Neural Machine Translation for English-Malayalam

[1] Vijay Sundar Ram R, [2] Sobha Lalitha Devi
[1,2]AU-KBC Research Centre, MIT Campus of Anna University
[1] sundar@au-kbc.org, [2]sobha@au-kbc.org

ABSTRACT

Neural Machine Translation systems produce state-of-art translation for high resource languages. It is yet a challenge in low-resource and morphologically rich languages. In this paper, we have discussed the existing techniques in handling the morphologically rich and low-resource languages and presented our experiments on developing English-Malayalam NMT system where we have processed the data using different techniques namely word segmentation using morphological analyser and applying Byte pair Encoding (BPE) technique. The results show a significant improvement by implementing the word segmentation using morphological analyser.

Keywords: Neural Machine Translation, Morphologically rich languages, Morph segmentation, Byte Pair Encoding.

INTRODUCTION

Machine Translation (MT) is one of the most dealt fields of Natural Language Processing (NLP), yet there is an on-going research to achieve near human translation. MT is the automated process of decoding the meaning of source text and recording the meaning in target language without loss of information. To achieve this task, there were systems with different approaches such as Dictionary based, Interlingual based, Example based MT (EBMT), Statistical Machine Translation (SMT), Analytical-Transfer-Generation based approach and the present Neural Machine Translation (NMT) approach.

Neural Machine Translation (NMT) started with the successful works by Kalchlorenner and Blunsom [8], Sutskever et al.,[21] and Cho et al., [3], where an encoder encoded the source sentence to a fixed-length vector, from which a decoder generated the translation. Sutskever et al.,[21] reported a NMT system built based on RNN with long short term memory (LSTM), this surpassed the performance of the previous start-of-art performance. These seq2seq models work well for short sentences, but do not perform well for long sentences due to the vanishing gradient problem. Bahdanau et al.,[2] presented an extended encoder-decoder to handle the problem of encoding of source sentence into a fixed-length vector. They used a bidirectional recurrent neural network (RNN) consisting of forward and backward RNN to focus around the word. Attention mechanism was introduced in the decoder to decide the part of the source sentence to pay attention. Luong et al.,[14] simplified the attention mechanism by considering the hidden states at the top layer of both encoder and decoder. This attention mechanism attends to the entire input sequence.

With improvements in attention mechanism Vaswani et al., [22] introduced a new architecture called transformer with encoder and decoder that relies solely on attention mechanism. The Transformer model relies on self-attention where all input sequence members are compared with each other, and modifies the corresponding output sequence position. Though these NMT systems (Bahdanau et al.,[2],Vaswani et al., [22]) has led to a greater improvement in translation of high resource languages, the translation of morphologically rich and low resource languages is a major challenge. In this paper, we discuss our NMT experiments in building English to Malayalam translation system. English-Malayalam is a low resource language pair and Malayalam is a morphologically rich language. Agglutination is also very high in Malayalam.

Further sections of the paper is organized as follows: In the following section, we discuss the different techniques to improve NMT in low resource and morphologically rich languages. We also summarize the different NMT works in Indian languages. Third section we describe briefly the characteristics of Malayalam language, which pose challenge in building a NMT system. In section 4, we describe our experimental setup and data preparation. Section 5 has the result and analysis. We conclude the paper with a conclusion sentence, where the gist of the work is presented.

RECENT WORKS

In this section we present the different approaches used to handle low resource and morphologically rich languages in NMT and present a brief summary of different NMT works in Indian languages.

A. Techniques to Mitigate the Low-Resource problem in NMT

Low resource of parallel data is a bottleneck in many language pairs. Different approaches were executed to overcome or reduce the problem. We briefly discuss these techniques in the following section.

Increasing the data using Back Translation

Senrich et al., [20] introduced back translation technique, where the monolingual data of the target language is translated to source language using available MT system and combined with the training data. This helps in improving the translation quality.

Phrase Table Injection

Zhao et al., [23] presented a method to combine the SMT and NMT by utilizing the phrase table generated in the SMT training. It is combined with the data in training the NMT.

Leveraging the Pre-trained models

Pre-trained models such as BERT, Glove, RoBERTa are commonly used in NMT to improve the quality of translation. These models are used in fine-tuning the NMT training.

Combining the Corpus

When similar languages are on the target side, in this technique, the knowledge is exploited to translate the mixed language better. Banerjee A et al., [1] has presented a technique, where English-Hindi and English-Marathi corpus are combined to train the NMT and English-Marathi corpora is used to fine-tune the NMT training.

Transfer Learning

Transfer learning is the process of applying an existing training Machine learning model to a new, but related problem. In pivot-based transfer learning, first they pre-train a source-pivot model with a source-pivot parallel corpus and a pivot-target model with a pivot-target parallel corpus. Then initialize the source-target model with the source encoder from the pre-trained source-pivot model and the target decoder from the pre-trained pivot-target model. Now the training with a source-target parallel corpus is continued. Kim et al., [9] has proposed three methods to increase the relation among source, pivot, and target languages in the pre-training: 1) step-wise training of a single model for different language pairs, 2) additional adapter

component to smoothly connect pre-trained encoder and decoder, and 3) cross-lingual encoder training namely autoencoding of the pivot language.

Domain term translation in most of languages is a challenging task and in Indian languages it is more challenging due to very less availability of parallel domain terms. Hema Ala et al.,[7] has proposed to handle domain terms in NMT using the back translation technique, where Domain specific back translation using monolingual and generates synthetic data. They have conducted experiments on Chemistry and Artificial Intelligence domains for Hindi and Telugu in both directions.

Unsupervised NMT (UNMT) is one of the up-coming techniques to overcome the low-resource problem. It is shown that UNMT works for source and target languages are similar and in same domain. Sai Koneru et al., [11] had presented an experiment on UNMT for Dravidian languages (Kannada, Tamil, Telugu and Malayalam) to English.

Ranathunga et al., [17] has presented a detailed survey on NMT works in low-resource languages.

B. Techniques to Improve NMT in Morphologically Rich Languages

Translation of morphologically rich languages using NMT has the following challenges, a) large number of inflected forms lead to a larger vocabulary and thus causes data sparsity. b) Generating sentences with correct linguistic agreement and expressing exact semantics of the input sentence is a challenge.

These challenges are handled using the following techniques; a) Breaking the word forms into sub-word units, so that the overall vocabulary size is reduced. b) Training with linguistic features such as lemma-tag strategy.

Sub-word units are generated using

A: Statistical methods such as Sub-word Text Encoder (STE) and Byte Pair Encoding (BPE)

B: Linguistically motivated word segmentation methods using tools such as Mofessor and Morphological analysers.

Dominik Machacek et al., [5] 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. Goyal et al., [6] 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., [4] 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., [13] and Saldanha et. al., [18] has presented their work in Tamil-Telugu translation. Mujadia et al., [15] has presented their work in Marathi-Hindi bidirectional translation.

CHARACTERISTICS OF MALAYALAM

Malayalam belongs to Dravidian language family. It is a morphological rich and highly agglutinative language. Agglutination is more in Malayalam than other Dravidian languages namely Kannada, Telugu and Tamil. Consider the following examples.

Ex1: അകപ്പെട്ടിരിക്കുകയാണെങ്കിലും (even if it is caught inside)

SPLIT WORDS: അകപ്പെട്ടിരിക്കുക + ആണ് + എങ്കിലും

അകപ്പെട്ടിരിക്കുക അകപ്പെട് (V)+ ഇ (VBP) + ഇരി (AUX) + ഉക (INF)

ആണ് ആണ്(COP)

എങ്കിലും എങ്കിലും (CONJ)

Ex 2 : നിലവാരത്തിലെത്തി (reached a standard)

SPLIT WORDS: നിലവാരത്തിൽ + എത്തി

നിലവാരത്തിൽ നിലവാരം(N) +ഇൽ(LOC)
എത്തി എത്ത് (V)+ ഇ(PAST)

Ex1 and Ex2 have two agglutinative words. In Ex1 ‘ ’, three words are agglutinated to form a single word. Similarly in Ex2., ‘ ’, has two words (an inflected noun and inflected verb) combined to form one word.
Malayalam has SOV sentence structure. Non-finite verbs bring clausal constructions in Malayalam. Person, Number, Gender (PNG) agreement with the subject and the finite verb is not in Malayalam. Copula verb is obligatory in Malayalam.

EXPERIMENTS

In this section, we discuss about the details of the parallel dataset, experimental setup for developing English-Malayalam NMT system and data preparation for three different experiments.

A. Dataset

We have used the PMIndia English-Malayalam corpus and English-Malayalam corpus developed from the manually translated Swayam course lectures in 31 courses. These courses include different domains namely Information Technology, Science and Technology, Management, Food Processing technology and Law. The statistics of the corpus is given the tables below.

Details	English (Source)	Malayalam (Target)
Number of Sentences	33,661	33,661
Number of Words	6,05,704	3,71,896
Number of unique words	25,129	92,052
Maximum Length of a Sentence (words)	98	62

Table (1)- Statistics of PMIndia Corpus

Details	English (Source)	Malayalam (Target)
Number of Sentences	1,43,433	1,43,433
Number of Words	25,78,337	16,24,744
Number of unique words	47,154	1,92,274
Maximum Length of a Sentence (words)	85	56

Table (2)- Statistics of Swayam Corpus

Details	English (Source)	Malayalam (Target)
Number of Sentences	1,77,094	1,77,094
Number of Words	31,84,041	19,96,640
Number of unique words	53,680	2,33,174
Maximum Length of a Sentence (words)	98	62

Table (3)- Statistics of Combined Corpus

Table(1) has the statistics of the PMIndia corpus; Table(2) has the statistics of the Swayam Corpus and Table(3) has the statistics of the combined corpus. In the third row in three tables, the number of words in English corpus is nearly twice of the Malayalam corpus and in the fourth

row of the tables, the unique words in Malayalam (includes inflected words) is nearly four times the number of unique words in English. In the fifth column of the tables, where the maximum length of a sentence is presented, the number of words in Malayalam slightly more than the half of the words in English. The information in these two rows clearly shows the morphological richness and high agglutination in Malayalam, which make the NMT training a challenging task.

B. Experiment Setup

We used OpenNMT-py toolkit for developing the English-Malayalam NMT system. 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

For combined corpus, 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.

C. Data Preparation

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

Word Level: The sentences in both the languages where tokenised and used for NMT training.

BPE: Byte Pair Encoding (BPE) proposed by Sennrich et al.,[19] was applied to the tokenised data. We used 2500 as BPE merge value for Malayalam.

MorphSeg: Malayalam being a morphologically rich and highly agglutinative language, we explored the word segmentation using morphological analyser. In this experiment we segmented only nouns and its suffixes. Morphological analyser built using paradigm and Finite state automata based approach was used [12].

RESULT AND ANALYSIS

We evaluated the translations from the three NMT models using BLEU score (Papineni et al., [16]. We used Sacre-bleu python library to calculate the BLEU scores. The results are presented in Table (4).

Model	BLEU Score
Word-Level	14.24
BPE	20.90
MorphSeg	25.53

Table (4)- BLEU score for English-Malayalam for different model.

The translation of using Word-Level model has more unknown word '<unk>' compared to the other three models. There were many partial translations.

In the second experiment, BPE, with reduction in the vocabulary size, the BLEU score of the translation from this model has increased significantly, compared to Word-Level model. And the translation was complete. The translation had transliterated forms of the words.

In the translation using MorphSeg model where the words are split into root and suffix, translation was better than the previous two models. But this had unknown words.

We have explained the translations further using the translations from three different models.

Ex 3:

Source Sentence: Some journals charge for fast processing of article.

Translated Sentences: ലേഖനത്തിൻറെ പ്രോസസ്സിംഗ് UNK ചില ജേർണലുകൾ.

BPE: ലേഖനത്തിന്റെ വേഗത്തിലുള്ള സംസ്കരണത്തിനായി ചില ജേർണലുകൾ ഈടാക്കുന്നു.

MorphSeg:ലേഖനത്തിന്റെ വേഗത്തിലുള്ള പ്രോസസ്സിംഗിന് ചില ജേർണലുകൾ ചാർജ് ഈടാക്കുന്നു.

Gold: ലേഖനത്തെ വേഗത്തിൽ പ്രോസസ്സിംഗ് ചെയ്യുവാൻ വേണ്ടി ചില ജേർണലുകൾ ചാർജ് ചുമത്തുന്നു.

In example 3, the translation using Word-Level model is partial and has unknown word. The translation using BPE model is complete but conveys a different meaning from the source sentence. The translation from MorphSeg is complete and meaningful.

Ex 4:

Source Sentence: The 2005 standard had 133 controls in eleven groups.

Translated Sentences: 2005 ൽ സ്റ്റാൻഡേർഡ് groups 600 ലധികം നിയന്ത്രണങ്ങളുണ്ട്.

BPE: 2005 സ്റ്റാൻഡേർഡ് പതിനൊന്ന് ഗ്രൂപ്പുകളി 133 നിയന്ത്രണങ്ങൾ ഉണ്ടായിരുന്നു.

MorphSeg: 2005 സ്റ്റാൻഡേർഡ് പതിനൊന്ന് ഗ്രൂപ്പുകളിൽ 133 നിയന്ത്രണങ്ങൾ ഉണ്ടായിരുന്നു.

Gold: 2005 സ്റ്റാൻഡേർഡ് പതിനൊന്ന് ഗ്രൂപ്പുകളിൽ 133 നിയന്ത്രണങ്ങൾ ഉണ്ടായിരുന്നു.

On analysing the translations in example 4, the output from Word-Level model is not complete and number 133 has changed to 600. Translation from MorphSeg model is good.

The above sample translations and the BLEU scores show that the word segmentation using morphological analyser improves the translation of English – Malayalam.

CONCLUSIONS

We have presented in this paper, the task of developing English-Malayalam Neural Machine Translation (NMT) system. English-Malayalam is a low resource language pair and Malayalam is a morphologically rich language with high agglutination. This poses a challenge in developing the NMT system. We have discussed the techniques to handle morphologically rich and low-resource languages. OpenNMT-py tool was used to build three different NMT model by training with data processed in three different methods namely Word-Level data, data processed with Byte Pair Encoding (BPE), segmenting word with morphological analyser. The evaluation of the translation from the different NMT models show that the NMT model trained after segmenting word with morphological analyser is performing better than the other models.

REFERENCES

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

[2] Banerjee, A.,Jain A., Mhaskar S., Deoghare S,D. Sehgal A., and Bhattacharya, P. (2021). Neural Machine Translation in Low-Resource Setting: a Case Study in English-Marathi Pair. *In Proceedings of the 18th Biennial Machine Translation Summit* - Volume 1: Research Track, MTSummit 2021 Virtual, pp 35-47

[3] Cho, K., van Merrienboer, B., Gulcehre, C., Bougares, F., Schwenk, H., and Bengio, Y. (2014).Learning phrase representations using RNN encoder-decoder for statistical machine translation. *In Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014).*

[4] Dewangan, S., Alva, S., Joshi, N., Bhattacharyya, P. (2021). Experience of neural machine translation between Indian languages. *Machine Translation* 35, 71–99

[5] 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*, pp. 277-284, Springer-Verlag, Cham, Switzerland, ISBN 978-3-030-00794-2

[6] 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*, pp 162-168

[7] Hema Ala, Vandan Mujadia, Dipti Misra Sharma. (2021). Domain Adaptation for Hindi-Telugu Machine Translation Using Domain Specific Back Translation. *In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pp 26-34

[8] Kalchbrenner, N. and Blunsom, P. (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.

[9] Kim, Y., Petrov, P., Petrushkov, P., Khadivi, S., and Ney, H. (2019). Pivot-based transfer learning for neural machine translation between non-English languages. *In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Pro˘cessing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 866–876, Hong Kong, China. Association for Computa˘tional Linguistics

[10] 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.

[11] Koneru, Sai; Liu, Danni; Niehues, Jan. (2021). Unsupervised Machine Translation On Dravidian Languages*, In 16th conference of the European Chapter of the Association for Computational Linguistics (EACL), Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages.*

[12] Lakshmi S., and Sobha Lalitha Devi (2013).”Malayalam Morphological Analyser”, *In processings of International Seminar on Current Trends in Dravidian Linguistics, May 27-29, 2013*

[13] Laskar SR., Paul B., Adhikary PK, Pakray P., Bandyopadhyay S. (2021), Neural Machine Translation for Tamil–Telugu Pair*. In Proceedings of the Sixth Conference on Machine Translation (WMT), pages 284–287*

[14] Luong M., Pham H., and Manning D. (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.

[15] Mujadia V. 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.*

[16] Papineni K., Roukos S., Ward T., and Zhu W (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.*

[17] Ranathunga, Surangika, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2021. *Neural machine translation for low-resource languages: A survey. CoRR, abs/2106.15115.*

[18] Saldanha R., Ananthanarayana V. S and Anand Kumar M and Parameswari K. (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

[19] Sennrich R., Haddow B., and Birch A. (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.

[20] Sennrich, R., Haddow, B., and Birch, A. (2016). 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, Berlin, Germany. Association for Computational Linguistics.

[21] Sutskever, I., Vinyals, O., and Le, Q. (2014). Sequence to sequence learning with neural networks. *In Advances in Neural Information Processing Systems (NIPS 2014)*

[22] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, U.; Polosukhin, I. (2017) Attention is All You Need. *In Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, CA, USA, 4–9

[23] Zhao, Y., Y. Wang, J. Zhang, and C. Zong (2018). Phrase table as recommendation memory for neural machine translation. *In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018*, July 13-19, 2018, Stockholm, Sweden., pp. 4609–4615.