Hindi to Dravidian Language Neural Machine Translation Systems

Vijay Sundar Ram and Sobha Lalitha Devi

AU-KBC Research Centre,
MIT Campus of Anna University, Chennai 60044
sobha@au-kbc.org

Abstract

Neural machine translation (NMT) has achieved state-of-art performance in high-resource language pairs, but the performance of NMT drops in low-resource conditions. Morphologically rich languages are yet another challenge in NMT. The common strategy to handle this issue is to apply sub-word segmentation. In this work, we compare the morphologically inspired segmentation methods against the Byte Pair Encoding (BPE) in processing the input for building NMT systems for Hindi to Malayalam and Hindi to Tamil, where Hindi is an Indo-Aryan language and Malayalam and Tamil are south Dravidian languages. These two languages are low resource, morphologically rich and agglutinative. Malayalam is more agglutinative than Tamil. We show that for both the language pairs, the morphological segmentation algorithm out-performs BPE. We also present an elaborate analysis on translation outputs from both the NMT systems.

1 Introduction

Machine translation has improved extensively using deep neural networks with the utilization of large dataset and high computational capacities. The successful works in Neural Machine Translation (NMT) started with the encoder-decoder based architecture presented by Kalchlorenner and Blunsom (2013), Sutskever et. al. (2014), and Cho et. al (2014). Sutskever et al (2014) built NMT system using Long short Term memory (LSTM) to overcome the fixed-length vector constraint in the previous architecture. Bahdanu et. al. (2015) introduced the attention mechanism, where bidirectional recurrent neural network (RNN) consisting of forward and backward RNN was used to focus around the word. This attention mechanism was simplified by considering the hidden states at the top layer of both encoder and decoder by Luong et. al. (2015). Transformer, an architecture where encoder and decoder completely relying on the attention machines was presented by Vaswani et. al. (2017).

Though these NMT systems have achieved a state-of-art performance in high-resource, closely related languages, its performance drop significantly in low-resource and morphologically rich languages. Some of the techniques employed to mitigate challenges in handling the low-resource languages are as follows; increasing the data using back translation, utilisation of phrase tables generated in SMT, leveraging the pre-trained models, combining the similar language data and using transfer learning. The morphological rich languages are handled using different sub-word segmentation techniques, which helps in reducing the vocabulary size and increasing the number of examples of each tokens. In this work, we compare the morphologically inspired segmentation methods against the Byte Pair Encoding (BPE) in processing the input for building NMT systems for Hindi (Hi) to Malayalam (ML) and Hindi to Tamil (TA), where Hindi is an Indo-Aryan language and Malayalam and Tamil are south Dravidian languages. These two languages are low-resource, morphologically rich and agglutinative.

Further the paper is organised as follows. In the following section, we present a summary on the different sub-word tokenisation works in NMT. This is followed by details on related works in

Proceedings of Recent Advances in Natural Language Processing, pages 1143–1150
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_121

Indian language NMT. In the third section, we describe briefly the characteristics of three languages, which highlight the challenges in building NMT systems for Hindi to Malayalam and Tamil. In section 4, we describe our experimental setup and data preparation. The result and analysis is presented in section 5. We conclude the paper with a conclusion section containing the gist of the work.

2 Related Works

The common strategy of handling the morphologically rich languages in NMT is to apply sub-word segmentation. This reduces the vocabulary size and increases the frequency of the tokens and improves the translation by handling rare words and unknown words, but it introduces grammatical errors. Sennrich et. al. (2016) presented the different word segmentation techniques which included simple character n-gram model and segmentation based on the byte pair encoding (BPE) comparison algorithm. BPE sub-word algorithm is one of the widely used sub-word tokenisation algorithm.

The other sub-word tokenisation algorithms include, WordPiece, SentencePiece, Mecab (a morphological analysed based Japanese tokeniser), Stanford Word Segmentation (a Chinese word segmentor based on Conditional Random Fields), OpenNMT Tokenizer and Moses tokenizer (normalise characters and separates punctuation from words).

There are various attempts in modifying the existing tokenization techniques and few are listed here. Wu and Zhao (2018) extended the BPE segmentation by including two other statistical measures namely accessor variety (AV) and description length gain (DLG). They evaluated it with German to English and Chinese to English translation.

Provilkov et. al. (2019) introduced BPE-dropout, where segmentation procedure of BPE was stochastically altered to produce multiple segmentations within the same fixed BPE framework.

Wang et. al. (2020) focussed on byte-level BPE (BBPE), where the text is tokenised into variable-length byte n-grams instead of character level sub-words.

Nonaka et. al. (2022) has presented a locally consistent parsing (LCP) stochastic string algorithm to achieve optimum compression instead of BPE compression, which has the drawback in generating multiple segments.

Tang et. al. (2020) performed a study on pure character based model in translating Finnish to English. They have demonstrated that the word level information is distributed over the entire character sequence and character at different position play different roles in learning linguistic knowledge.

Deguchi et. al. (2020) performed tokenisation of sentences by using sub-word units induced from bilingual sentences. Here the tokenisation of sentences is performed by considering its translation.

Nguyen et. al. (2020) proposed an approach, where the heterogeneous translation units were used to build in Russian to Vietnamese NMT. They considered linguistic characteristics of syntactic Russian and analytic Vietnamese.

Machacek et. al. (2019) compared the linguistically motivated method morfessor and derivational dictionaries based method and statistical methods such as STE and BPE in German to Czech translation. Their experiments showed the non-linguistically motivated method performed better.

In this sub section, we present a gist of the NMT works published in Indian languages. Goyal et al. (2020) has presented Hindi to English NMT, where they generalised the embedding layer of the Transformer model to incorporate linguistic features such as PoS, lemma, and morphological features. There was a significant increase in the BLEU scores. Dewangan et al. (2021) has presented an elaborate NMT experiments to understand the poor performance of the Dravidian languages compared to Indo-Aryan languages. They used Byte Pair Encoding (BPE) method to understand the BPE in Indian languages. From their study, they presented that the optimal value for BPE merge for Indian languages is between 0-5000, which is low compared to that observed for European languages.

WMT21 had a similar language task, which has boosted the research to explore the use of shared vocabulary in NMT. Laskar et. al. (2021) and Saldanha et. al. (2021) has presented their work in Tamil-Telugu translation. Mujadia et al. (2020) has presented their work in Marathi-Hindi bidirectional translation.

In the next section, we present a brief note on the characteristics of the languages considered.

3 Resources Characteristics of the Languages

As mentioned earlier, in this work, we describe the NMT experiments in Hindi to Malayalam and Hindi to Tamil translation, where Hindi is an Indo-Aryan language and Malayalam and Tamil are Dravidian languages. The three languages are similar in the following features: verb final, relatively free word order, morphologically rich in inflections. And these languages are dissimilar in agglutination. Malayalam and Tamil have agglutination and Hindi does not have. Malayalam has more agglutination than Tamil. The other differences are as follows.

Hindi and Tamil have number, gender and person agreement, whereas Malayalam does not have. Hindi is an ergative language. In the ergative constructions, finite verb has agreement with the object. Malayalam and Tamil are nominative-accusative languages.

Malayalam and Tamil have distinctive case markers, whereas in Hindi, case marker 'se' occurs as instrumental, accusative and ablative case marker. This leads to one to many in case mapping between Hindi to Malayalam and Tamil. In Hindi, plural marker is affixed to the noun and case markers are written separately. In the case of pronouns, case markers are also affixed to the pronouns. In Malayalam and Tamil, both plural markers and case markers are affixed to the nouns.

Copula verb is obligatory in Hindi and Malayalam whereas in Tamil it can be dropped.

Malayalam and Tamil has distinctive 3rd person pronouns (avan, aval, avar, athu), whereas in Hindi, 'vaha' is used for all 3rd person singular pronouns.

The clausal construction in Hindi varies with Malayalam and Tamil. In Hindi, the clausal constructions are introduced by relative-correlatives such as (jo-vo, agar-tho, jisa-usa, jisne-usne, jab-thab etc). In Malayalam and Tamil, the clausal constructions are introduced by non-finite verbs namely, relative participle verb, conditional, infinite verb and verbal participle verb. It is further explained with the following example 1.

Ex 1:
HI: agar barish ayege tho paani
rain(N) come(V)+Future water(N)
milegaa.
get(V)+Future
Here 'agar' and 'tho' are the relative-correlative

ML: mazha peythaal, vellam
rain(N) rain(V)+cond water(N)
labikkum.
get(V)+future

TA: mazhai peythaal, thanneer
rain(N) rain(V)+cond water(N)
kidaikkum.
get(V)+future
(If it rain, we will get water.)

In the above example 1, conditional sentence is presented in Hindi, Malayalam and Tamil. In Hindi the conditional clause is introduced with the relative-correlative 'agar-tho', whereas in Malayalam and Tamil it is introduced by the non-finite verb using the suffix '-aal'.

Negation in verb phrase in Hindi varies with Malayalam and Tamil. In Hindi, the negation occurs before the finite verb and in Malayalam and Tamil, it occurs as an auxiliary verb. Consider the following example 2.

Ex 2:
HI: vaha nahi aaya.
He(Pn) not(neg) come(V)+past+3sc
ML: avan vannilla
He(Pn) come(V)+INF+aux (neg)
TA: avan varavillai (vara+illai)
He(Pn) come(V)+INF+aux (neg)

In example 2, the difference in construction of negation verb in Hindi and Malayalam and Tamil is clearly seen with the position of the negation.

These variations between Hindi and Malayalam and Tamil in clausal structure, case markers, pronouns and verb construction introduce challenge in Hindi to Dravidian language translation. In the next section, we describe the corpus and the experimental setup.

4 Experiment

In this section, we discuss about the details of the parallel dataset, experimental setup for developing Hindi to Malayalam and Hindi to Tamil NMT systems and data preparation for three different experiments.

4.1 Dataset

We have used Hindi-Malayalam and Hindi-Tamil corpus, built using the manually translated Swayam course lectures. Swayam is a massive online course platform by Government of India, which offers variety of courses in various domains such as Engineering, Business Management, Humanities, Programming, Business, Mathematics, Science and Technology, Health, Law etc. We have used parallel sentences from the lectures of 52 courses from different domains, namely, Science and Technology, Food Processing technology, Information Technology, Business Management, Plant pathology and Law. The statistics of the corpus is given the tables below.

S.No	Details	Hindi (Source)	Malayalam (Target)
1	Number of Sentences	158318	158318
2	Number of Words	3421259	1932170
3	Number of unique words	98945	257848
4	Maximum Length of a Sentence (words)	80	61

Table 1: Statistics of Hindi-Malayalam Corpus.

S.No	Details	Hindi (Source)	Tamil (Target)
1	Number of Sentences	165172	165172
2	Number of Words	3565959	2214121
3	Number of unique words	104613	186413
4	Maximum Length of a Sentence (words)	80	66

Table 2: Statistics of Hindi-Tamil Corpus.

Table 1 has the statistics of the Hindi-Malayalam parallel corpus; Table 2 has the statistics of the Hindi-Tamil parallel corpus. In the both tables 1 and 2, in the second row, the number of words in Hindi is one and half times more than the number of words in Malayalam and Tamil. In table 1, the number of unique words in Malayalam is one and half times more than the unique words in Hindi. In table 2, the number of unique words in Tamil is one and half times more than the unique words in Hindi. The information in these two rows clearly shows the morphological richness and high agglutination in Malayalam and Tamil, which make the NMT training a challenging task. The difference in the number of unique words in Malayalam and Tamil shows the high agglutination in Malayalam compared to Tamil.

4.2 Experiment Setup

We used OpenNMT-py toolkit for developing the Hindi-Malayalam and Hindi-Tamil NMT systems. The architecture of the model used is a Bi-direction RNN Encoder-Decoder with attention mechanism. The gated units used are Bi-LSTM. We used Loung attention mechanism. The model was trained till 2,00,000 training steps. The details of the parameters for NMT training is below.

Embedding size: 500; RNN for encoder and decoder: bi-LSTM; Bi-LSTM dimension: 500; encoder - decoder layers: 2; Attention: Luong; label smoothing: 1.0; dropout: 0.30; Optimizer: Adam

With the above setup, we trained three different NMT models by varying the training corpus. The three different experiments were, 1) Word Level, 2) Sub-word segmented data using Byte pair Encoding (BPE), 3) Word Segmentation using Morphological analyser

From the parallel dataset, 3000 sentences were randomly chosen for fine-turning the NMT training and another 1000 sentences were randomly chosen for testing. The same set of training, validation and test data were used for all the three experiments.

4.3 Data Preparation

The data was processed in three different methods as described below:

Word Level: The sentences in the three languages where tokenised with a white space and punctuations were separated from the words. The processed sentences were used for NMT training in both Hindi to Malayalam and Hindi to Tamil NMT training.

BPE: Byte Pair Encoding (BPE) proposed by Sennrich et al. (2016) was applied to the tokenised data. We used 3000 as BPE merge value for Malayalam and Tamil and for Hindi we used 5000 as BPE merge value.

Morph-Seg: The sentences in all the three languages, namely, Hindi, Malayalam and Tamil are processed with morphological analyser to split the words into root and suffix. The words in the sentence are replaced by the morphologically segmented root and suffixes to prepare the data. Morphological analysers built using paradigm and Finite state automata based approach was used for the three languages. For Hindi, we used morphological analyser available in the following link, https://ltrc.iiit.ac.in/morph/index.htm. Malayalam morphologically analyser used in present in Lakshmi and Sobha (2013). Tamil morphological analyser used is present in Sobha et. al. (2013).

5 Results and Analysis

We evaluated the translations from the three NMT models for both Hindi to Malayalam and Hindi to Tamil using BLEU score (Papineni et al. 2002). We used Sacre-bleu python library to calculate the BLEU scores. The results are presented in Table (3).

S.No	Details	Hindi to Malayalam (BLEU Score)	Hindi to Tamil (BLEU Score)
1	Word-Level	5.519	13.413
2	BPE	10.866	17.492
3	Morph-Seg	17.983	24.642

Table 3: BLEU Score for Hindi to Malayalam and Hindi to Tamil from different models

The BLEU scores show that the morphological segmentation has significantly improved the translation in both Hindi-Malayalam and Hindi-Tamil.

On analysis of the translation output from the three different experiments in both Hindi to Malayalam and Hindi to Tamil, our observations are as follows,

Word-Level: Many named entities, technical words and verb phrases occurred as unknown word (<unk>).

BPE: Translated sentences were complete but most of these translations were not the exact translation.

Translations convey a different sense due to the choice of the verb generation.

There were also words omitted in the translation.

Technical words and rare words were handled, but there were errors in it.

Morph-Seg: Clausal sentences were translated correctly than the other two systems.

Verb phrase generation was exact, though there were errors.

More closer to exact translation, but there were unknown words.

Technical words, Named Entities and rare words occurring as <unk> is the problem, but it is comparatively less than the word-level system.

We have explained the translation output with examples in the further part of this section.

Ex 3.a (HI to ML):

Hindi-Input: लोग, दृष्टिकोण अनुमानों पर सरल कार्रवाई कर सकते हैं.

(People can take simple actions on attitude projections.)

Malayalam Translations:

Word-Level: ആളുകൾക്ക് <unk> ലളിതമായ നടപടിയെടുക്കാൻ കഴിയും.

BPE: ആളുകൾക്ക് മനോഭാവങ്ങൾ ലളിതമാക്കാൻ കഴിയും.

(People can take simple actions on attitude.)

Morph-Seg: ആളുകൾക്ക് മനോഭാവം കണക്കിലെടുത്ത് ലളിതമായ പ്രവർത്തനം നടത്താൻ കഴിയും.

(People can take simple actions on attitude projections.)

Ex 3.b (HI to TA)

Hindi-Input: म्यूटेशन आनुवंशिक में मिल सकते हैं .

(Mutations can be found in genetics.)

Tamil Translations:

Word-Level: பிறழ்வுகள் மரபணு <unk> இருக்கலாம்.

BPE: பிறழ்வுகள் மரபணு மரபணுவில் இருக்கலாம் .

(Mutations can occur in the genetics genetics.)

Morph-Seg: பிறழ்வுகள் மரபணுவில் கிடைக்கலாம்.

(Mutations can be found in genetics.)

Ex 3.a has Hindi to Malayalam translation and Ex 3.b has Hindi to Tamil translation. The word-level translation has <unk>. Though BPE and Morph-Seg translation outputs are proper sentences. Morph-Seg translation has exact translation. BPE translation in both languages has a different sense from the source sentence.

Ex4: Clausal Sentence

Hindi-Input: इतिहास उस दौर से शुरू होता है जब लोग लिखने की कला जानते थे.

(History begins from the time when people knew the art of writing.)

Tamil Translations:

Word-Level: மக்கள் எழுதும் கலையை மக்கள் <unk> வரலாறு தொடங்குகிறது.

BPE: மக்கள் எழுதும் கலத்திலிருந்து தொடங்கும்போது வரலாறு தொடங்குகிறது.

(History begins from the cell when people writing.)

Morph-Seg: மக்கள் எழுதும் கலையை மக்கள் அறிந்த போது வரலாறு தொடங்குகிறது.

(History begins from the time when people knew the art of writing.)

In Ex 4, the Hindi sentence has a relative participle clause. The clause construction was correctly translated by the Morp-Seg system. It has generated the relative participle verb, 'அறிந்த' (aRintha).

Ex 5: Sentence with series of NPs:

Hindi-Input: ग्राउंड रखरखाव उपकरण, जैसे लॉन मोवर, रोलर्स, लाइम पाउडर मशीन, मार्कगि मशीन, घास काटने वाली तलवारें, दरांती, श्रब मास्टर, कटर आदि.

(Ground Maintenance Equipments like Lawn Mower, Rollers, Lime Powder Machine, Marking Machine, Mower, Sickle, Shrub Master, Cutter etc.)

Tamil Translations:

Word-Level: தரை பராமரிப்பு சாதனங்கள், பாலைவனங்கள், <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> போன்றவை .

BPE: தரை பராமரிப்பு சாதனங்கள், ஒவ்வாமைகள், லீக் தூள் இயந்திரங்கள், புல்வெளி இயந்திரங்கள், புல்வெளிகள், ஆப்பிரிக்கா, ஆப்பிரிக்கா, கேரட் மற்றும் பலர் போன்ற தரை பராமரிப்பு சாதனங்கள்.

(Ground care equipment like Ground care equipment, Allergies, Leek powder machines, Lawn machines, Lawns, Africa, Africa, Carrot and many others.)

Morph-Seg: புல்வெளிகள், புல்வெளிகள், லாரிகள், சுண்ணாம்பு பொடிகள், இயந்திரங்கள், இயந்திரங்கள், இயந்திரங்கள், இயந்திரங்கள், வெட்டும் போன்ற தரை பராமரிப்பு உபகரணங்கள் அகும்.

(Ground care equipment like such as Lawn care equipment such as lawnmowers, lawnmowers, trucks, lime powders, machines, machines, machines, machines, mowers etc.)

In example 5, the Hindi sentence has series of noun phrases. The three systems gave improper translation for this sentence. The Word-Level system gave series of <unk>, the BPE has generated output with many words which are not in the input sentences such as ‘,Africa’, ‘Carrot’ etc. Morph-Seg, most of the noun phrases was partially translated, and only the head of the NPs were translated.

The following two examples demonstrate, technical words handled by BPE system. The first example (Ex.6.a) has the correct word replacement and the second example Ex.6.b has wrong word replacement.

Ex.6.a

Hindi-Input: कुछ स्यूडोमोनाड्स समस्या पैदा कर सकते हैं.

(Some pseudomonads can cause problems.)

BPE Tamil translation: சில சூடோமோனாட்கள் சிக்கலை உருவாக்கலாம்.

(Some pseudomonads may also develop problems.).

Ex.6.b:

Hindi-Input: 5% मैलाथियान, 1% लिडन ये सभी चूहे के विनाश के लिए प्रभावी हैं.

(5% malathion, 1% lindane all these are effective for rat extermination.)

BPE Tamil translation: 5% மில்லியன்கள், 1% இணைப்பு இந்த சுண்ணாம்பு அழிவுக்கு பயனுள்ளதாக இருக்கும்.

(5% millions, 1% patch is useful for this lime destruction.)

Examples 6.a and 6.b has Hindi to Tamil translations. In Ex.6 the word, ‘स्यूडोमोनाड्स' (pseudomonads) has been translated correctly to 'சூடோமோனாட்கள்' (seudomonad + plural suffix) with plural suffix. Whereas in example 6.b, there are two technical terms ‘malathion’ and ‘lindane’ in the Hindi sentences, in the translation, the word ‘malathion’ has been wrongly translated to ‘மில்லியன்கள்’ (millions) and the ‘lindane’ is missing in the translation. And ‘rat’ has occurred as ‘lime’.

From the above analysis, we observed that morph-segmentation of data in both Hindi to Malayalam and Hindi to Tamil has improved the

translation. The translation of rare words occurs as <unk> has to be corrected.

6 Conclusion

We have presented our experiments in building Neural Machine Translation system for Hindi to Malayalam and Hindi to Tamil, where we compare the morphologically inspired segmentation methods against the Byte Pair Encoding (BPE) in processing the input for building NMT systems. Hindi is an Indo-Aryan language and Malayalam and Tamil are Dravidian languages. All the three languages are morphologically rich language. Malayalam and Tamil have agglutination. We have briefly explained the characters of these languages. We have compared the translation output from the Word-Level (base line) system and NMT systems trained with these two different sub-word processed data. Word-Level system had unknown words and verb generation was not proper. BPE system translation outputs were complete sentences but these translations were not exact translation. The sense of the sentences varied from the source sentence. BPE system handled unknown words. It also had errors. Translation from Morph-Seg systems had a significantly high BLEU score. The sense of translated sentences was close to the source sentences. Unknown words are a challenge in this system, but it is comparatively less than the Word-Level system.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *In Proceedings of 3rd International Conference on Learning Representations, ICLR 2015*, San Diego, United States.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *In Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*.

Hiroyuki Deguchi, Masao Utiyama, Akihiro Tamura, Takashi Ninomiya and Eiichiro Sumita. 2020. Bilingual Subword Segmentation for Neural Machine Translation, *In Proceedings of 28th International Conference on Computational Linguistics*, Barcelona, Spain, pages 428–74297

Shubham Dewangan, Shreya Alva, Nitish Joshi, Pushpak Bhattacharyya. 2021. Experience of neural machine translation between Indian languages. *Machine Translation* 35, 71–99

Dominik Macháček, Jonáš Vidra, Ondřej Bojar. 2018. Morphological and Language-Agnostic Word Segmentation for NMT. *In Proceedings of the 21st International Conference on Text, Speech and Dialogue—TSD 2018*, pages 277-284, Springer-Verlag, Cham, Switzerland, ISBN 978-3-030-00794-2

Goyal, Vikrant and Kumar, Sourav and Sharma, Dipti Misra. 2020. Efficient Neural Machine Translation for Low-Resource Languages via Exploiting Related Languages. *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 162–168

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. *In Proceedings of the ACL Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1700–1709. Association for Computational Linguistics.

Guillaume Klein, François Hernandez, Vincent Nguyen, and Jean Senellart. 2020. The opennmt neural machine translation toolkit: 2020 edition. *In Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (AMTA 2020)*, pages 102–109.

Lakshmi Sridhar, and Sobha Lalitha Devi. 2013. Malayalam Morphological Analyser. *In proceedings of International Seminar on Current Trends in Dravidian Linguistics*, pages 27–29

Sahinur Rahman Laskar, Bishwaraj Paul, Prottay Kumar Adhikary, Partha Pakray, Sivaji Bandyopadhyay. 2021. Neural Machine Translation for Tamil–Telugu Pair. *In Proceedings of the Sixth Conference on Machine Translation (WMT)*, pages 284–287

Minh-Thang Luong Hieu Pham Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. *In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.

Vandan Mujadia and Dipti Sharma. 2020. NMT based Similar Language Translation for Hindi - Marathi. *In Proceedings of the Fifth Conference on Machine Translation*, pages 414–417, Online. Association for Computational Linguistics.

Keita Nonaka, Kazutaka Yamanouchi, Tomohiro I, Tsuyoshi Okita, Kazutaka Shimada and Hiroshi Sakamoto. 2022. A Compression-Based Multiple Subword Segmentation for Neural Machine

Translation, *Electronics* 2022, 11(7), 1014; https://doi.org/10.3390/electronics11071014

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. *In Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. 2020. BPE-Dropout: Simple and Effective Subword Regularization. *In Proceedings of 58th Annual Meeting of the Association for Computational Linguistics*, pages 1882--1892

Richard Saldanha, Ananthanarayana V. S, Anand Kumar Madasamy, and Parameswari Krishnamurthy. 2021. NITK-UoH: Tamil-Telugu Machine Translation Systems for the WMT21 Similar Language Translation Task. *In Proceedings of the Sixth Conference on Machine Translation (WMT)*, pages 299–303

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.

Sobha Lalitha Devi, Marimuthu, K, Vijay Sundar Ram, Bakiyavathi, T and Amudha, K. 2013. Morpheme Extraction in Tamil using Finite State Machines. *In Proceedings of Morpheme Extraction Task at FIRE*.

lya Sutskever, Oriol Vinyals, Quoc V. Le. 2014. Sequence to sequence learning with neural networks. *In Advances in Neural Information Processing Systems (NIPS 2014*), 3104–3112

Gongbo Tang, Rico Sennrich, and Joakim Nivre. 2020. Understanding Pure Character-Based Neural Machine Translation: The Case of Translating Finnish into English, *In Proceedings of 28th International Conference on Computational Linguistics*, pages 4251–4262

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. *In Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, CA, USA, pages 4–9

Changhan Wang, Kyunghyun Cho, Jiatao Gu. 2019. Neural Machine Translation with Byte-Level Subwords, *CoRR*,abs/1909.03341, http://arxiv.org/abs/1909.03341

Yingting Wu, Hai Zhao. 2018. Finding Better Subword Segmentation for Neural Machine Translation, In: *CoRR*,abs/1807.09639, http://arxiv.org/abs/1807.09639