Optimal Word Segmentation for Neural Machine Translation into Dravidian Languages

Prajit Dhar **Arianna Bisazza** **Gertjan van Noord**
University of Groningen
{p.dhar, a.bisazza, g.j.m.van.noord}@rug.nl

Abstract

Dravidian languages, such as Kannada and Tamil, are notoriously difficult to translate by state-of-the-art neural models. This stems from the fact that these languages are morphologically very rich as well as being low-resourced. In this paper, we focus on subword segmentation and evaluate Linguistically Motivated Vocabulary Reduction (LMVR) against the more commonly used SentencePiece (SP) for the task of translating from English into four different Dravidian languages. Additionally we investigate the optimal subword vocabulary size for each language. We find that SP is the overall best choice for segmentation, and that larger subword vocabulary sizes lead to higher translation quality.

1 Introduction

Dravidian languages are an important family of languages spoken by about 250 million of people primarily located in Southern India and Sri Lanka (Steever, 2019). Kannada (KN), Malayalam (MA), Tamil (TA) and Telugu (TE) are the four most spoken Dravidian languages with approximately 47, 34, 71 and 79 million native speakers, respectively. Together, they account for 93% of all Dravidian language speakers. While Kannada, Malayalam and Tamil are classified as South Dravidian languages, Telugu is a part of South-Central Dravidian languages. All four languages are SOV (Subject-Object-Verb) languages with free word order. They are highly agglutinative and inflectionally rich languages. Additionally, each language has a different writing system. Table 1 presents an English sentence example and its Dravidian-language translations.

The highly complex morphology of the Dravidian languages under study is illustrated if we compare translated sentence pairs. The analysis of our parallel datasets (section 4.1, Table 3) shows for instance that an average English sentence contains almost ten times as many words as its Kannada equivalent. For the other three languages, the ratio is a bit smaller but the difference with English remains considerable. This indicates why it is important to consider word segmentation algorithms as part of the translation system.

In this paper we describe our work on Neural Machine Translation (NMT) from English into the Dravidian languages Kannada, Malayalam, Tamil and Telugu. We investigated the optimal translation settings for the pairs and in particular looked at the effect of word segmentation. The aim of the paper is to answer the following research questions:

- Does LMVR, a linguistically motivated word segmentation algorithm, outperform the purely data-driven SentencePiece?
- What is the optimal subword dictionary size for translating from English into these Dravidian languages?

In what follows, we review the relevant previous work (Sect. 2), introduce the two segmenters (Sect. 3), describe the experimental setup (Sect. 4), and present our answers to the above research questions (Sect. 5).

2 Previous Work

2.1 Translation Systems

Statistical Machine Translation One of the earliest automatic translation systems for English into a Dravidian language was the English→Tamil system by Germann (2001). They trained a hybrid rule-based/statistical machine translation system that was trained on only 5k English-Tamil parallel sentences. Ramasamy et al. (2012) created SMT systems (phrase-based and hierarchical) which were trained on a dataset of 190k parallel

Proceedings of the 8th Workshop on Asian Translation, pages 181–190
Bangkok, Thailand (online), August 5-6, 2021. ©2021 Association for Computational Linguistics

EN	He was born in Thirukkuvalai village in Nagapattinam District on 3rd June, 1924.
KN	ಅವರು ನಾಗಪಟ್ಟಣಂ ಜಿಲ್ಲೆಯ ತಿರುಕ್ಕುವಲಯ್ ಗ್ರಾಮದಲ್ಲಿ 1924ರ ಜೂನ್ 3ರಂದು ಜನಿಸಿದ್ದರು. avaru nāgapaṭṭaṇam jilleya tirukkuvalay grāmadalli 1924ra jūn 3randu janisiddaru.
ML	1924ല് നാഗപട്ടണം ജില്ലയിലെ തിരുക്കുവളൈ ഗ്രാമത്തിലാണ് അദ്ദേഹം ജനിച്ചത് 1924l nāgapaṭṭaṇam jillayile tirukkuvaḷai grāmattilāṇ addēham janiccat.
TA	நாகப்பட்டினம் மாவட்டம் திருக்குவளைக் கிராமத்தில் அவர் 1924-ஆம் ஆண்டு ஜூன் மாதம் 3-ஆம் தேதி பிறந்தார். nāgappaṭṭinam māvaṭṭam tirukkuvaḷaik kirāmattil avar 1924-ām āṇṭu jūn mātam 3-ām tēti pirantār.
TE	ఆయన నాగపట్టణం జిల్లా తిరుక్కువాలై గ్రామంలో 1924 జూన్ 3న జన్మించారు. āyana nāgapaṭṭaṇam jillā tirukkuvālai grāmanlō 1924 jūn 3na janmincāru.

Table 1: Example sentence in English along with its translation and transliteration in the four Dravidian languages.

sentences (henceforth referred to as UFAL). They also reported that applying pre-processing steps involving morphological rules based on Tamil suffixes improved the BLEU score of the baseline model to a small extent (from 9.42 to 9.77). For the Indic languages multilingual tasks of WAT-2018, the Phrasal-based SMT system of Ojha et al. (2018) with a BLEU score of 30.53.

Subsequent papers also focused on SMT systems for Malayalam and Telugu with some notable work including: (Anto and Nisha, 2016; Sreelekha and Bhattacharyya, 2017, 2018) for Malayalam and (Lingam et al., 2014; Yadav and Lingam, 2017) for Telugu.

Neural Machine Translation On the neural machine translation (NMT) side, there have been a handful of NMT systems trained on English→Tamil. On the aforementioned Indic languages multilingual tasks of WAT-2018, Sen et al. (2018), Dabre et al. (2018) reported only 11.88 and 18.60 BLEU scores, respectively, for English→Tamil. The poor performance of these systems compared to the 30.53 BLEU score of the SMT system (Ojha et al., 2018) showed that those NMT systems were not yet suitable for translating into the morphologically rich Tamil.

However, the following year, Philip et al. (2019) outperformed Ramasamy et al. (2012) on the UFAL dataset with a BLEU score of 13.05 (the previous best score on this test set was 9.77). They report that techniques such as domain adaptation and back-translation can make training NMT systems on low-resource languages possible. Similar findings was also reported by Ramesh et al. (2020) for Tamil and Dandapat and Federmann (2018) for Telugu .

To the best of our knowledge and as of 2021, there has not been any scientific publication involving translation to and from Kannada, except for Chakravarthi et al. (2019). One possible reason for this could be the fact that sizeable corpora involving Kannada (i.e. in the order of magnitude of at least thousand sentences) have been readily available only since 2019, with the release of the JW300 Corpus (Agić and Vulić, 2019).

Multilingual NMT Since 2018 several studies have presented multilingual NMT systems that can handle English → Malayalam, Tamil and Telugu translation (Dabre et al., 2018; Choudhary et al., 2020; Ojha et al., 2018; Sen et al., 2018; Yu et al., 2020; Dabre and Chakrabarty, 2020). In particular, Sen et al. (2018) presented results where the BLEU score improved when comparing monolingual and multilingual models. Conversely, Yu et al. (2020) found that NMT systems that were multi-way (Indic ↔ Indic) performed worse than English ↔ Indic systems.

To our knowledge, no work so far has explored the effect of the segmentation algorithm and dictionary size on the four languages: Kannada, Malayalam, Tamil and Telugu.

3 Subword Segmentation Techniques

Prior to the emergence of subword segmenters, translation systems were plagued with the issue of

Name	Domain	Available in:			
		Kannada	Malayalam	Tamil	Telugu
Bible	Religion	18	1		14
ELRC	COVID-19		<1	<1	<1
GNOME	Technical	<1	<1	<1	<1
JW300	Religion	70	45	52	45
KDE	Technical	1	<1	<1	<1
NLPC	General			<1	
OpenSubtitles	Cinema		26	3	3
CVIT-PIB	Press		5	10	10
PMIndia	Politics	10	4	3	8
Tanzil	Religion		18	9	
Tatoeba	General	<1	<1	<1	<1
Ted2020	General	<1	<1	<1	1
TICO-19	COVID-19			<1	
Ubuntu	Technical	<1	<1	<1	<1
UFAL	Mixed			11	
Wikimatrix	General		<1	10	18
Wikititles	General			1	

Table 2: Composition of training corpora. The numbers indicate the relative size (in percentages) of the corresponding part for that language.

out-of-vocabulary (OOV) tokens. This was particularly an issue for translations involving agglutinative languages such as Turkish (Ataman and Federico, 2018) or Malayalam (Manohar et al., 2020). Various segmentation algorithms were brought forward to circumvent this issue and in turn, improve translation quality.

Perhaps the most widely used algorithm in NMT to date is the language-agnostic Byte Pair Encoding (BPE) by Sennrich et al. (2016). Initially proposed by Gage (1994), BPE was repurposed by Sennrich et al. (2016) for the task of subword segmentation, and is based on a simple principle whereby pairs of character sequences that are frequently observed in a corpus get *merged* iteratively until a predetermined dictionary size is attained. In this paper we use a popular implementation of BPE, called **SentencePiece (SP)** (Kudo and Richardson, 2018).

While purely statistical algorithms are able to segment any token into smaller segments, there is no guarantee that the generated tokens will be linguistically sensible. Unsupervised morphological induction is a rich area of research that also aims at learning a segmentation from data, but in a linguistically motivated way. The most well-known example is Morphessor with its different variants (Creutz and Lagus, 2002; Kohonen et al., 2010; Grönroos et al., 2014). An important obstacle to applying Morfessor to the task of NMT is the lack of a mechanism to determine the dictionary size. To address this, Ataman et al. (2017) proposed a modification of Morfessor FlatCat (Grönroos et al., 2014), called **Linguistically Motivated Vocabulary Reduction (LMVR)**. Specifically, LMVR imposes an extra condition on the cost function of Morfessor Flatcat so as to favour vocabularies of the desired size. In a comparison of LMVR to BPE, Ataman et al. (2017) reported a +2.3 BLEU improvement on the English-Turkish translation task of WMT18.

Given the encouraging results reported on the agglutinative Turkish language, we hypothesise that translation into Dravidian languages may also benefit from a linguistically motivated segmenter, and evaluate LMVR against SP across varying vocabulary sizes.

4 Experimental Setup

4.1 Training Corpora

The parallel training data is mostly taken from the datasets available for the MultiIndicMT task from WAT 2021. If a certain dataset is not available from the MultiIndicMT training repository, we resorted to extract that dataset from OPUS (Tiedemann, 2012) or WMT20. Table 2 reports on the datasets that we used along with their domain and their source.

After extracting and cleaning the data (see below), approximately 8 million English tokens and their corresponding target language tokens are selected as our training corpora. We fixed the number of source tokens across language pairs in or-

Target Language	Tokens(k)	EN Tokens(k)	Sentences(k)	Source/Target Token Ratio
Kannada	817	7791	361	9.53
Malayalam	1153	7973	458	6.91
Tamil	1171	7854	345	6.71
Telugu	1027	7872	385	7.67

Table 3: Approximate sizes (in thousands) of the parallel training corpora

der to compare the efficacy of a segmentation technique across the languages without a size bias. Table 3 presents the statistics on the corpora for all language pairs. One takeaway from the table is that there is a very large difference in the token sizes between English and the Dravidian languages. On average, there are 6 to 9 times more tokens on the English side of a corpus than on its Dravidian language translation. This shows that all our Dravidian languages are morphologically *very* complex, but there are also important differences among them, with Kannada having the highest source/target ratio, considerably higher than the more widely studied Tamil language.

4.2 Pre-Processing

Sentence pairs with identical source and target sides, or with more than 150 tokens are removed. The target language texts are then normalized using the Indic NLP Library[1]. Afterwards, either SP[2] or LMVR[3] is used to segment both source and target sentences. To further reduce noise in the datasets, we discard sentences pairs with either (i) a target to source length ratio above 0.7 or (ii) a language match threshold below 85% according to the lang-id tool (Lui and Baldwin, 2011), and (iii) duplicate sentence pairs.

4.3 NMT Training

We developed our NMT systems using Fairseq (Ott et al., 2019). We adopt the Transformer-Base implementation (BASE) with a few modifications following the architecture setup of Philip et al. (2019) and Dhar et al. (2020). These modifications include: setting both encoder and decoder layers to 6, embedding dimensions to size 1024 and number of attention heads to 8. Training is performed using batches of 4k tokens, using a label-smoothed cross entropy loss. The hidden layers are of 1024 dimensions and layer normalization is applied before each encoder and decoder layer. Dropout is set to 0.001 and weight decay to 0.2. Our loss function is cross-entropy with label smoothing of 0.3. The models are trained for a maximum of 100 epochs with early stopping criterion set to 5.

4.4 Dictionary Size

The segmentation algorithms are trained on the training data described in Section 4.1. We experiment with the following subword dictionary sizes: 1k, 5k, 10k, 15k, 20k, 30k, 40k and 50k. In all experiments, we learn separate subword dictionaries for the source and target languages, for two reasons: (i) LMVR is a linguistically motivated morphology learning algorithm that models the composition of a word based on the transitions between different morphemes and their categories. Therefore, training jointly on two languages would not be a principled choice. (ii) Prior studies such as (Dhar et al., 2020) have reported better translation scores for English-Tamil using SP models that were separately trained on the source and target sides.

5 Results

The NMT systems are evaluated and tested on the official development and test sets, respectively from WAT21. These evaluation sets are sourced from the PMIndia dataset (Haddow and Kirefu, 2020). During validation, models are evaluated by BLEU on the segmented data, whereas final test scores are computed on the un-segmented and detokenized sentences (de-tokenization is performed with the Indic NLP library tool). In addition to BLEU (Papineni et al., 2002), we also report on CHRF score (Popović, 2015), which is based on character n-grams and is therefore more suitable to assess translation quality in morphologically complex languages.[4] We report the macro-averaged

[1] http://anoopkunchukuttan.github.io/indic_nlp_library/

[2] https://github.com/google/sentencepiece

[3] https://github.com/d-ataman/lmvr

[4] We compute BLEU scores with SacreBLEU (Post, 2018), and CHRF scores with chrF++.py https://github.com/

Target Language	Dictionary Size	BLEU		CHRF		Jaccard Similarity (%)	
		SP	LMVR	SP	LMVR	Types	Tokens
Kannada	1k	10.4	6.2	48.3	40.6	17.0	2.5
	5k	13.0	5.9	**50.2**	40.7	14.8	0.6
	10k	**13.9**	6.8	49.6	42.8	13.1	0.4
	15k	13.4	6.4	48.8	41.8	10.7	0.3
	20k	13.0	7.3	48.3	43.4	10.6	0.3
	30k	12.6	6.6	47.4	42.4	10.1	0.2
	40k	12.3	7.4	46.5	43.9	9.5	0.2
	50k	12.0	6.8	46.0	42.7	9.0	0.2
Malayalam	1k	8.1	8.8	47.4	46.1	15.6	3.3
	5k	11.2	12.6	52.3	50.5	16.6	1.3
	10k	14.6	15.9	55.3	50.5	14.2	0.8
	15k	17.0	18.6	57.9	54.9	14.2	0.7
	20k	19.2	19.7	60.1	55.2	12.0	0.6
	30k	23.4	23.8	63.6	58.3	11.8	0.5
	40k	24.5	27.3	**63.7**	60.2	11.3	0.5
	50k	24.4	**28.5**	63.6	60.9	11.3	0.5
Tamil	1k	10.4	8.1	48.3	45.7	16.7	2.4
	5k	13.2	8.2	50.6	46.2	15.7	0.6
	10k	15.6	10.0	51.8	48.7	14.2	0.3
	15k	20.1	10.9	53.6	49.1	11.7	0.2
	20k	21.8	12.4	54.5	50.0	11.8	0.2
	30k	23.8	11.3	55.3	49.2	11.6	0.2
	40k	22.8	10.5	54.0	48.8	11.2	0.2
	50k	**27.3**	9.1	**55.9**	47.3	10.8	0.2
Telugu	1k	5.3	11.8	40.7	45.9	16.8	4.5
	5k	5.6	10.8	44.6	43.5	17.8	1.6
	10k	6.2	12.8	45.4	45.6	15.3	1.1
	15k	10.4	14.1	50.1	47.6	15.7	1.0
	20k	11.1	23.7	50.8	54.7	13.7	0.7
	30k	14.1	23.8	54.0	58.3	14.2	0.7
	40k	18.6	18.8	58.1	50.7	14.2	0.7
	50k	19.3	**24.5**	**59.4**	54.6	14.1	0.6

Table 4: BLEU and CHRF scores for English-to-X NMT, using different segmenters and varying subword vocabulary size. SP refers to the purely statistical SentencePiece segmenter, LMVR to Linguistically Motivated Vocabulary Reduction. Dictionary size refers to the size of both the source and target subword dictionaries. Rightmost columns show the Jaccard similarity (percentage) for the types and tokens from the segmenter outputs.

document level F3-score.Results are presented in Table 4.

SP clear winner for Kannada and Tamil: SP presented the highest BLEU and CHRF scores for Kannada and Tamil. When we compare the best systems for both SP and LMVR, large differences are observed. For Kannada differences of +6 BLEU and +7.4 are observed and for Tamil the differences are +14.9 for BLEU and +5.9 for CHRF.

Mixed results for Telugu and Malayalam: However, we find no clear winner for the other two languages. When observing only BLEU scores, LMVR appears to have the upper hand, with an improvement of +2.8 BLEU and +4.5 BLEU for Malayalam and Telugu, respectively. However the results are flipped when we look at the CHRF scores. SP systems here report higher scores, with

m-popovic/chrF

+3.5 improvement in Malayalam and +1.1 for Telugu. Given the morphological richness of our target languages, we take CHRF as the more reliable score, and conclude that the purely statistical segmenter SP is a better choice for translation into Dravidian languages in our setup.

Larger dictionary sizes better: When observing the effect of the dictionary size, we find that the size 50k gives the highest BLEU scores for Malayalam, Tamil and Telugu. This is in contrast with studies such as (Philip et al., 2019; Sennrich and Zhang, 2019) who suggest to use a smaller dictionary size for low-resource settings. For these language pairs, we see a steady increase in BLEU and CHRF as we increase the dictionary size. For Kannada, the best results are obtained for much smaller dictionary sizes, but in contrast with the other three languages, the differences between the scores for other dictionary sizes is much smaller. For instance, looking at the CHRF scores of SP, the numbers decrease from 48.3 to 46.0, whereas for instance for Malayalam, these numbers range from 47.4 to 63.6.

Kannada hardest to translate: When comparing more in general translation difficulty across target languages, Kannada appears to be the most challenging language by far. A possible explanation for this difference is the genre distribution of our datasets (cf. Table 2): While the test sets are from PMIndia (a mixture of background information, news and speeches), the majority of our Kannada training data consists of religion related texts. Another possible confounding factor is that we based our NMT configuration on prior work that focused only on English-Tamil (Philip et al., 2019; Dhar et al., 2020), and this may be sub-optimal for the other Dravidian languages despite the similar training data size.

6 Analysis

6.1 Different Subtokens generated

Table 4 presents the Jaccard similarity (JS) between the segmenter outputs between LMVR and SP. The outputs are either the types (dictionaries) or the tokens in the training sentences. A JS of 0 denotes that none of the subwords were the same in the sentences being compared, while a score of 100 denotes a complete match (i.e, they are identical). As visible from the scores, though there is some sharing of types between the segmenters (ranging from 9-17%), there is no such sharing of subwords in the training data, with a maximum JS score of only around 4% for the smallest dictionary sizes. In fact, these values reduce even further as the dictionary size are increased. For the largest dictionary size (50k), almost no subtoken sharing occurs.

6.2 Effect of Unknown Subwords

We carried out an analysis on the effect of unknown subwords found in the development set after the application of a given segmentation algorithm.We present these statistics in Figure 1. Few details stand out:

High percentage of unknown subwords in Kannada with LMVR While development sets encoded with SP reported the lowest percentage of unknowns, it is the complete opposite for the ones encoded with LMVR (0.2% vs 15% on average). This could have played a role in the lowest CHRF scores achieved by the LMVR systems on Kannada.

LMVR sensitive to dictionary size This is observed in particular for Kannada and Malayalam, where the increase in dictionary size leads to higher numbers of unknown subwords. Conversely for SP, increasing the dictionary size causes no major change in the number of unknowns found for these two languages. On the other hand, SP is more susceptible to the dictionary size for Tamil while Telugu, in general, does not present any such trends.

Overall we find no strong correlation between system performance and percentage of unknown subwords. By contrast, and quite surprisingly so, our best NMT systems for Malayalam, Tamil and Telugu are those with larger dictionary sizes and higher percentage of unknowns in the development set.

6.3 Effect of subword lengths

We also looked at the effect of the segmenter on the subword length. Given a language and segmenter, we calculate the average length of a subword (in characters) for the training sets. In Figure 2 we plot the distribution of the average subword lengths for all our settings. Few observations are apparent,

- For every language and dictionary size, LMVR results in shorter subwords. Taking dictionary size of 50k as an example, the dif-

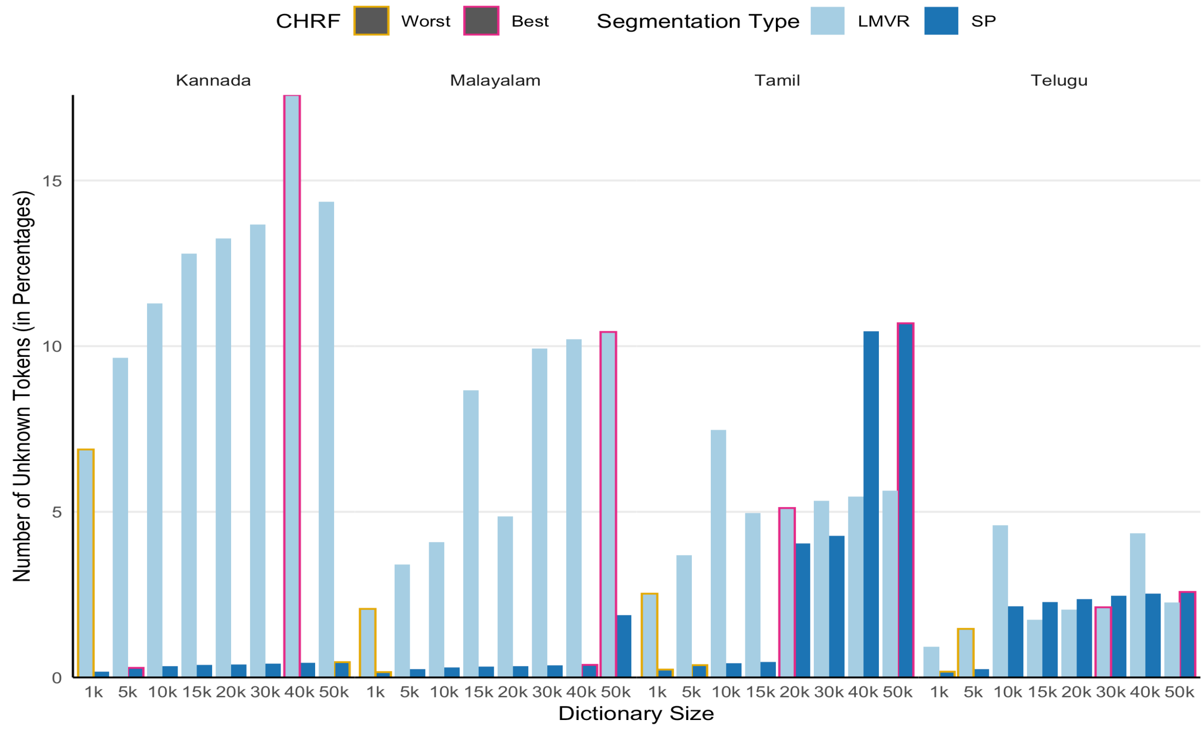

Figure 1: Number of unknown tokens (in percentages) in the development set vs Dictionary size for each language and segmentation type. Also systems that reported the lowest and highest CHRF scores (on the development set) for each language and segmentation are marked.

ference between LMVR and SP ranges from 1.2 for Malayalam to 1.7 for Tamil.

- As the dictionary size increases, we see the distributions spreading out. As the dictionary size decreases, the distributions become more centered. This is particularly seen for LMVR. As the dictionary size increases, the distributions of the SP systems spread out more than their LMVR counterparts.

- While it makes sense that the average subword length increases as we increase the dictionary size (from 3 to 5), the apparent widening in the difference between SP and LMVR is not so easily explained.

In the end however, we find no discernible connection between the subword length and the performance of a segmenter. Across all languages, we see similar trends of how the distrubtions change, but this does not seem to affect the translation quality, as seen in the difference in the CHRF scores.

7 Conclusion

We presented our work on Neural Machine Translation from English into four Dravidian languages (Kannada, Malayalam, Tamil and Telugu). Several experiments were carried out to find out whether a linguistically motivated subword segmenter (LMVR) is more suitable than a purely statistical one (SentencePiece) for translating into the morphologically complex Dravidian languages, while using a Transformer architecture. While BLEU results were mixed on Malayalam and Telugu, CHRF scores clearly suggest that SentencePiece remains the best option for all of our tested language pairs.

We also found interesting differences among the four target languages. Though they all belong to the same language family and share various linguistic phenomena, they are different with respect to source/target token ratio (Table 3), and the rate of unknown subwords in the development set (Figure 1). Whether this is due to linguistic characteristics or to genre differences in the training corpora remains hard to gauge.

Finally, we invite future researchers to carry out research on Dravidian languages, especially Kannada. Compared to the plethora of work found for other languages, the work on Dravidian languages is lagging behind. As our results show, there remains a large space for improvements, particularly

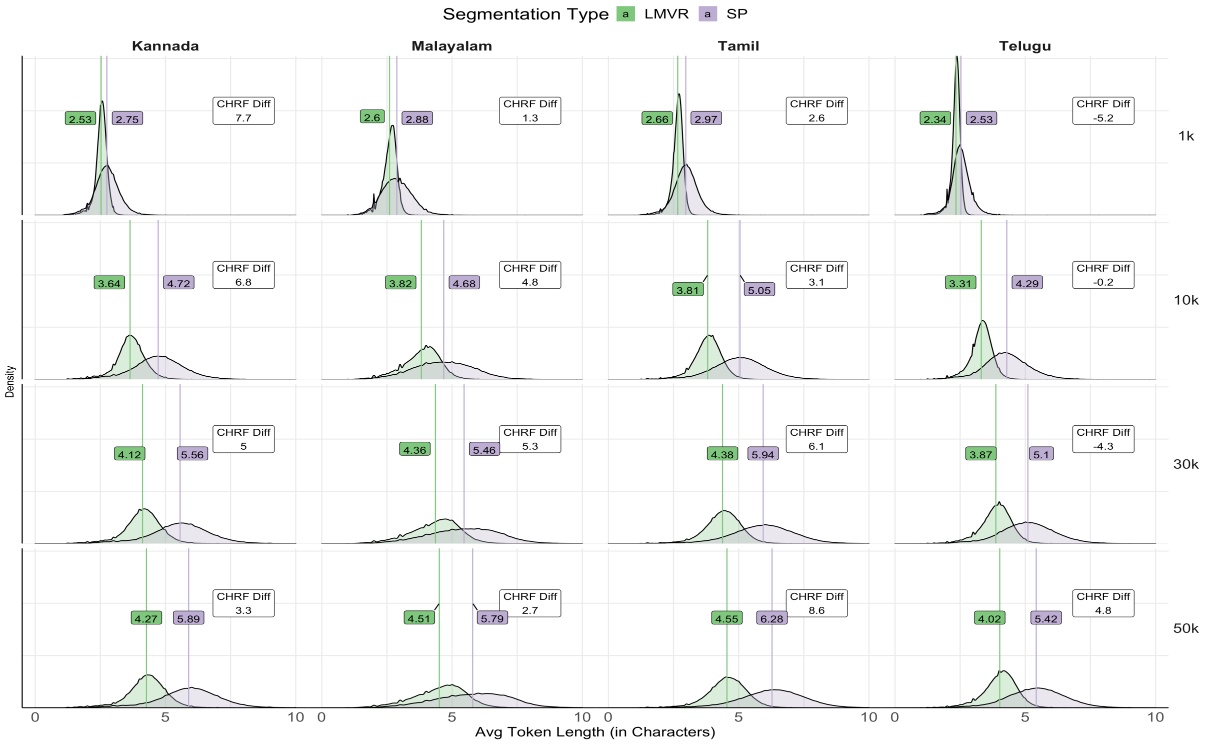

Figure 2: The Probability density function plot showing the distribution of the average subword length for a given segmenter and language on the training sets. The colored boxes denote the mean of the respective distributions. Also included are the differences in the CHRF scores between SP and LMVR.

when translating *into* these languages.

References

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Ancy Anto and K. Nisha. 2016. Text to speech synthesis system for english to malayalam translation. *2016 International Conference on Emerging Technological Trends (ICETT)*, pages 1–6.

Duygu Ataman and Marcello Federico. 2018. Compositional representation of morphologically-rich input for neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 305–311, Melbourne, Australia. Association for Computational Linguistics.

Duygu Ataman, Matteo Negri, Marco Turchi, and Marcello Federico. 2017. Linguistically motivated vocabulary reduction for neural machine translation from turkish to english. *The Prague Bulletin of Mathematical Linguistics*, 108(1):331 – 342.

Bharathi Raja Chakravarthi, Mihael Arcan, and John P. McCrae. 2019. Comparison of Different Orthographies for Machine Translation of Under-Resourced Dravidian Languages. In *2nd Conference on Language, Data and Knowledge (LDK 2019)*, volume 70 of *OpenAccess Series in Informatics (OASIcs)*, pages 6:1–6:14, Dagstuhl, Germany. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Himanshu Choudhary, Shivansh Rao, and Rajesh Rohilla. 2020. Neural machine translation for low-resourced Indian languages. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3610–3615, Marseille, France. European Language Resources Association.

Mathias Creutz and Krista Lagus. 2002. Unsupervised discovery of morphemes. In *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning*, pages 21–30. Association for Computational Linguistics.

Raj Dabre and Abhisek Chakrabarty. 2020. NICT‘s submission to WAT 2020: How effective are simple many-to-many neural machine translation models? In *Proceedings of the 7th Workshop on Asian Translation*, pages 98–102, Suzhou, China. Association for Computational Linguistics.

Raj Dabre, Anoop Kunchukuttan, Atsushi Fujita, and Eiichiro Sumita. 2018. NICT's participation in WAT 2018: Approaches using multilingualism and recurrently stacked layers. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation: 5th Workshop on Asian Translation:*

5th Workshop on Asian Translation, Hong Kong. Association for Computational Linguistics.

Sandipan Dandapat and Christian Federmann. 2018. Iterative data augmentation for neural machine translation: a low resource case study for english–telugu. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, pages 287–292, Alacant, Spain.

Prajit Dhar, Arianna Bisazza, and Gertjan van Noord. 2020. Linguistically motivated subwords for English-Tamil translation: University of Groningen's submission to WMT-2020. In *Proceedings of the Fifth Conference on Machine Translation*, pages 126–133, Online. Association for Computational Linguistics.

Philip Gage. 1994. A new algorithm for data compression. *C Users J.*, 12(2):23–38.

Ulrich Germann. 2001. Building a statistical machine translation system from scratch: How much bang for the buck can we expect? In *Proceedings of the ACL 2001 Workshop on Data-Driven Methods in Machine Translation*.

Stig-Arne Grönroos, Sami Virpioja, Peter Smit, and Mikko Kurimo. 2014. Morfessor FlatCat: An HMM-based method for unsupervised and semi-supervised learning of morphology. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1177–1185, Dublin, Ireland.

Barry Haddow and Faheem Kirefu. 2020. Pmindia – a collection of parallel corpora of languages of india.

Oskar Kohonen, Sami Virpioja, and Krista Lagus. 2010. Semi-supervised learning of concatenative morphology. In *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, pages 78–86, Uppsala, Sweden. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Keerthi Lingam, E. Ramalakshmi, and Srujana Inturi. 2014. English to telugu rule based machine translation system: A hybrid approach. *International Journal of Computer Applications*, 101(2):19–24. Full text available.

Marco Lui and Timothy Baldwin. 2011. Cross-domain feature selection for language identification. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 553–561, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Kavya Manohar, A. R. Jayan, and Rajeev Rajan. 2020. Quantitative analysis of the morphological complexity of malayalam language. In *Text, Speech, and Dialogue*, pages 71–78, Cham. Springer International Publishing.

Atul Kr. Ojha, Koel Dutta Chowdhury, Chao-Hong Liu, and Karan Saxena. 2018. The RGNLP machine translation systems for WAT 2018. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation: 5th Workshop on Asian Translation: 5th Workshop on Asian Translation*, Hong Kong. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Kishore Papineni, S. Roukos, T. Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Jerin Philip, Shashank Siripragada, Upendra Kumar, Vinay Namboodiri, and C V Jawahar. 2019. Cvit's submissions to wat-2019. In *Proceedings of the 6th Workshop on Asian Translation*, pages 131–136, Hong Kong, China. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

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 (MTPIL-2012)*, pages 113–122.

Akshai Ramesh, Venkatesh Balavadhani Parthasa, Rejwanul Haque, and Andy Way. 2020. An error-based investigation of statistical and neural machine translation performance on Hindi-to-Tamil and English-to-Tamil. In *Proceedings of the 7th Workshop on Asian Translation*, pages 178–188, Suzhou, China. Association for Computational Linguistics.

Sukanta Sen, Kamal Kumar Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2018. IITP-MT at WAT2018: Transformer-based multilingual indic-English neural machine translation system. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation: 5th Workshop* *on Asian Translation: 5th Workshop on Asian Translation*, Hong Kong. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 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)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

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

S Sreelekha and P Bhattacharyya. 2017. A case study on english-malayalam machine translation. *ArXiv*, abs/1702.08217.

S Sreelekha and P Bhattacharyya. 2018. Morphology injection for English-Malayalam statistical machine translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Sanford B Steever. 2019. *The Dravidian Languages*. Routledge.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

K Deepthi Yadav and L Lingam. 2017. Rule based machine translation of complex sentences from english to telugu. *International Journal of Research*, 4(9):790–800.

Zhengzhe Yu, Zhanglin Wu, Xiaoyu Chen, Daimeng Wei, Hengchao Shang, Jiaxin Guo, Zongyao Li, Minghan Wang, Liangyou Li, Lizhi Lei, Hao Yang, and Ying Qin. 2020. HW-TSC's participation in the WAT 2020 indic languages multilingual task. In *Proceedings of the 7th Workshop on Asian Translation*, pages 92–97, Suzhou, China. Association for Computational Linguistics.