NMT based Similar Language Translation for Hindi - Marathi

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

Abstract

This paper describes the participation of team F1toF6 (LTRC, IIIT-Hyderabad) for the WMT 2020 task, similar language translation. We experimented with attention based recurrent neural network architecture (seq2seq) for this task. We explored the use of different linguistic features like POS and Morph along with back translation for Hindi-Marathi and Marathi-Hindi machine translation.

1 Introduction

Machine Translation (MT) is the field of Natural Language Processing which aims to translate a text from one natural language (i.e Hindi) to another (i.e Marathi). The meaning of the resulting translated text must be fully preserved as the source text in the target language.

For the translation task, different types of machine translation systems have been developed and they are mainly Rule based Machine Translation (RBMT)(Forcada et al., 2011), Statistical Machine Translation (SMT) (Koehn, 2009) and Neural Machine Translation (NMT) (Bahdanau et al., 2014).

Statistical Machine Translation (SMT) aims to learn a statistical model to determine the correspondence between a word from the source language and a word from the target language. Neural Machine Translation is an end to end approach for automatic machine translation without heavily 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 used for low resource and similar language machine translation.

Data	Sents	Token	Type
Hindi (Parallel)	38,246	7.6M	39K
Marathi (Parallel)	38,246	5.6M	66K
Hindi (Mono)	80M	-	-
Marathi (Mono)	3.2M	-	-

Table 1: Hindi-Marathi WMT2020 Training data

This paper describes our experiments for the task of similar language translation of WMT-2020. We focused only on Hindi-Marathi language pair for the translation task (both directions). The origin of these two languages are the same as they are Indo-aryan languages(wikipedia, 2020). Hindi is said to have evolved from Sauraseni Prakrit (wikipedia Hindi, 2020) whereas Marathi is said to have evolved from Maharashtri Prakrit (wikipedia Marathi, 2020). They also have evolved as two major languages in different regions of India.

In this work, we focused only on recurrent neural network with attention based sequence to sequence architecture throughout all experiments. 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 and back translation to leverage synthetic data for machine translation task in both directions. In the similar language translation task of WMT-2020, we participated as team named “f1plusf6”.

2 Data

We utilised parallel and monolingual corpora provided for the task on Hindi<->Marathi language pairs. Table-1 describes the training data (parallel

Proceedings of the 5th Conference on Machine Translation (WMT), pages 414–417

and monolingual) on which we carried out all experiments. We deliberately excluded Indic WordNet data from the training after doing manual quality check. As this is a constrained task, our experiments do not utilise any other available data.

3 Pre-Processing

As a first pre-precessing step we use IndicNLP Toolkit[1] along with an in-house tokenizer to tokenize and clean both Hindi and Marathi corpora (train, test, dev and monolingual).

3.1 Morph + BPE Segmentation

Marathi and Hindi are morphologically rich languages and from the Table-1, based on the the comparative token/type ratio, one can find that Marathi is a more agglutinative language than Hindi. Translating from morphologically-rich agglutinative languages is more difficult due to their complex morphology and large vocabulary. To address this issue, we have come up with a segmentation method which is based on morph and BPE segmentation (Sennrich et al., 2016b) as a pre-processing step.

In this method, we utilised unsupervised Morfessor (Virpioja et al., 2013) to train a Morfessor model on monolingual data for both languages. 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 each sentence.

(1) aur jab maansaahaaree
pakshee lothon par jhapate ,
tab abraam ne unhen uda diya .
'And when the carnivorous birds swooped on the carcasses, Abram blew them away.'

(2) aur jab maansaa##haaree
pakshee loth##on par jhapat##e ,
tab ab##raam ne unhen uda diya .
'And when the carnivorous birds swooped on the carcasses, Abram blew them away.'

(3) aur jab maan@@ saa##haaree
pakshee loth##on par jha@@ pat##e ,
tab ab##raam ne unhen uda diya .
'And when the carnivorous birds swooped on the carcasses, Abram blew them away.'

We demonstrate this method with a Hindi sentence as given in Example-1. Example -1, shows Hindi text with romanized text and the corresponding English translation for better understanding. The Example-2 shows the same sentence with Morfessor based segmentation with token ##. Here we notice that Morfessor model has segmented the Hindi words into meaningful stems and suffixes. i.e maansaahaaree=maansaa + haaree(meat + who eats). We would like to use it in our experiments to tackle the difficulties that arise due to complex morphology at the source language in machine translation tasks. On top of this morph segmented text we applied BPE (Sennrich et al., 2016a) as given in Example-3. Here @@ is sub-word separator for byte pair based segmentation and ## is the separator for morph based segmentation.

3.2 Features

For Hindi to Marathi translation, we carried out experiments using Part of Speech (POS) tags as a word level as well as a subword level feature as described in (Sennrich and Haddow, 2016). We use LTRC shallow parser[2] toolkit to get POS tags.

4 Training Configuration

Recurrent Neural Network (RNN) based machine translation models work on encoder-decoder based architecture. Here, the encoder takes the input (source sentence) and encodes it into a single vector (called as a context vector). Then the decoder takes this context vector to generate an output sequence (target sentence) by generating a word at a time(Sutskever et al., 2014). Attention mechanism is an extension to this sequence to sequence architecture to avoid attempting to learn a single vector. Instead, based on learnt attention weights, it focuses more on specific words at the source end and generates a word at a time. More details can be found here (Bahdanau et al., 2014), (Luong et al., 2015).

For our experiments, we utilize sequence to sequence NMT model with attention for all of our experiments with following configuration.

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

[2] http://ltrc.iiit.ac.in/analyzer/

Model	Feature	BPE (Merge ops)	BLEU
BiLSTM + LuongAttn	Word level	-	19.70
BiLSTM + LuongAttn	Word + Shared Vocab (SV)+ POS	-	20.49
BiLSTM + LuongAttn	BPE	10K	20.1
BiLSTM + LuongAttn	BPE+SV+MORPH Segmentation	10K	20.44
BiLSTM + LuongAttn	BPE+SV+MORPH+POS	10K	**20.62**
BiLSTM + LuongAttn	BPE+SV+MORPH+POS + BT	10K	16.49

Table 2: BLEU scores on Development data for Hindi-Marathi

Model	Feature	BPE (Merge ops)	BLEU
BiLSTM + LuongAttn	Word level	-	21.42
BiLSTM + LuongAttn	Word + Shared Vocab (SV)	-	23.84
BiLSTM + LuongAttn	BPE	20K	24.56
BiLSTM + LuongAttn	BPE+SV+MORPH Segmentation	20K	25.36
BiLSTM + LuongAttn	BPE+SV+MORPH+POS	20K	**25.55**
BiLSTM + LuongAttn	BPE+SV+MORPH+POS + BT	20K	23.80

Table 3: BLEU scores on Development data for Marathi-Hindi

- Morph + BPE based subword segmentation, POS tags as feature
- Embedding size : 500
- RNN for encoder and decoder: bi-LSTM
- Bi-LSTM dimension : 500
- encoder - decoder layers : 2
- Attention : luong (general)
- copy attention(Gu et al., 2016) on dynamically generated dictionary
- label smoothing : 1.0
- dropout : 0.30
- Optimizer : Adam
- Beam size : 4 (train) and 10 (test)

As these are two similar languages, share writing scripts and large sets of named entities, we used shared vocab across training. We used Opennmt-py (Klein et al., 2020) toolkit with above configuration for our experiments.

5 Back Translation

Back translation is a widely used data augmentation method for low resource neural machine translation(Sennrich et al., 2016a). We utilised monolingual data (i.e of Marathi) and a NMT model trained on given training data for a direction (i.e, Marathi to Hindi) to enrich training data of the opposite directional NMT training (i.e, Hindi - Marathi) by populating synthetic data. We used around 5M back translated pairs (after perplexity based pruning with respect to sentence length) for both translation directions.

Using 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.

6 Result

Table-2 and Table-3 show performance of systems with different configuration in terms of BLEU score(Papineni et al., 2002) for Hindi-Marathi and Marathi-Hindi respectively on the validation data. We achieved 20.62 and 25.55 development and 5.94 and 18.14 test BLEU scores for Hindi-Marathi and Marathi-Hindi systems respectively.

The results show that for low resource similar language settings, MT models based on sequence to sequence neural network can be improved with linguistic information like morph based segmentation and POS features. The results also show that morph based segmentation along with

byte pair encoding improves BLEU score for both directions. But Marathi-Hindi directed translation shows considerable improvement. Therefore our method shows improvement while translating from morphologically richer language (Marathi) to comparatively less morphologically richer language (Hindi).

The results also suggest that the use of back translated synthetic data for low resource language pairs reduces the overall performance marginally. The reason for this could be, due to low quantity of training data for NMT models, they could be over learning and back translation could be helping to do better generalization.

7 Conclusion

We conclude from our experiments that linguistic feature driven NMT for similar low resource languages is a promising approach. We also believe that morph+BPE based segmentation is a potential segmentation method for morphologically richer languages.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Mikel L Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25(2):127–144.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor OK Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1631–1640.

wikipedia Hindi. 2020. Shauraseni prakrit - wikipedia. https://en.wikipedia.org/wiki/Shauraseni_Prakrit. (Accessed on 08/15/2020).

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.

Philipp Koehn. 2009. *Statistical machine translation*. Cambridge University Press.

Minh-Thang Luong, Hieu Pham, and 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.

wikipedia Marathi. 2020. Maharashtri prakrit - wikipedia. https://en.wikipedia.org/wiki/Maharashtri_Prakrit. (Accessed on 08/15/2020).

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.

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. 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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 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.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Sami Virpioja, Peter Smit, Stig-Arne Grönroos, Mikko Kurimo, et al. 2013. Morfessor 2.0: Python implementation and extensions for morfessor baseline.

wikipedia. 2020. Indo-aryan languages - wikipedia. https://en.wikipedia.org/wiki/Indo-Aryan_languages. (Accessed on 08/17/2020).