English-Marathi Neural Machine Translation for LoResMT 2021

Vandan Mujadia vandan.mu@research.iiit.ac.in
Dipti Misra Sharma dipti@iiit.ac.in
Machine Translation - Natural Language Processing Lab
Language Technologies Research Centre
Kohli Center on Intelligent Systems
International Institute of Information Technology
Hyderabad

Abstract

In this paper, we (team - oneNLP-IIITH) describe our Neural Machine Translation approaches for English-Marathi (both direction) for LoResMT-2021[1]. We experimented with transformer based Neural Machine Translation and explored the use of different linguistic features like POS and Morph on subword unit for both English-Marathi and Marathi-English. In addition, we have also explored forward and backward translation using web-crawled monolingual data. We obtained 22.2 (overall 2^{nd}) and 31.3 (overall 1^{st}) BLEU scores for English-Marathi and Marathi-English on respectively.

1 Introduction

Machine Translation (MT) is a field of Natural Language Processing which aims to translate a text from one natural language (i.e English) to another (i.e Marathi). The meaning of the source text must be fully preserved in the resulting translated text in the target language. Recent years have seen significant quality advancements in machine translation with the advent of Neural Machine Translation. For the translation task, different types of machine translation systems have been developed and they are mainly categorized into Rule based Machine Translation (RBMT)(Forcada et al., 2011), Statistical Machine Translation (SMT) (Koehn, 2009) and Neural Machine Translation (NMT) (Bahdanau et al., 2014).

Rule based Machine Translation (RBMT) translates on the basis of grammatical rules. It involves a grammatical analysis of the source language and the target language. Based on the analysis, it generates the translated sentence (Dwivedi and Sukhadeve, 2010). Statistical Machine Translation (SMT) is based on statistical models, which analyse large parallel and monolingual text and tries to determine the correspondence between a source language word and a target language word. NMT (Bahdanau et al., 2014) is an end to end approach for automatic machine translation without heavy hand crafted feature engineering. Due to recent advances, NMT has been receiving heavy attention and achieved state of the art performance in the task of language translation. With this work, we intend to check how NMT systems could be developed for low resource machine translation.

[1] https://sites.google.com/view/loresmt/

This paper describes our experiments for LoResMT-2021[2](Ojha et al., 2021). The third edition of LoResMT-2021 aims at building MT systems for low-resource language pairs on COVID-related texts. For our work, we focused only on English-Marathi language pair (both directions) and participated for categories where in first, we only used given parallel training data (constrained) and in second, we utilized available parallel corpora from different sources for English-Marathi and English-Hindi (unconstrained).

In this work, we experimented only with Transformer (Vaswani et al., 2017) based Neural Machine Translation throughout. Along with it, we also explored the morph (Virpioja et al., 2013) induced sub-word segmentation with byte pair encoding (BPE)(Sennrich et al., 2016b) to enable open vocabulary translation. We used POS tags as linguistic feature for English-Marathi direction along with forward and back translation to leverage synthetic data for machine translation. We also explored the use of English-Hindi parallel data for English-Marathi as origin of these two languages are the same and they are Indo-aryan languages (wikipedia, 2021). Hindi is said to have evolved from Sauraseni Prakrit (wikipedia Hindi, 2021) whereas Marathi is said to have evolved from Maharashtri Prakrit (wikipedia Marathi, 2021) and they both use the same writing script - Devanagari[3]. In LoResMT-2021, we participated as team "oneNLP-IIITH".

2 Data

Data (Language)	#Sentences	#Token	#Type
Train - English (Parallel)	20,933	0.3M	28K
Train - Marathi (Parallel)	20,933	0.29M	42K
Validation - English (Parallel)	500	12K	3.7K
Validation - Marathi (Parallel)	500	10K	4.7K
English (Monolingual)	8K	0.1M	200K
Marathi (Monolingual)	21K	0.2M	39K

Table 1: English-Marathi LoResMT-2021 Training data (for Constrained)

Data (Language)	#Sentences	#Token	#Type
Train - English (Parallel)	7M	13M	0.5M
Train - Hindi (Parallel)	7M	5.6M	0.9K
Train - English (Parallel)	1.8M	2.5M	0.1K
Train - Marathi (Parallel)	1.8M	2.2M	0.6K
English (Monolingual)	0.1M	-	-
Marathi (Monolingual)	0.1M	-	-

Table 2: Other Utilised data (for Unconstrained)

We utilized provided parallel and monolingual corpora for the Machine Translation task on English<->Marathi language pairs. Table-1 describes the training (parallel and monolingual) and validation data (parallel) after cleaning (i.e removed parallel data from training which are also in validation). We carried out constrained experiments on this data. For unconstrained experiments we use additional parallel dataset from samanantar (Ramesh et al., 2021). For back

[2] https://sites.google.com/view/loresmt/

[3] https://en.wikipedia.org/wiki/Devanagari

and forward translation, we web-crawled monolingual data for both English and Marathi to aid relatively new NLP domain Covid. Table-2 describes this additional dataset in terms of number of sentences, token and type.

3 Data Pre-Processing

For data pre-precessing, we used IndicNLP Tool[4] with in-house tokenizer to tokenize and clean both English and Marathi corpora (train, test, valid and monolingual) as a first step. Following subsections explain other pre-processing steps for our MT experiments.

3.1 Morph + BPE Segmentation

Based on token/type ratio, Marathi is morphologically richer compared to English from Table-1. Translating from morphologically-rich agglutinative languages is more difficult due to their complex morphology and large vocabulary. We address this issue with a segmentation method which is based on morphology and BPE segmentation (Sennrich et al., 2016b) as a pre-processing step as prescribed in (Mujadia and Sharma, 2020). We utilized unsupervised Morfessor (Virpioja et al., 2013) by training it on monolingual data for Marathi. We then applied this trained Morfessor model on our corpora (train, test, validation) to get meaningful stem, morpheme, suffix segmented sub-tokens for each word in a sentence. Subsequently, we applied the subword algorithm on top of the morph segmentation as shown in Figure-1. For English, we only applied subword segmentation throughout the experiments.

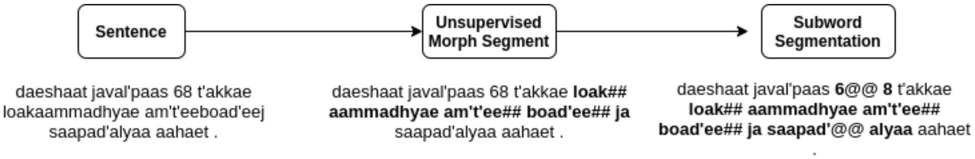
Sentence
Unsupervised Morph Segment
Subword Segmentation
daeshaat javal'paas 68 t'akkae loakaammadhyae am't'eeboad'eej saapad'alyaa aahaet .
daeshaat javal'paas 68 t'akkae loak## aammadhyae am't'ee## boad'ee## ja saapad'alyaa aahaet .
daeshaat javal'paas 6@@ 8 t'akkae loak## aammadhyae am't'ee## boad'ee## ja saapad'@@ alyaa aahaet .

Figure 1: Morph and Subword based pre-processing for a Marathi sentence. Here ## denotes UMorph based segmentation and @@ denotes subword based segmentation

3.2 Features

We carried out experiments using Part of Speech (POS) tag as a word and subword level feature (Sennrich and Haddow, 2016) only for English. We used Spacy (Honnibal et al., 2020) toolkit to get POS tags for English and used them by concatenating their embedding with word embedding for NMT training as shown in Figure-2.

3.3 Hindi centric parallel data

For unconstrained experiments, we experimented and studied the use of available parallel data. Along with the English-Marathi parallel data, we utilized a small chunk of English-Hindi parallel data from Samanantar corpus (Ramesh et al., 2021) as Hindi is a close and related language to Marathi. We appended the English-Hindi parallel data to the existing English-Marathi data and maintained 1:1 ratio of them for overall training.

3.4 Forward and Back Translation

Back translation is a widely used data augmentation method for low resource neural machine translation (Sennrich et al., 2016a). Here, we utilized the provided and web crawled monolin-

[4] http://anoopkunchukuttan.github.io/indic nlp library/

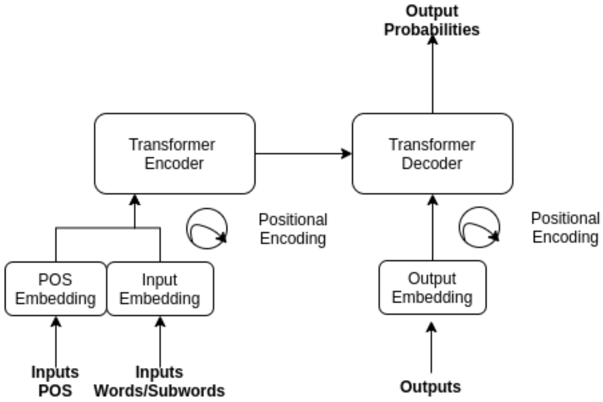

Figure 2: Modeling POS tags as feature along with word embedding for English in Transformer Network

gual data (for Marathi and English). We used around 0.1M forward and back translated pairs for both translation directions.

4 Training Configuration

Throughout all experiments, we used Transformer sequence to sequence architecture with the following configuration for constrained and unconstrained experiments.

- **Constrained**
  Morph + BPE based subword segmentation, POS tags as feature, Embedding size : 512 Transformer for encoder and decoder, rnn_size 512 feature_Embedding 100 (only for POS), heads 4 encoder - decoder layers : 2, label smoothing : 1.0, dropout : 0.30, Optimizer : Adam, Beam size : 4 (train) and 10 (test), training steps : 20K

- **Unconstrained**
  Morph + BPE based subword segmentation, Embedding size : 512 Transformer for encoder and decoder, RNN_size 512, heads 8 encoder - decoder layers : 6, label smoothing : 1.0, dropout : 0.30, Optimizer : Adam, Beam size : 4 (train) and 10 (test), training steps : 20K

For these experiments, we used shared vocab across trainings. We used Opennmt-py (Klein et al., 2020) toolkit with above configuration for our experiments.

Using the above described configuration, we performed experiments based on different parameter (feature) configurations. We trained and tested our models on word level, BPE level and morph + BPE level for input and output. We also used POS tagger and experimented with shared vocabulary across the translation task. The results are discussed in following Result section.

5 Result

Table-3 and Table-4 show performance of our systems with different configurations in terms of BLEU score (Papineni et al., 2002) for English-Marathi and Marathi-English respectively on the validation and Test data. We achieved highest 17.9 development and 22.2 test BLEU scores for English-Marathi and highest 32.88 development and 31.6 test BLEU scores for

Type	Feature	BPE	Valid	Test
C	Word	-	13.03	-
C	BPE	7.5K	13.25	-
C	Morph + BPE + POS	-	14.17	10.4
C	Morph + BPE + POS	7.5K	14.03	-
C	Morph + BPE + POS	15K	14.54	11.5
C	Morph+BPE+POS + BT(1L sent)	15K	14.89	**14.0**
UC	BPE + (Eng-Mar ExtData)	10K	11.73	-
UC	BPE+(Eng-Mar&Eng-Hin ExtData ExtData)	10K	13.73	21.5
UC	BPE+(Eng-Mar&Eng-Hin ExtData)+F-BT	10K	**16.25**	**22**
UC	Morph+BPE+(Eng-Mar&Eng-Hin ExtData)+F-BT	10K	**17.90**	**22.2**

Table 3: BLEU scores for English-Marathi. Here C stands Constrained and UC for Unconstrained, BPE stands for byte pair encoding (subword), Morph for Morphological segment and POS for Part of Speech and F-BT for forward and backward translation

Type	Feature	BPE	Valid Data	Test Data
C	BPE	10K	19.11	16.2
C	BPE	7.5K	19.47	16.4
C	Morph+BPE	7.5K	19.67	**16.7**
UC	BPE + (Eng-Mar ExtData)	7.5K	20.10	20.7
UC	BPE+(Eng-Mar&Eng-Hin ExtData)	10K	**29.80**	**30.6**
UC	BPE+(Eng-Mar&Eng-Hin ExtData)+F-BT	10K	**32.88**	**31.6**

Table 4: BLEU scores for Marathi-English. Here C stands Constrained and UC for Unconstrained, BPE stands for byte pair encoding (subword), Morph for Morphological segment and F-BT for forward and backward translation

Marathi-English systems respectively.

The results show that for low resource settings, transformer network based MT models can be improved with linguistic information like morph and POS features. The results also indicate that morph based segmentation along with byte pair encoding improves BLEU score and can be used for morph rich languages. The results also suggest that performance drastically improves when model is exposed to more parallel data (for unconstrained setting). Our experiments suggest that use of English-Hindi parallel data gives performance boost by 3.0+ BLEU points for English-Marathi and almost 10.0+ BLEU points for Marathi-English. Also, forward and back translated synthetic data obtained from same Covid domain improves quality of NMT models marginally, as they could be helping models to do better generalization. From the Test results (Table-3 and Table-4), we stand at overall 2^{nd} and 1^{st} for English-Marathi and Marathi-English respectively.

6 Conclusion

From our experiments, we conclude that linguistic feature driven NMT for low resource languages is a promising approach and use of similar language training data gives a significant boost in performance to the low resource language.

References

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Dwivedi, S. K. and Sukhadeve, P. P. (2010). Machine translation system in indian perspectives. *Journal of computer science*, 6(10):1111.

Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Sánchez-Martínez, F., Ramírez-Sánchez, G., and Tyers, F. M. (2011). Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25(2):127–144.

Honnibal, M., Montani, I., Van Landeghem, S., and Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python.

Klein, G., Hernandez, F., Nguyen, V., and Senellart, J. (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.

Koehn, P. (2009). *Statistical machine translation*. Cambridge University Press.

Mujadia, V. and Sharma, D. M. (2020). Nmt based similar language translation for hindi-marathi. In *Proceedings of the Fifth Conference on Machine Translation*, pages 414–417.

Ojha, A. K., Liu, C.-H., Kann, K., Ortega, J., Satam, S., and Fransen, T. (2021). Findings of the LoResMT 2021 Shared Task on COVID and Sign Language for Low-Resource Languages. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (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.

Ramesh, G., Doddapaneni, S., Bheemaraj, A., Jobanputra, M., AK, R., Sharma, A., Sahoo, S., Diddee, H., J, M., Kakwani, D., Kumar, N., Pradeep, A., Deepak, K., Raghavan, V., Kunchukuttan, A., Kumar, P., and Khapra, M. S. (2021). Samanantar: The largest publicly available parallel corpora collection for 11 indic languages.

Sennrich, R. and Haddow, B. (2016). Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, pages 83–91.

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

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

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Virpioja, S., Smit, P., Grönroos, S.-A., Kurimo, M., et al. (2013). Morfessor 2.0: Python implementation and extensions for morfessor baseline.

wikipedia (2021). Indo-aryan languages - wikipedia. https://en.wikipedia.org/wiki/Indo-Aryan_languages.

wikipedia Hindi (2021). Shauraseni prakrit - wikipedia. https://en.wikipedia.org/wiki/Shauraseni_Prakrit.

wikipedia Marathi (2021). Maharashtri prakrit - wikipedia. https://en.wikipedia.org/wiki/Maharashtri_Prakrit.