Transformer-based Neural Machine Translation System for Tamil – English

Amit Kumar, Anil Kumar Singh
Department of Computer Science & Engineering
Indian Institute of Technology (B.H.U.)
Varanasi, India
{amitkumar.rs.cse17, aksingh.cse}@iitbhu.ac.in

Abstract

This paper describes the Machine Translation (MT) system submitted by the NLPRL team for the Tamil – English Indic Task at WAT 2019. We presented the Neural Machine Translation (NMT) system based on the Transformer approach. Training and performance of the model are evaluated on the En-Tam corpus (An English-Tamil Parallel Corpus) collected by researchers at UFAL (Institute of Formal and Applied Linguistics). The evaluation of the model done using Adequacy, BLEU, RIBES, and AM-FM scores, and the model improves translation in terms of Adequacy, RIBES and AM-FM as compared to the baseline.

1 Introduction

Asia[1] is home to billions of people who speaks about 2,300 languages. The population of the continent is about six times that of Europe. A majority of Asians speak languages which are, in terms of language resources and tools, low to medium resource languages. The causes of this may be historical, economic, social and political, but this fact has technical implications. There is a need to develop Machine Translation (MT) systems to bridge the communication gap between peoples of Asian countries, not just between Asian and European countries. There are continued efforts in this direction, but the lack of resources poses a challenge, which requires innovative solutions. The work presented here is not very innovative, but can be treated as an incremental step in this direction.

We discuss here our submission to the Indic Task for Tamil – English language pair (Ramasamy et al., 2012a) at the 6th workshop on Asian Translation or WAT 2019 (Nakazawa et al., 2019). Neural Machine Translation (NMT) (Sutskever et al., 2014) has been revolutionary for MT in the past few years.

Tamil comes under the family of Dravidian languages, spoken mostly in a southern state (Tamil Nadu) of India. If we consider a standard sentence in Tamil, the order is usually subject-object-verb (SOV), but object-subject-verb (OSV) is also common. While English follows subject-verb-object (SVO), therefore, Tamil-English language pairs can be considered distant language pairs. The two have very different word order, apart from other differences. Therefore, a major requirement of MT system for this language pair is to handle word order better.

2 Related work

In the last few decades, a number of works have been done on Machine Translation (MT), the initial attempt was made in the 1950s (Booth, 1955). A number of approaches have been tried out by researchers, for example, rule-based MT (Poornima et al., 2011), hybrid-based MT (Salunkhe et al., 2016), and data-driven MT (Wong et al., 2006). All of these approaches have their own advantages and disadvantages.

Rule-based approaches (Kasthuri and Kumar, 2013) cover rules based on linguistic knowledge about source and target languages in the form of dictionaries and grammars, and it covers the morphological, syntactic and semantic characteristics of each language, respectively.

Data-driven approaches rely on corpus analysis and processing. It covers Statistical Machine Translation (SMT) (Ramasamy et al., 2012b), Example-based Machine Translation (EBMT) (Carl and Way, 2003) and Neural Machine Translation (NMT) (Sutskever et al., 2014). SMT works on a large parallel corpus and does translation based on a statistical model. It relies on a combi-

[1] https://www.worldatlas.com/

Proceedings of the 6th Workshop on Asian Translation, pages 171–174
Hong Kong, China, November 4, 2019.

nation of language model as well as a translation model with decoding algorithms. On the the other hand, EBMT uses available translated examples to perform translation based on analogies. This is executed by detecting examples that coincide with the input. Then the alignment is performed to locate those parts of the translation that can be reused. Neural Machine Translation (NMT) (Sutskever et al., 2014) came into the prominence around 2014. (Choudhary et al., 2018) train an NMT model using pre-trained word-embedding (Al-Rfou' et al., 2013) along with subword units using Byte-Pair-Encoding (BPE) (Sennrich et al., 2015). Several models have been trained on various datasets and have given promising results.

Hybrid-based MT (Simov et al., 2016) is the combination of rule-based methods and any of the data-driven approaches.

Our paper describes experiments on using the transformer architecture (Vaswani et al., 2017) that we tried with English and Tamil language pair and it achieves a better result than the shared task baseline.

3 System Description

This section covers the dataset, preprocessing, and the experimental setup required for our systems.

3.1 Datasets

For the Indic Task, we use the EnTam Corpus collected by researchers at UFAL (Ramasamy et al., 2012a). EnTam Corpus contains development, training, and test data. The training data includes around 160,000 lines of parallel corpora. The data belongs to three domains: Cinema, News, and the Bible. The development and test data contain 1000 and 2000 lines of parallel corpora, respectively. Before performing training, we preprocess the data using SentencePiece library[2].

3.2 Preprocessing

NMT models usually operate on a fixed size vocabulary. Unlike most unsupervised word segmentation algorithms, which assume an infinite vocabulary, SentencePiece trains the segmentation model such that the final vocabulary size is fixed, e.g. 8000 (8K), 16k, or 32k. We tried SentencePiece on vocabulary sizes of 50,000 and 5,000 symbols. Indic sentences have a large vocabulary due to complex morphology, but size of the training data is limited. Hence, to deal with Indic corpora, we decided to use a vocabulary size of 5,000 symbols for source and target byte-pair encoding, respectively.

3.3 Experimental Setup

We trained two models, namely, Tamil – English and English – Tamil. For training the model, We use fairseq, a sequence modelling toolkit [3]. Our models are based on Transformer network. The number of encoder and decoder layers is set to 5. Encoder and decoder have embedding dimensions of 512. Embeddings are shared between encoder, decoder, and output, i.e., our model requires shared dictionary and embedding space. The embedding dimensions of encoder and decoder in the feed-forward network are set to 2048. The number of encoder and decoder attention heads are set to 2. The models are regularized with dropout, label smoothing and weight decay, with the corresponding hyper-parameters being set to 0.4, 0.2 and 0.0001, respectively. Models are optimized with Adam using $\beta_1 = 0.9$ and $\beta_2 = 0.98$. We perform the experiments on an Nvidia Titan Xp GPU.

4 Results and Analysis

RIBES (Isozaki et al., 2010), BLEU (Papineni et al., 2002), and AM-FM (Banchs et al., 2015) scores of our submitted systems are shown in Table 1, Table 2, and Table 3 resepectively. WAT 2019 organizers evaluate all the submitted system using Adequacy, BLEU, RIBES, and AM-FM scores, as shown in Figure 1 and Figure 2. It is known that Tamil and English follow different word orders, therefore we have to focus on word order for translation. On considering word order, our system performs well on RIBES metric, as shown in Figure 2. If we go through AM-FM score in Figure 2, our system still works well, keeping in view the preservation of semantic meaning and syntactic structure. Overall, if we consider Adequacy score, System beats the baseline model and top performer for English-to-Tamil among all the submitted systems as shown in Figure 1 and Figure 2.

5 Conclusion

In this paper, we report our submitted system. We train our system for Tamil-to-English and English-

[2] https://github.com/google/sentencepiece

[3] https://github.com/pytorch/fairseq

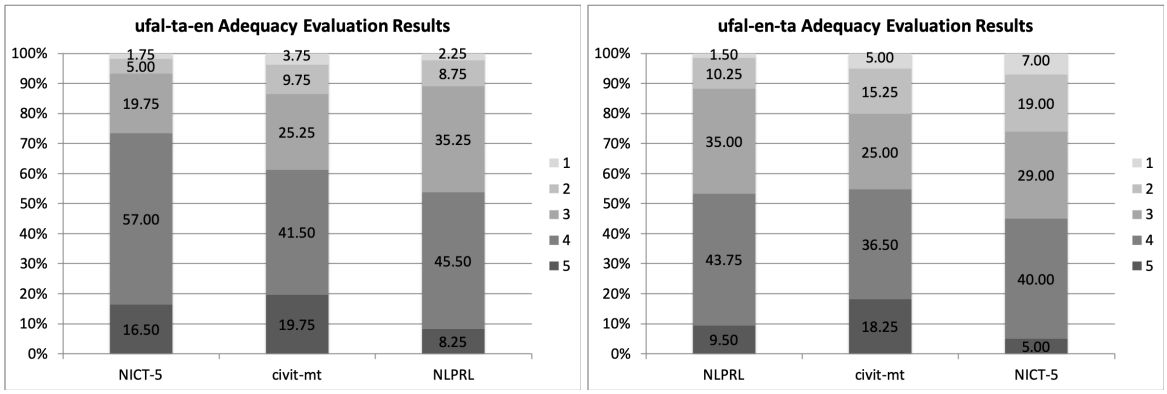

Figure 1: Official bar chart showing Adequacy Evaluation for Tamil–English and English–Tamil Indic languages shared task at WAT 2019.

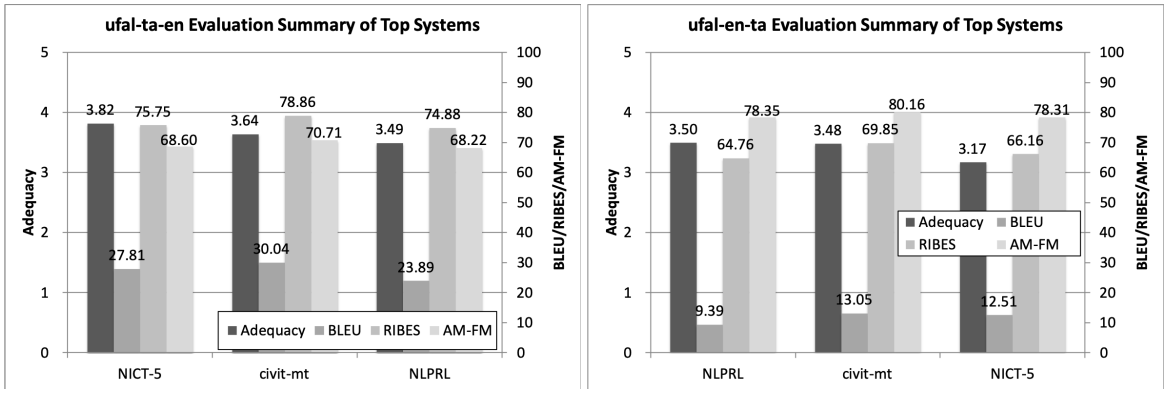

Figure 2: Official bar chart showing Adequacy, BLEU, RIBES and AM-FM scores of top systems submitted in the Tamil–English and English–Tamil Indic languages shared task at WAT 2019.

System	Baseline	Our System
Tamil–English	0.728999	0.748829
English–Tamil	0.634551	0.647579

Table 1: RIBES score of Tamil–English and English–Tamil System submitted by our team at WAT 2019.

System	Baseline	Our System
Tamil–English	24.46	23.89
English–Tamil	11.73	9.39

Table 2: BLEU score of Tamil–English and English–Tamil System submitted by our team at WAT 2019.

System	Baseline	Our System
Tamil–English	0.663930	0.682170
English–Tamil	0.769600	0.783550

Table 3: AM-FM score of Tamil–English and English–Tamil System submitted by our team at WAT 2019.

to-Tamil language pairs. The system is based on Transformer-based Neural Machine Translation. We evaluate our system using Adequacy, BLEU, RIBES, and AM-FM. Based on the official scores of Adequacy released by WAT 2019, We found that our system performs well on preserving word order and semantic-syntactic features on translation and performs better than the baseline.

References

Rami Al-Rfou', Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual NLP. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192, Sofia, Bulgaria. Association for Computational Linguistics.

Rafael E Banchs, Luis F DHaro, and Haizhou Li. 2015. Adequacy–fluency metrics: Evaluating mt in the continuous space model framework. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(3):472–482.

Andrew Donald Booth. 1955. Machine translation of languages, fourteen essays.

Michael Carl and Andy Way. 2003. *Recent advances in example-based machine translation*, volume 21. Springer Science & Business Media.

Himanshu Choudhary, Aditya Kumar Pathak, Rajiv Ratan Saha, and Ponnurangam Kumaraguru. 2018. Neural machine translation for English-Tamil. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 770–775, Belgium, Brussels. Association for Computational Linguistics.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 944–952, Cambridge, MA. Association for Computational Linguistics.

M. Kasthuri and S. Britto Ramesh Kumar. 2013. Rule based machine translation system from english to tamil. In *Proceedings of the 2013 International Conference on Information Technology and Applications*, ITA '13, pages 158–163, Washington, DC, USA. IEEE Computer Society.

Toshiaki Nakazawa, Chenchen Ding, Raj Dabre, Hideya Mino, Isao Goto, Win Pa Pa, Nobushige Doi, Yusuke Oda, Anoop Kunchukuttan, Shantipriya Parida, Ondej Bojar, and Sadao Kurohashi. 2019. Overview of the 6th workshop on Asian translation. In *Proceedings of the 6th Workshop on Asian Translation*, Hong Kong. Association for Computational Linguistics.

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 on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

C Poornima, V Dhanalakshmi, KM Anand, and KP Soman. 2011. Rule based sentence simplification for english to tamil machine translation system. *International Journal of Computer Applications*, 25(8):38–42.

Loganathan Ramasamy, Ondřej Bojar, and Zdeněk Žabokrtský. 2012a. Morphological processing for english-tamil statistical machine translation. In *Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages (MTPIL-2012)*, pages 113–122.

Loganathan Ramasamy, Ondřej Bojar, and Zdeněk Žabokrtskỳ. 2012b. Morphological processing for english-tamil statistical machine translation. In *Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages*, pages 113–122.

Pramod Salunkhe, Aniket D Kadam, Shashank Joshi, Shuhas Patil, Devendrasingh Thakore, and Shrikant Jadhav. 2016. Hybrid machine translation for english to marathi: A research evaluation in machine translation:(hybrid translator). In *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, pages 924–931. IEEE.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Kiril Simov, Petya Osenova, and Alexander Popov. 2016. Towards semantic-based hybrid machine translation between Bulgarian and English. In *Proceedings of the 2nd Workshop on Semantics-Driven Machine Translation (SedMT 2016)*, pages 22–26, San Diego, California. Association for Computational Linguistics.

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.

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 *Advances in neural information processing systems*, pages 5998–6008.

Fai Wong, Mingchui Dong, and Dongcheng Hu. 2006. Machine translation using constraint-based synchronous grammar. *Tsinghua Science and Technology*, 11(3):295–306.