MUCS@ - Machine Translation for Dravidian Languages using Stacked Long Short Term Memory

Asha Hegde
Dept of Computer Science
Mangalore University
hegdekasha@gmail.com

Ibrahim Gashaw
Dept of Computer Science
Mangalore University
ibrahimug1@gmail.com

H. L. Shashirekha
Dept of Computer Science
Mangalore University
hlsrekha@gmail.com

Abstract

The Dravidian language family is one of the largest language families in the world. In spite of its uniqueness, Dravidian languages have gained very less attention due to scarcity of resources to conduct language technology tasks such as translation, Parts-of-Speech tagging, Word Sense Disambiguation etc. In this paper, we, team MUCS, describe sequence-to-sequence stacked Long Short Term Memory (LSTM) based Neural Machine Translation (NMT) models submitted to "Machine Translation in Dravidian languages", a shared task organized by EACL-2021. The NMT models are applied for translation using English-Tamil, English-Telugu, English-Malayalam and Tamil-Telugu corpora provided by the organizers. Standard evaluation metrics namely Bilingual Evaluation Understudy (BLEU) and human evaluations are used to evaluate the model. Our models exhibited good accuracy for all the language pairs and obtained 2nd rank for Tamil-Telugu language pair.

1 Introduction

Human being is a social entity and they love to communicate with each other and live together from ancient times. They communicate with each other through different means of communication and exchange their data/information/thoughts with each other. Initially, sign language was the means of communication and it was used to exchange thoughts with each other, but now, in this era of technology, there are many ways to communicate out of which language plays a major role. There are more than thousands of languages used all over the world, even in India, due to different religion, culture and tradition. Among Indian languages Dravidian is a language family consisting of four long literary languages, namely, Tamil, Kannada, Telugu and Malayalam along with Tulu and Kodava which are small literary languages. All these languages except Kodava have their own script. Further, these languages consists of 80 different dialects[1] namely Brahui, Kurukh, Malto, Kui, Kuvi, etc. Dravidian Languages are mainly spoken in southern India, Sri Lanka, some parts of Pakistan and Nepal by over 222 million people (Hammarström et al., 2017). It is thought that Dravidian languages are native to the Indian subcontinent and were originally spread throughout India[1]. Tamil have been distributed to Burma, Indonesia, Malaysia, Fiji, Madagascar, Mauritius, Guyana, Martinique and Trinidad through trade and emigration. With over two million speakers, primarily in Pakistan and two million speakers in Afghanistan, Brahui is the only Dravidian language spoken entirely outside India (Ethnologue)[2]. Rest of the Dravidian languages are extensively spoken inside India and by south Indians settled throughout the world. These languages share similar linguistic features and few of them are listed below (Unnikrishnan et al., 2010):

- Verbs have a negative as well as an affirmative voice.
- Root word can be extended using one or more suffixes or prefixes according to the comfort level of the speaker.
- Phonology of all Dravidian languages follow similar strategy.
- Though these languages follow Subject-Object-Verb word order, they are considered as free word order languages as meaning of the sentence will not change when the word order is changed.

[1] https://www.shh.mpg.de/870797/dravidian-languages
[1] https://www.mustgo.com/worldlanguages/dravidian-language-family
[2] http://self.gutenberg.org/articles/eng/Languages_of_Pakistan

Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, pages 340–345

- Gender classification is made based on suffix.
- Nouns are declined, showing case and number.
- These languages have their own alphabets related to the Devanagari alphabet that is used for Sanskrit.

Though these languages have rich collection of resources, they are still considered as under-resourced languages[4] due to the availability of very less digital resources and tools. Most of the south Indians speak/understand only their native language as a means of communication. Further, because of migration people need to learn the local languages to survive. In this regard, Language Translation (LT) technology gains importance as it provides the easy means of communication between people when language barrier is a major issue. Various LT technologies are available, namely, manual translation or human translation, Machine aided translation and automatic translation or Machine Translation (MT). As MT is fast, inexpensive, reliable and consistent compared to other LT technologies, they are gaining popularity.

1.1 Machine Translation Approaches

MT is an area in Natural Language Processing and does translation of information from one natural language to another natural language by retaining the meaning of source context. Initially, MT task was treated with dictionary matching techniques and upgraded slowly to rule-based approaches (Dove et al., 2012). In order to address information acquisition, corpus-based methods have become popular and bilingual parallel corpora have been used to acquire knowledge for translation (Britz et al., 2017). Hybrid MT approaches have also become popular along with corpus-based approaches, as these approaches promise state-of-the-art performance.

The recent shift to large-scale analytical techniques has led to substantial changes in the efficiency of MT. It has attracted the attention of MT researchers through a corpus based approach. NMT has now become an important alternative to conventional Statistical Machine Translation based on phrases (Patil and Davies, 2014). It is the task of translating text from one natural language (source) to another natural language (target) using the most popular architectures namely Encoder-Decoder, Sequence-to-Sequence or Recurrent Neural Network (RNN) models (Sutskever et al., 2014). In addition, all parts of the neural translation model are trained jointly (end-to-end) to optimise translation efficiency unlike traditional translation systems (Bahdanau et al., 2014). In an NMT system, a bidirectional RNN, known as encoder is used to encode a source sentence and another RNN known as decoder is used to predict the target language terms. With multiple layers, this encoder-decoder architecture can be built to improve the translation performance of the system. The rest of the paper is organized as follows: Related work is presented in Section 2 followed by Methodology and Dataset in Section 3. Experiments and Results are given in Section 4 and the paper concludes in Section 5.

2 Related Work

Many attempts are being carried out by the researchers to give special attention for Dravidian language family and several researches have made noticeable work in this direction (Chakravarthi et al., 2019a). NMT is a promising technique to establish sentence level translation and suitable preprocessing techniques will share their contribution for better performance. Observing this (Choudhary et al., 2018) have applied Byte Pair Encoding (BPE) for their model to resolve Out-of-Vocabulary problem. They used EnTam V2.0 dataset that contains English-Tamil parallel sentences. Their model exhibited improvement in BLEU score of 8.33 for Bidirectional LSTM+Adam (optimizer) + Bahdanau (attention) + BPE + word Embedding. Multi-modal multilingual MT system utilizing phonetic transcription was implemented by (Chakravarthi et al., 2019b) using under-resourced Dravidian languages. As a part of this work they have released MMDravi - a Dravidian language dataset that comprises of 30,000 sentences. They have conducted both Statistical Machine Translation (SMT) and NMT and the NMT model outperformed SMT in terms of Bilingual Evaluation Under Study (BLEU) score. (Pareek et al., 2017) have proposed a novel Machine Learning based translation approach for Kannada-Telugu Dravidian language pair considering wikipedia dataset. Further, they considered English-Kannada and English-Telugu language pairs for illustrating the efficacy of their model. They have proposed n-grams based connecting phrase extraction and these extracted

[4]https://en.unesco.org/news/towards-world-atlas-languages

phrases are trained in multilayered neural network. For testing, the phrases are extracted from test dataset and alignment score is computed and then these phrases are used for post-processing. They have observed a considerable accuracy of 91.63%, 90.65% and 88.63% for English-Kannada, English-Telugu and Kannada-Telugu, respectively. Dictionary based MT for Kannada to Telugu is developed by (Sindhu and Sagar, 2017) and as a part of this work, a bilingual dictionary with 8000 words is developed. Providing suffix mapping table at the tag level, this model uses dictionary for translating word by word without giving much attention on the correlation between words and has shown more than 50% accuracy.

3 Methodology

Though many attempts are being carried out to develop MT system for Dravidian languages, development of full fledged MT system for Dravidian languages is an open ended problem as these languages are agglutinative and morphologically rich (Chakravarthi, 2020). In the proposed work, sequence-to-sequence stacked LSTM is used to build the translation system using the dataset provided by the organizers[5].

3.1 Sequence-to-Sequence Architecture

Sequence-to-sequence architecture is basically used for response generation and it is suitable when source and target sentences are almost of same length. Further, it is used to find the relationship between two distinct language pairs in MT model. This architecture consists of two parts, namely, i) the encoder that accepts the text of the source language as input and generates its intermediate representation and (ii) the decoder that generates the output based on the encoding vector and previously generated words. Since, the model utilises the source representation information and the previously generated words to predict the next-word, this distributed representation allows the sequence-to-sequence model to generate appropriate mapping between the input and the output (Li et al., 2016). Therefore, this model has shown good performance compared to conventional statistical MT systems and rule based MT systems. Suppose, S is source consisting of n sentences s_1, s_2, s_3, ... , s_n and R is target. Encoder transforms source sentences into fixed dimension vectors A_{s1},A_{s2}, A_{s3}, ... , A_{sn} and the decoder uses conditional probability to produce the predicted sentences b_1, b_2, b_3, ... , b_n word-by-word. While decoding, next word is predicted using previously predicted word vectors and source sentence vectors in equation 1 and equation 2 is derived from the equation 1. Each term in the distribution is represented with a softmax over all the words in the vocabulary (Neubig, 2017).

$$P(R\,|S) = P(R\,|A_{s1}, A_{s2}, ..., A_{sn}) \quad (1)$$

$$P(R\,|S) = P(b_1, b_2, ...b_n; s_1, s_2, ..., s_n) \quad (2)$$

3.2 Stacked Long Short Term Memory

LSTM is an RNN architecture used to address the issues of RNN, namely, long term dependence and gradient disappearance. Like RNN, LSTM has a chain like system, but the structure of the repeating module is distinct from RNN (Tai et al., 2015). LSTM has the ability to add and remove information to the cell state by carefully regulating the structure called gates. There are four gates in a module in place of a single neural network layer. Forget gate is the first gate that shows one of the major features of LSTM network. It helps in determining the portion that should be forgotten by the network. In this architecture, forgetting is as important as learning. If network is able to forget unimportant information and just retain important details, this will reduce the memory requirement of the system. Next gate is based on sigmoid layer and known as input gate that helps to determine which values are changed by the system. This gate is useful in deciding the role of current layer by choosing the right inputs to improve learning. The third gate is a input modulation gate (tanh) that generates a new candidate value vector that could be added in the same module to the state. Finally, the fourth gate ie., output gate determines the production. This is also a function of *tanh* that produces the state for the next modules.

Conventional LSTM model has single hidden LSTM layer followed by a standard feed forward output layer whereas the stacked LSTM model has multiple LSTM layers along with multiple hidden layers. Stacking of LSTM makes the model more deeper, more accurate and gives the correct meaning of deep learning (Barone et al., 2017). In recent years, stacked LSTM model is becoming a stable approach for problems with sequence prediction. A stacked LSTM architecture can be described as

[5]https://dravidianlangtech.github.io/2021/index.html

Table 1: Details of the given dataset used for training and validating the model

Corpus name	# of parallel sentences used for training	# of parallel sentences used for validation	# of words
English-Telugu	23,222	2000	13,05,105
English-Tamil	28,417	2000	17,44,747
English-Malayalam	3,82,868	2000	2,55,97,816
Tamil-Telugu	17,155	2000	9,40,921

LSTM model comprised of multiple LSTM layers. The LSTM layer above provides a sequence output rather than a single value output to the LSTM layer below, specifically, one output per input time step rather than one output for all input time steps (Barone et al., 2017). This information is shown in Figure 1 and it consists of two LSTM layers namely Layer 1 and Layer 2. Further, in Figure 1 X_i is the encoder input, Y_i is the decoder input and Z_i is the predicted sequence (decoder output) obtained using multilayered LSTM architecture.

3.3 Dataset

In MT, dataset and preparation of the dataset for training the translation model play a major role. To perform successful translation, this data preparation process should be carried out at different levels. In this work, we have used English-Telugu, English-Tamil, English-Malayalam and Tamil-Telugu the dataset. Details of the dataset are shown in Table 1.

3.4 Dataset Preprocessing

Insights of the dataset makes the translation task more effective and efficient. As many sentences in the datasets were duplicated both in train and test sets, it becomes essential to clean, analyse and correct the dataset before using it for the translation experiments. It was also observed that few sentences are repeated more than 5 times and this confuses the model to learn and recognise various new characteristics and overfits the model by leading to wrong output. Train and test set containing the same sentences may give better prediction for test set but poor prediction for new sentences. To resolve these problems, unique sentences are selected from the corpus. Further, duplication of sentences with same source and target text are removed to avoid uncertainty during training. Null lines cause mapping issues while training the corpora for translation. Hence, null lines and extra spaces are removed . The remaining parallel text is considered for building the translation model.

Table 2: Hyper parameters used in sequence-to-sequence stacked LSTM model

Hyper Parameters used	Values
latent dimension	512
optimizer	adam
learning rate	0.001
dropout	0.03
epochs	100
batch size	128

4 Experiments and Result

Sequence-to-Sequence stacked LSTM model is used to build translation model for the given dataset. This model is developed with multiple layers to enhance learning ability of the model. Further, in this work we used one-hot encoding embedding technique for both source and the target text. The experiments are conducted in Google Colab to resolve resource issues during translation and is conducted for various hyper parameter values and best values are shown in Table 2.

4.1 Result

In this work, translation is carried out separately for the given corpora, namely, English-Tamil, English-Telugu, English-Malayalam and Tamil-Telugu. Further, this model is evaluated using BLEU as well as human evaluation against gold dataset which is provided by the organizers and performance measure of the models are as shown in Table 3. Though there are many challenges with the test dataset, considerable results are obtained for all the corpora.

4.2 Analysis

In this work, Tamil-Telugu translation model achieved 2nd rank with a BLEU score of 0.43. Further, the models exhibited BLEU scores of 1.66, 0.29 and 0.48 for English-Tamil, English-Telugu and English-Malayalam corpora respectively. In this sequence-to-sequence model, number of LSTM layers is varied to get better translation performance and for 3 LSTM layers models achieved good accuracy. When the number of LSTM layers is increased beyond this break even

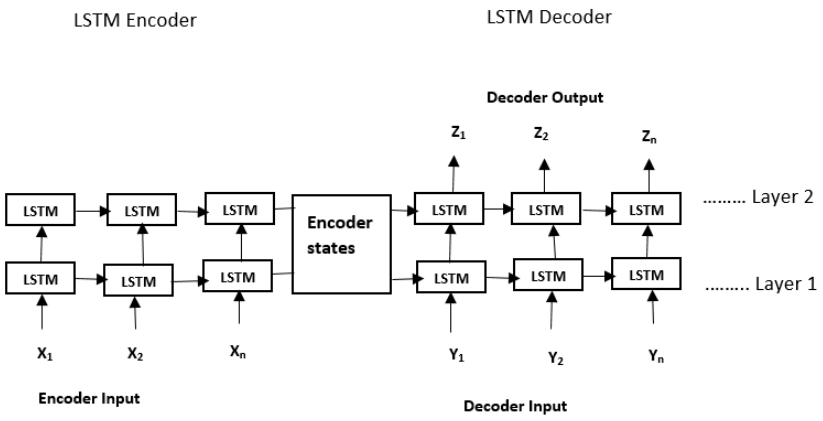

Figure 1: Stacked LSTM architecture with two layers

Table 3: Performance measure of Sequence-to-Sequence stacked LSTM model

Corpus Name	Validation Accuracy	Validation Loss	BLEU Score
English-Tamil	37.77	1.04	1.66
English-Telugu	32.87	1.32	0.29
English-Malayalam	42.55	0.88	0.48
Tamil-Telugu	**33.26**	**1.27**	**0.43**

point performance of the models started decreasing.

This sequence-to-sequence stacked LSTM models have exhibited considerable BLEU score as the size of the corpora is small to conduct efficient translation. The morphological richness and agglutinativeness of the language pairs used in translation increases the translation complexity. In addition, test sets provided for these tasks are challenging due to the presence of special characters. Further, long sentences, duplicate sentences and null lines in the test sets had their share to the translation complexities.

5 Conclusion and Future Work

This paper describes the models submitted to "Machine Translation in Dravidian languages" shared task to perform translation of English-Tamil, English-Telugu, English-Malayalam and Tamil-Telugu language pairs. A close analysis of the corpora is carried out before performing the preprocessing and the corpora are found challenging to establish translation. Though all the corpora belong to low-resource languages and being rich in morphology, this sequence-to sequence stacked LSTM model exhibits the satisfactory results for all the corpora. All the models exhibited considerable performance and our model obtained 2nd rank for Tamil-Telugu language pair. In future, we would like to explore different NMT techniques that helps to translate low-resource languages efficiently.

References

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

Antonio Valerio Miceli Barone, Jindřich Helcl, Rico Sennrich, Barry Haddow, and Alexandra Birch. 2017. Deep architectures for neural machine translation. *arXiv preprint arXiv:1707.07631*.

Denny Britz, Anna Goldie, Minh-Thang Luong, and Quoc Le. 2017. Massive exploration of neural machine translation architectures. *arXiv preprint arXiv:1703.03906*.

Bharathi Raja Chakravarthi. 2020. *Leveraging orthographic information to improve machine translation of under-resourced languages*. Ph.D. thesis, NUI Galway.

Bharathi Raja Chakravarthi, Mihael Arcan, and John P McCrae. 2019a. Comparison of different orthographies for machine translation of under-resourced dravidian languages. In *2nd Conference on Language, Data and Knowledge (LDK 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Bernardo Stearns, Arun Jayapal, S Sridevy, Mihael Arcan, Manel Zarrouk, and John P McCrae. 2019b. Multilingual multimodal machine translation for Dravidian languages utilizing phonetic transcription. In *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*. European Association for Machine Translation.

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.

Catherine Dove, Olga Loskutova, and Ruben de la Fuente. 2012. What's your pick: RbMT, SMT or hybrid. In *Proceedings of the tenth conference of the Association for Machine Translation in the Americas (AMTA 2012). San Diego, CA*.

Harald Hammarström, Robert Forkel, and Martin Haspelmath. 2017. Glottolog 3.0. *Max Planck Institute for the Science of Human History*.

Xiaoqing Li, Jiajun Zhang, and Chengqing Zong. 2016. Towards zero unknown word in neural machine translation. In *IJCAI*, pages 2852–2858.

Graham Neubig. 2017. Neural machine translation and sequence-to-sequence models: A tutorial. *arXiv preprint arXiv:1703.01619*.

Piyush Kumar Pareek, K Swathi, Puneet Shetteppanavar, et al. 2017. An efficient machine translation model for Dravidian language. In *2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, pages 2101–2105. IEEE.

Sumant Patil and Patrick Davies. 2014. Use of google translate in medical communication: evaluation of accuracy. *Bmj*, 349:g7392.

DV Sindhu and BM Sagar. 2017. Dictionary based machine translation from Kannada to Telugu. In *IOP conference series: materials science and engineering*, volume 12182.

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.

Kai Sheng Tai, Richard Socher, and Christopher D Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075*.

P Unnikrishnan, PJ Antony, and KP Soman. 2010. A novel approach for English to South Dravidian language statistical machine translation system. *the International Journal on Computer Science and Engineering*, 2(8):2749–2759.