unit trained models, we also tested the translation models trained on tourism & health domains on an agriculture domain test set of 1000 sentences (see Table 5 for results). *In this cross-domain translation scenario, the BPE level model outperforms the OS-level and word-level models*

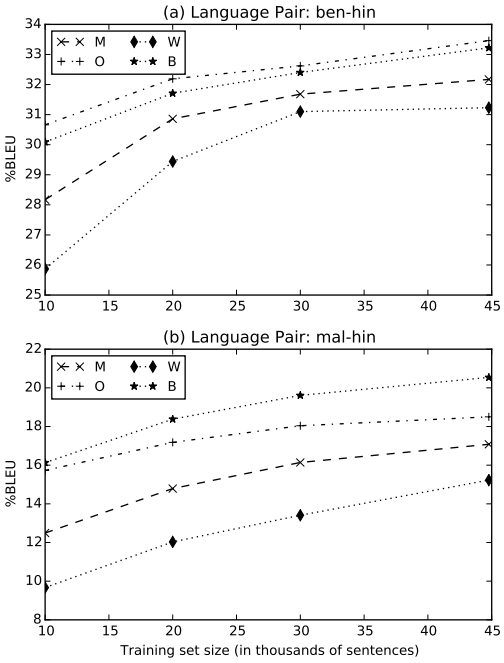

Figure 1: Effect of training data size on translation accuracy for different basic units

for most language pairs. The Konkani-Marathi pair alone shows a degradation using the OS level model. The BPE model is almost on par with the OS level model for Telugu-Malayalam and Hindi-Malayalam.

5.5 Effect of training data size

For different training set sizes, we trained SMT systems with various representation units (Figure 1 shows the learning curves for two language pairs). BPE level models are better than OS, morpheme and word level across a range of dataset sizes. Especially when the training data is very small, the OS and BPE level models perform significantly better than the word and morpheme level models. For Malayalam-Hindi, the BPE level model is better than the OS level model at utilizing more training data.

5.6 Joint bilingual learning of BPE units

In the experiments discussed so far, we learnt the BPE vocabulary separately for the source and target languages. In this section, we describe our experiments with jointly learning BPE vocabulary over source and target language corpora as suggested by Sennrich et al. (2016). The idea is to learn an encoding that is consistent across source and target languages and therefore helps alignment. We expect a significant number of common BPE units between related languages. If source and target languages use the same writing system, then a joint model is created by learning BPE over concatenated source and target language corpus. If the writing systems are different, then we transliterate one corpus to another by one-one character mappings. This is possible between Indic scripts. But this scheme cannot be applied between Urdu and Indic scripts as well as between Korean Hangul and Japanese Kanji scripts.

	Best$_{prev}$	**JB**$_{1k}$	**JB**$_{2k}$	**JB**$_{3k}$	**JB**$_{4k}$
ben-hin	O (33.46)	**33.54**	33.23	33.54	33.35
pan-hin	O (**72.51**)	72.41	72.35	72.13	72.04
kok-mar	B$_{1k}$ (23.84)	**24.01**	23.76	23.8	23.86
mal-tam	B$_{4k}$ (8.74)	8.6	**8.82**	8.74	8.72
tel-mal	B$_{3k}$ (**9.12**)	8.47	8.84	8.89	8.92
hin-mal	B$_{1k}$ (10.96)	**11.19**	11.09	11.1	10.96
mal-hin	B$_{1k}$ (**21.23**)	20.79	21.22	21.12	21.06
bul-mac	B$_{2k}$ (**22.27**)	22.11	22.17	21.58	22.24
dan-swe	B$_{3k}$ (36.61)	36.15	**36.86**	36.51	36.71
may-ind	O (61.24)	**61.26**	60.98	61.11	60.66

Table 6: Translation accuracies for Joint BPE models trained with different number of merge operations (BLEU). The **Best**$_{prev}$ indicates the best performing units and their accuracy scores from Tables 3 and 4 shown for comparison.

Table 6 shows the results of the joint BPE model for language pairs where such a model is built. We do not see any major improvement over the monolingual BPE model due to the joint BPE model.

6 Why are BPE units better than others?

The improved performance of BPE units compared to word-level and morpheme-level representations is easy to explain: with a limited vocabulary they **address the problem of data sparsity**. But character level models also have a limited vocabulary, yet they do not improve translation performance except for very close languages. Character level models learn character mappings effectively, which is sufficient for translating related languages which are very close to each other (translation is akin to transliteration in these cases). But they are not sufficient for translating related languages that are more divergent. In this case, translating cognates, morphological affixes, non-cognates *etc.* require a larger context. So, BPE and OS units — which **provide more**

Src-Tgt	Word	Morph	BPE	OS	Char
ben-hin	0.40	0.58	0.60	0.62	0.71
pan-hin	0.50	0.64	0.69	0.70	0.72
kok-mar	0.66	0.63	0.64	0.67	0.74
mal-tam	0.46	0.56	0.70	0.71	0.77
tel-mal	0.45	0.52	0.62	0.64	0.78
hin-mal	0.39	0.46	0.52	0.58	0.79
mal-hin	0.37	0.45	0.54	0.60	0.71

Table 7: Pearson's correlation coefficient between lexical similarity and translation accuracy (both in terms of LCSR at character level). *This was computed over the test set between: (i) sentence level lexical similarity between source and target sentences and (ii) sentence level translation accuracy between hypothesis and reference.*

	hin	mar	mal
OS	tI, stha	mA, nA	kka, nI
Suffix	ke, me.m	ChyA, madhIla	unnu, .e∼Nkill.m
Word	paryaTaka, athavA	prAchIna, aneka	bhakShaN.m, yAtra

Table 8: Examples of BPE units for Indian languages. (ITRANS transliteration shown)

context — outperform character units.

A study of the correlation between lexical similarity and translation quality makes this evident (See Table 7). We see that character models work best when the source and target sentences are lexically very similar. The additional context decouples OS and BPE units from lexical similarity. Words and morphemes show the least correlation since they do not depend on lexical similarity.

Why does BPE performs better than OS which also provides a larger contextual window for translation? While orthographic syllables represent just approximate syllables, we observe that BPE units also **represent higher level semantic units like frequent morphemes, suffixes and entire words**. Table 8 shows a few examples for some Indian languages. So, BPE level models can learn semantically similar translation mappings in addition to lexically similar mappings. In this way, BPE units enable the translation models to **balance the use of lexical similarity with semantic similarity**. This further decouples the translation quality from lexical similarity as seen from the table. BPE units also have an **additional degree of freedom** (choice of vocabulary size), which allows tuning for best translation performance. This could be important when larger parallel corpora are available, allowing larger vocabulary sizes.

7 Conclusion & Future Work

We show that translation units learnt using BPE can outperform all previously proposed translation units, including the best-performing orthographic syllables, for SMT between related languages when limited parallel corpus is available. Moreover, BPE encoding is writing system independent, hence it can be applied to any language. Experimentation on a large number of language pairs spanning diverse language families and writing systems lend strong support to our results. We also show that BPE units are more robust to change in translation domain. They perform better for morphologically rich languages and extremely data scarce scenarios.

BPE seems to be beneficial because it enables discovery of translation mappings at various levels simultaneously (syllables, suffixes, morphemes, words, *etc.*). We would like to further pursue this line of work and investigate better translation units. This is also a question relevant to translation with subwords in NMT. NMT between related languages using BPE and similar encodings is also an obvious direction to explore.

Given the improved performance of the BPE-unit, tasks involving related languages *viz.* pivot based MT, domain adaptation (Tiedemann, 2012) and translation between a *lingua franca* and related languages (Wang et al., 2012) can be revisited with BPE units.

Acknowledgments

We thank the Technology Development for Indian Languages (TDIL) Programme and the Department of Electronics & Information Technology, Govt. of India for their support. We also thank the reviewers for their feedback.

References

Pushpak Bhattacharyya, Mitesh Khapra, and Anoop Kunchukuttan. 2016. Statistical machine translation between related languages. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*.

Ondřej Bojar, Vojtěch Diatka, Pavel Rychlý, Pavel Straňák, Vít Suchomel, Aleš Tamchyna, and Daniel Zeman. 2014. HindEnCorp – Hindi-English and Hindi-only Corpus for Machine Translation. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*.

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

Rohan Chitnis and John DeNero. 2015. Variable-length word encodings for neural translation models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Junyoung Chung, Kyunghyun Cho, and Yoshua Bengio. 2016. A character-level decoder without explicit segmentation for neural machine translation. In *Proceedings of the Meeting of the Association for Computational Linguistics*.

Nadir Durrani and Sarmad Hussain. 2010. Urdu word segmentation. In *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Nadir Durrani, Hassan Sajjad, Alexander Fraser, and Helmut Schmid. 2010. Hindi-to-Urdu machine translation through transliteration. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*.

Murray B Emeneau. 1956. India as a lingustic area. *Language* .

Philip Gage. 1994. A new algorithm for data compression .

Martin Haspelmath. 2001. The european linguistic area: Standard average european. In Martin Haspelmath, editor, *Language typology and language universals: An international handbook*, Walter de Gruyter.

Bushra Jawaid, Amir Kamran, and Ondřej Bojar. 2014. Urdu monolingual corpus. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University in Prague. http://hdl.handle.net/11858/00-097C-0000-0023-65A9-5.

Girish Nath Jha. 2012. The TDIL program and the Indian Language Corpora Initiative. In *Proceedings of the Language Resources and Evaluation Conference*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*.

Anoop Kunchukuttan and Pushpak Bhattacharyya. 2016a. Faster decoding for subword level phrase-based smt between related languages. In *Third Workshop on NLP for Similar Languages, Varieties and Dialects*.

Anoop Kunchukuttan and Pushpak Bhattacharyya. 2016b. Orthographic syllable as basic unit for smt between related languages. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Anoop Kunchukuttan, Ratish Puduppully, and Pushpak Bhattacharyya. 2015. Brahmi-Net: A transliteration and script conversion system for languages of the Indian subcontinent. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies: System Demonstrations*.

Anoop Kunchukuttan, Ratish Pudupully, Rajen Chatterjee, Abhijit Mishra, and Pushpak Bhattacharyya. 2014. The IIT Bombay SMT System for ICON 2014 Tools contest. In *Proceedings on the NLP Tools Contest at International Conference on Natural Language Processing*.

I Dan Melamed. 1995. Automatic evaluation and uniform filter cascades for inducing n-best translation lexicons. In *Proceedings of Third Workshop on Very Large Corpora*.

Preslav Nakov and Jörg Tiedemann. 2012. Combining word-level and character-level models for machine translation between closely-related languages. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the Meeting of the Association for Computational Linguistics*.

Uwe Quasthoff, Matthias Richter, and Christian Biemann. 2006. Corpus portal for search in monolingual corpora. In *Proceedings of the fifth international conference on language resources and evaluation*.

Loganathan Ramasamy, Ondřej Bojar, and Zdeněk Žabokrtský. 2012. Morphological Processing for English-Tamil Statistical Machine Translation. In *Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages*.

Martine Irma Robbeets. 2005. *Is Japanese Related to Korean, Tungusic, Mongolic and Turkic?*. Otto Harrassowitz Verlag.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the Meeting of the Association for Computational Linguistics*.

Peter Smit, Sami Virpioja, Stig-Arne Grönroos, and Mikko Kurimo. 2014. Morfessor 2.0: Toolkit for statistical morphological segmentation .

Karumuri Subbarao. 2012. *South Asian Languages: A Syntactic Typology*. Cambridge University Press.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of Advances in Neural Information Processing Systems*.

Sarah Thomason. 2000. Linguistic areas and language history. In John Nerbonne and Jos Schaeken, editors, *Languages in Contact*, Editions Rodopi B.V., Brill.

Jörg Tiedemann. 2009a. Character-based PBSMT for closely related languages. In *Proceedings of the 13th Conference of the European Association for Machine Translation*.

Jörg Tiedemann. 2009b. News from opus-a collection of multilingual parallel corpora with tools and interfaces. In *Proceedings of the Conference on Recent Advances in Natural Language Processing*.

Jörg Tiedemann. 2012. Character-based pivot translation for under-resourced languages and domains. In *EACL*.

Jörg Tiedemann and Preslav Nakov. 2013. Analyzing the use of character-level translation with sparse and noisy datasets. In *Proceedings of the Conference on Recent Advances in Natural Language Processing*.

Nikolai Trubetzkoy. 1928. Proposition 16. In *Actes du premier congres international des linguistes La Haye*.

David Vilar, Jan-T Peter, and Hermann Ney. 2007. Can we translate letters? In *Proceedings of the Second Workshop on Statistical Machine Translation*.

Sami Virpioja and Stig-Arne Grönroos. 2015. Lebleu: N-gram-based translation evaluation score for morphologically complex languages. In *Proceedings of the Workshop on Machine Translation*.

Alexander Vovin. 2010. *Korea-Japonica: A Re-Evaluation of a Common Genetic Origin*. University of Hawaii Press.

Pidong Wang, Preslav Nakov, and Hwee Tou Ng. 2012. Source language adaptation for resource-poor machine translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*.

Y. Wu, M. Schuster, Z. Chen, Q. V. Le, and M. Norouzi. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *ArXiv e-prints: abs/1609.08144* .