INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

R&D SHOWCASE 2021
Technology, Social Impact

NMT based Similar Language Translation for Hindi - Marathi

Objectives

Poster describes participation of team F1toF6 (LTRC, IIIT-Hyderabad) for the WMT 2020 task, similar language translation. Following are the points

- To investigate the applicability of current MT techniques for similar language translation
- To check the usage and effectiveness of different linguistic features like POS and Morph for Indic language settings
- The effect of Back Translation under similar language low resource machine translation settings

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. This paper describes our experiments 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. Hindi is said to have evolved from Sauraseni Prakrit whereas Marathi is said to have evolved from Maharashtri Prakrit. 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.

- Experimented with attention based recurrent neural network architecture (seq2seq) for Hindi-Marathi and Marathi-Hindi machine translation.
- Explored the use and effectiveness of different linguistic features like POS and Morph
- Tried applying Back Translation to improve translation quality under low resource settings

Data

Used provided parallel and monolingual corpora. Table-1 describes it in detail.

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

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.

Pre-Processing

- Tokenization and cleaning of Hindi and Marathi using Toolkit[a] and in-house tokenizer.
- A novel segmentation method, based on morph and byte pair encoding [1]. Used Morfessor [2] to get meaningful stem, morpheme and suffix segmented sub-tokens. Explained with a Hindi sentence as given in Example-1,2,3.
- Linguistic Features : We use LTRC shallow parser[b] toolkit to get POS tags.

[a]http://anoopkunchukuttan.github.io/indic_nlp_library/
[b]http://ltrc.iiit.ac.in/analyzer/

(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.'

Figure 1:Pre-processing : Segmentation Example for Hindi (in Roman script)

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 ##.Example-3. Here @@ is sub-word separator for byte pair based segmentation and ## is the separator for morph based segmentation.

Training Configuration

- Morph + BPE based subword segmentation+POS
- Embedding size : 500
- RNN for encoder and decoder: bi-LSTM
- Bi-LSTM dimension : 500
- encoder - decoder layers : 2
- Attention : luong (general)
- copy attention[3] on dynamically generated dictionary
- label smoothing : 1.0
- dropout : 0.30
- Optimizer : Adam
- Beam size : 10

Results

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

Figure 2:Results

Conclusion

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

Back Translation

Used around 5M back translated pairs (after perplexity based pruning with respect to sentence length) for both translation directions as synthetic data.

Result

Figure-2 shows the performance of systems with different configuration in terms of BLEU score[4] 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.

References

[1] Rico Sennrich, Barry Haddow, and Alexandra Birch. 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, 2016.

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

[3] Jiatao Gu, Zhengdong Lu, Hang Li, and Victor OK Li. 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, 2016.

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

Vandan Mujadia, Dipti Misra Sharma

LANGUAGE TECHNOLOGIES RESEARCH CENTER (LTRC)