NITK-UoH: Tamil-Telugu Machine Translation Systems for the WMT21 Similar Language Translation Task

Richard Saldanha,
Ananthanarayana V. S and **Anand Kumar M**
Department of Information Technology,
National Institute of Technology Karnataka
NH 66, Srinivasnagar, Surathkal, Mangalore
Karnataka 575025, India
richardsaldanha.207it005@nitk.edu.in
anvs@nitk.edu.in
m_anandkumar@nitk.edu.in

Parameswari Krishnamurthy
Centre for Applied Linguistics
and Translation Studies,
University of Hyderabad
Prof. CR Rao Road
Gachibowli, Hyderabad
Telangana 500046, India
pksh@uohyd.ac.in

Abstract

In this work, two Neural Machine Translation (NMT) systems have been developed and evaluated as part of the bidirectional Tamil-Telugu similar languages translation subtask in WMT21. The OpenNMT-py toolkit has been used to create quick prototypes of the systems, following which models have been trained on the training datasets containing the parallel corpus and finally the models have been evaluated on the dev datasets provided as part of the task. Both the systems have been trained on a DGX station with 4 - V100 GPUs.

The first NMT system in this work is a Transformer based 6 layer encoder-decoder model, trained for 100000 training steps, whose configuration is similar to the one provided by OpenNMT-py and this is used to create a model for bidirectional translation. The second NMT system contains two unidirectional translation models with the same configuration as the first system, with the addition of utilizing Byte Pair Encoding (BPE) for subword tokenization through the pre-trained MultiBPEmb model. Based on the dev dataset evaluation metrics for both the systems, the first system i.e. the vanilla Transformer model has been submitted as the Primary system. Since there were no improvements in the metrics during training of the second system with BPE, it has been submitted as a contrastive system.

1 Introduction

Tamil is a language, predominantly spoken in Tamil Nadu, a state in Southern India, along with countries with a large Tamil speaking diaspora such as Sri Lanka, Malaysia and Singapore, to name a few. Telugu on the other hand is the official language of two Southern states in India, namely Andhra Pradesh and Telangana. It is also spoken among the Telugu speaking immigrant population in the USA, Canada and the UK. Both languages belong to the Dravidian family of languages which comprise of Tamil, Telugu, Kannada and Malayalam as the major languages spoken in South India. Despite belonging to the same family of languages, there are many differences between Tamil and Telugu, such as the script used for writing and linguistic differences in terms of phonology, morphology, syntax among others. Tamil belongs to the Southern branch of Dravidian languages, which has a rich literary tradition spanning more than 2000 years. Telugu, on the other hand, belongs to the South Central branch of Dravidian languages and has a considerable amount of different linguistic characteristics when compared to Tamil as described by Krishnamurthy (2019).

As part of the similar language translation's subtask for Dravidian Languages, namely Tamil (TA) and Telugu (TE), we have attempted to build Neural Machine Translation (NMT) models using the OpenNMT-py toolkit [1], which helps to generate quick prototypes for the NMT models with the desired configurations. The first NMT system (submitted as the primary system) in this work is a Transformer based 6 layer encoder-decoder model which provides a single model for bidirectional translation between Tamil and Telugu using the datasets provided for this shared task. The second NMT system (submitted as the contrastive system) consists of two unidirectional translation models with the same configuration as the first system, but with the addition of utilizing Byte Pair Encoding (BPE) for subword tokenization using the pre-trained MultiBPEmb model (Heinzerling and Strube, 2018).

The rest of the work is described in sections that pertain to the related work, data, system descrip-

[1] https://opennmt.net/OpenNMT-py/main.html

Proceedings of the Sixth Conference on Machine Translation (WMT), pages 299–303
November 10–11, 2021.

Dataset Type	Dataset Name	Number of samples
Parallel Aligned TA-TE pairs (Training)	PM India	26009
Parallel Aligned TA-TE pairs (Training)	News	11038
Parallel Aligned TA-TE pairs (Training)	MKB	3100
Parallel Aligned TA-TE pairs (Dev)	Dev	1261
Non Aligned TA-TE sets (Test)	Test	1735 (per language set)

Table 1: Dataset statistics for parallel aligned Tamil-Telugu pairs used as train and dev (validation) datasets along with non aligned samples used as the test set.

Dataset Type	Dataset Name	Language	Longest Line Length
Training	PM India	TA	659
Training	News	TA	1524
Training	MKB	TA	412
Dev	Dev	TA	923
Test	Test	TA	1544
Training	PM India	TE	718
Training	News	TE	1356
Training	MKB	TE	376
Dev	Dev	TE	1004
Test	Test	TE	757

Table 2: Dataset statistics for Longest Line.

tion, results and conclusion.

2 Rationale for Selecting the Models and Related Work

There has been a significant amount of work done on developing machine translation systems for Indian languages, with some notable examples for Dravidian languages such as Tamil and Malayalam described in Kumar et al. (2019). This shared task provides a unique challenge in terms of the constraint on the parallel aligned language pair data made available for training. The other challenges include the linguistically rich and domain specific content present in the Prime Minister of India (PMI) and the Mann ki baat (MKB) datasets, where topics related to India's domestic and foreign policy issues can be found.

In order to address the challenge of lengthy input (samples containing more than 300 space delimited tokens), the Transformer model described by Vaswani et al. (2017) was adopted. This model provides the multi head attention mechanism which helps retain context for longer length sentence samples. To reduce the vocabulary, reduce the training time and possibly improve the translation quality (through sub word tokenization), a MultiBPEmb model trained with a vocabulary of 100000 tokens from 275 languages has been utilised (Heinzerling and Strube, 2018).

Other methods to improve translation quality, that have not been explored as part of this work are the use of back translation using monolingual corpus or corpora, on the lines of the one described by Sennrich et al. (2016). Factored NMT (which uses data tagged on the basis of morphology and Parts of Speech (POS)) such as the one described by García-Martínez et al. (2016) is another possible candidate suitable for the kind of challenge provided by the similar language translation task, as the use of POS and morphological information can reduce the number of tokens and make the models more generalizable in terms of predictions.

3 Data

The datasets used in the NMT systems for this work are the parallel aligned Tamil and Telugu (TA-TE) language pairs provided as part of the Dravidian Language sub task of the Similar Language Translation shared task[2]. Some statistics about the dataset are outlined in Table 1.

3.1 Dataset preprocessing

Due to the moderate size of the training dataset, which contains 40147 samples, along with the topic

[2] https://wmt21similar.cs.upc.edu/

Model Configuration Name	Model Configuration Value
Corpus Weights for PMI dataset	23
Corpus Weights for News dataset	19
Corpus Weights for MKB dataset	3
Source and Target Sequence Length	1600
Save checkpoint after steps	500
Number of training steps	100000
Number of validation steps	5000
Training batch size	4096
Dev(validation) batch size	16
Optimizer	Adam
Number of Encoder Decoder Layers	6 (each)
Number of Attention heads	8

Table 3: Training Configuration for Transformer based Encoder-Decoder Model (Primary System).

overlap of sentence samples between the training and dev datasets as well as test set (to a certain extent) on topics such as the Indian Prime Minister's statements on domestic issues and foreign policies in the PM India dataset, the entire training dataset has been utilized in its original form.

The length wise statistics of the dataset (in terms of space delimited tokens) is given in Table 2, this was taken as the deciding factor in fixing the maximum input length as 1600 for the NMT systems developed. The tokenization for the primary system was done as space delimited tokens which yielded a shared Tamil-Telugu vocabulary of 194860 tokens. On the other hand on using the MultiBPEmb model for subword tokenization gave a vocabulary of 14056 tokens for Tamil (TA) and 13170 tokens for Telugu (TE), which included some words in English as well.

4 System Description

As mentioned in section 1, the PyTorch based toolkit OpenNMT-py has been used to create rapid prototypes for NMT models (the motivations for the same can be seen in section 2), which have then been trained on the datasets provided, validated against the provided dev sets and finally translations for the test sets described in section 3 have been obtained and submitted to the committee for evaluating the Similar Language Translation task.

A DGX station with 4 - V100 GPUs have been used to train the models utilized in this task. A Transformer based 6 layer encoder-decoder model on the lines of the NMT system described by Vaswani et al. (2017), was trained for 100000 training steps as the first NMT system to be evaluated. The configuration for this model is the same as that provided by OpenNMT-py. In order to save time, a single bidirectional translation model for TA-TE language pair has been created, which can translate from Tamil to Telugu and vice versa. The datasets used in this system were doubled in terms of the number of samples when compared to the second NMT system (constrastive submission), by reversing the position of the TA-TE language pair and appending them to the original datasets. No special tagging identifiers were used as the Tamil and Telugu scripts are distinct.

Basic space delimited tokenization was applied on the datasets, which resulted in a combined TA-TE vocabulary of 194860 tokens being generated, the relevant key configuration for this model are listed in Table 3.

The corpus weights help assign varied importance to the particular datasets used in this task, the values for these weights were determined after visual analysis of the dev(validation) dataset which indicated the dev dataset's contents had a greater overlap with PMI, News and (Mann ki Baat - which roughly translates to "From the heart") MKB in that particular order. The training time for the entire model was 18 hours.

The second NMT system consists of two unidirectional translation models with the same configuration as the first system, with the addition of utilizing Byte Pair Encoding (BPE) for subwords using the pretrained MultiBPEmb model (Heinzerling and Strube, 2018). The intuition behind using BPE was to reduce the vocabulary size using subword tokenization. The choice of the pre trained BPE model was based on the relevance of content

System Name	Source Language	Target Language	BLEU	RIBES	TER
Primary System (Transformer Based)	TA	TE	4.321	7.4	99.1
Contrastive System (Transformer Based + BPE subword)	TA	TE	0.003	0.0	130.6
Primary System (Transformer Based)	TE	TA	3.908	9.0	98.7
Contrastive System (Transformer Based + BPE subword)	TE	TA	0.029	3.0	105.0

Table 4: Dev dataset BLEU, RIBES and TER Corpus level scores using the VizSeq library.

System Name	Source Language	Target Language	BLEU	RIBES	TER	System Rank
Primary System	TA	TE	6.09	17.03	-	1
Contrastive System	TA	TE	0.00	0.03	-	9
Primary System	TE	TA	6.55	19.61	98.356	4
Contrastive System	TE	TA	0.04	1.00	-	9

Table 5: Test dataset BLEU, RIBES, TER scores and BLEU based System Rank in the Shared Task

used for BPE model training, languages supported and size of the vocabulary. Heinzerling and Strube (2018) describes a MultiBPE model with a 100000 vocabulary which was deemed suitable for this task as it supported Tamil and Telugu, was trained on WikiNews and could use a single vocabulary like the first NMT system used in this work. During training it was found that the translations for the Dev set couldn't distinguish between Tamil and Telugu subwords correctly, due to the failure in vocabulary matching for the candidates used in the evaluation and possibly due to the vocabulary shared between the languages. Hence, this system was trained twice generating two unidirectional models for TA-TE and TE-TA translations. The training time for each model was 5 hours, which is less when compared to the primary system due to the number of samples used (the primary system uses double the number of samples) and the vocabulary size (the contrastive system has a smaller and fixed vocabulary as a pre trained BPE model has been used).

5 Results

The evaluation metrics used to evaluate the systems in this task are BiLingual Evaluation Understudy (BLEU) score as described by Papineni et al. (2002), Rank-based Intuitive Bilingual Evaluation (RIBES) score as described by Isozaki et al. (2010) and Translation Error Rate (TER) as described by Snover et al. (2006).

Corpus level metrics for the dev dataset were computed using the VizSeq python library which is an implementation of several metrics described by Wang et al. (2019).The metrics for the dev dataset are listed in Table 4.

Based on the evaluation metrics of the Dev (validation) dataset translations for both the systems evaluated in this work, the first system i.e. the vanilla Transformer model has been submitted as the Primary system. Since there were no improvements in the metrics (the reason for it can be seen in section 6), during training of the second system which consists of the Transformer model along with the use of MultiBPEmb model for sub word tokenization, hence the second system has been submitted as a contrastive system.

Table 5 lists the evaluation metrics[3] applied on the test dataset and the BLEU based system rank in the shared task provided by the evaluation committee [4,5].

6 Conclusion and Future Work

The analysis of the evaluation metrics, from section 5, on the dev dataset indicates that the primary system, which is a Transformer based Encoder-

[3]The results of the TER metrics for the test set translations have been marked as - (refer Table 5), when the values exceed 100.0

[4]https://mzampieri.com/workshops/wmt/2021/TA_TE.pdf

[5]https://mzampieri.com/workshops/wmt/2021/TE_TA.pdf

Decoder model, performs better than the contrastive system which contains Transformer based NMT models with BPE for subword tokenization. The reason for this is possibly due to the lack of vocabulary matching the candidates being evaluated and also due to the shared vocabulary of the MultiBPEmb model. The choice of a pre trained MultiBPE model was to reduce effort on the embeddings, but in hindsight training the MultiBPE model using the given datasets or fine tuning the pre trained MultiBPE model on the given datasets would have been a better choice.

As seen from the evaluation of translations obtained using the Dev and Test datasets using BLEU, RIBES and TER metrics in section 5, there is a considerable scope of improvement in the scenario where a constraint is placed on the number of datasets containing parallel corpus language pair samples, that can be used for training. The possible reason for the low BLEU scores in the primary system is the relatively small number of samples used along with the presence of a large variety in the linguistic forms present in the datasets. In the case of the contrastive system, the low BLEU scores can be attributed to the use of the pre trained MultiBPE model (a pre trained BPE model fine tuned on the given datasets would have helped improved the scores). Some approaches that have the potential to improve the results are, the use of back translation using monolingual corpus (through training corpus augmentation and providing more training examples for the model to learn), utilizing domain specific corpora from the shared machine translation task for Indian Languages described in section 2. Factored NMT, an NMT which uses input tagged on the basis of morphology and Parts of Speech (POS) to reduce the number of tokens, the use of alternative BPE models trained on content which are a close match to the dataset used in the shared task, are other promising alternatives.

References

Mercedes García-Martínez, Loïc Barrault, and Fethi Bougares. 2016. Factored neural machine translation architectures. In *International Workshop on Spoken Language Translation (IWSLT'16)*.

Benjamin Heinzerling and Michael Strube. 2018. BPEmb: Tokenization-free Pre-trained Subword Embeddings in 275 Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 2989–2993, Miyazaki, Japan. European Language Resources Association (ELRA).

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.

Parameswari Krishnamurthy. 2019. Development of telugu-tamil transfer-based machine translation system: An improvization using divergence index. *Journal of Intelligent Systems*, 28(3):493–504.

M. Anand Kumar, B. Premjith, Shivkaran Singh, S. Rajendran, and K. P. Soman. 2019. An overview of the shared task on machine translation in indian languages (mtil) – 2017. *Journal of Intelligent Systems*, 28(3):455–464.

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, Barry Haddow, and Alexandra Birch. 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.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231.

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.

Changhan Wang, Anirudh Jain, Danlu Chen, and Jiatao Gu. 2019. Vizseq: A visual analysis toolkit for text generation tasks. `https://arxiv.org/pdf/1909.05424.pdf`. (Accessed on 08/04/2021).