International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 06 | June 2021 www.irjet.net p-ISSN: 2395-0072

A GRU Based Marathi to Gujarati Machine Translator

Bhavna Dewani

[1]*M. Tech., Dept. of Computer Engineering, K. J. Somaiya College of Engineering, Mumbai*

Abstract - *One of the most frequent challenges while translating any regional language to another, the limited resources and corpora for dialects, pose a threat to non-English translations. Machine translation is spearheading research areas of Language Translation through Computer Linguistics. And through Neural Machine Translation, the accuracy and scope of machine translation are improved tenfold. This paper introduces a GRU-based NMT system for Marathi-Gujarati translation.*

Key Words: Neural Machine Translation, Recurrent Neural Network, BLEU score.

1.INTRODUCTION

Language has been an important and distinctive property of human beings. It's more than words and nuances. It's a way of communicating ideas. Speech and text are one of the fastest forms of communication for humans. There's a wide range of human languages in every corner of the world, each with its own grammar, dialects, words etc. The differences in the processing of a language are what makes them complex. Although any language is complex, understanding and learning it is not a hardship once developed through the basics. This process is no different than toddlers learning the first-ever language of their lives. Toddlers are taught to learn by breaking down the language into steps (algorithms) for better understanding. The combination of these is used to teach computers about languages through NLP.

Natural Language Processing includes the logic to involve language in the virtual world. The end result of this process gives a user-friendly interface and systems to interact better with humans. NLP being the center of three major fields-algorithm, intelligence and linguistics, as shown in Fig 1., involves various challenges. We will dig into Machine Translation (MT) that gears up the use of programming to translate one language, text or speech into another [1].

Machine Translation is not a new idea, it is used first in 1940 since the Cold War of the U.S. and Russia. It was first used in 1933 by Artsrouni, and by Troyanskii [2]. Artsruni designed a paper tape storage system in the form of a dictionary to find synonyms in another language.

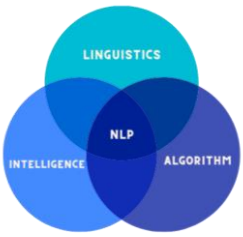

Fig -1: Natural Language Processing

Troyanskii proposed the three-stage machine translation process where the first step person responsible for the source understanding undertakes the analysis of words into their raw form. The second step involves the conversion of sequences of source base form into target base forms. The third step was to convert the target base forms into its normal form.

In 1949, MT was used to understand wartime cryptography techniques and statistical methods [2]. Much work of MT was done by the 1960s in advance linguistics. In 1966, the ALPAC reported that human translator beats the qualitative purpose MT. Then, IBM took the buck in 80s began research on Translation using statistical machine translation to achieve the goal of automatic translation. Up until 2014, many approaches were used but Kyung Hyun Cho [3] made a revolutionary change. They build a black box system using deep learning to improvise the translation every time by using the trained model data. This revolution in MT is known as Neural Machine Translation [4].

The issue here with NMT is, it is not very well-developed for the regional languages. So, we propose this solution a Gated Recurrent Unit (GRU)-based Machine Translation System for the translation of two of the most popular as well as polar languages i.e., Marathi and Gujarati.

The flow of the paper is as follows - Section 2 provides a literature survey of MT Approaches. Section 3 includes an overview and methodology of the proposed system. Section 4 presents the proposed system implementation. Section 5 produces the observation & evaluation matrix of the experiment and the result of the proposed system. Finally, we conclude with the observations and future work.

2. BACKGROUND

2.1 Neural Machine Translation

Using the neural network models in order to create a model for machine translation is commonly known as Neural Machine Translation. Earlier MT emphasized on the data-driven approach. However, the statistical approach brought the tuning of the translation of each module. But NMT attempted to build a single large neural network that reads a sentence and outputs a correct translation. In other words, NMT is an end-to-end system which can learn directly and map the output text accordingly.

NMT consists of two primary components - encoders and decoders [5]. The encoder summarizes the input sentence in the source language in thought vector. This thought vector is fed as an input to the decoder to elaborate it as the output sentence as shown in Fig. 2.

2.2 Recurrent Neural Networks

RNNs are neural networks designed for sequence modules. They are generalized as feed-forward networks for sequential data [6]. For e.g., a recurrent neural network can be thought of as the addition of loops to the architecture. To learn broader abstractions from the input sequences, the recurrent connections add memory to the network or another state.

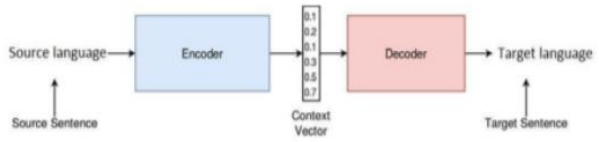

Fig -2: Conceptual Structure of Neural Machine Translator

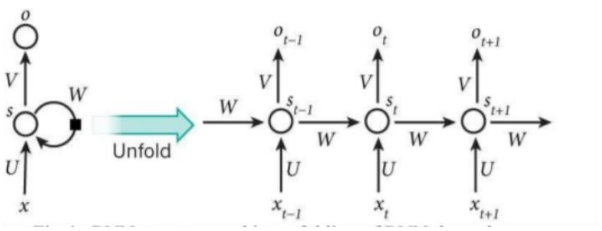

Fig-3. RNN structure and its unfolding of RNN through time.

The structure of the RNN is as shown in Fig. 3. It shows a single step calculation of an RNN structure at time t. For calculating the output at time t, it takes input from step t-1. And it goes without saying that for calculating the output at time t+1, it will take input from t. This shows that RNN is very useful for sequential data. This is known as the unfolding of the recurrent neural network. The goal of RNN is to evaluate conditional probability ρ*(T1, ..., Tt S1, ..., St)* where *(S1, ..., St)* is an input sentence in the source language and *(T1, ..., Tt)* is its corresponding output sentence in the target language.

2.3 LSTM & GRU

Long Short-Term Memory & Gated Recurrent Unit are types of Neural Network, both having a gating mechanism.

In normal RNN, there is a single layer to pass the states like input and hidden. Whereas, LSTM has more gates. There is also a cell state. that it fundamentally addresses the problem of keeping or resetting context across sentences and regardless of the distance between such context resets. LSTMs and GRU help in utilizing gates in different manners to address the problem of long-term dependencies [7].

LSTM, a variant of RNN is well-renowned for its long-range temporal dependencies that help in learning problems. Thus, RNNs are sometimes replaced with LSTMs in MT networks as they may succeed better in this setting. However, LSTM is a complex model while GRU is a much-simplified model as compared to any other. It has fewer parameters than LSTM [8]. That's why we propose a GRU-based model for our Machine Translation Project. Fig. 4. shows the internal cell structure of LSTM and GRU. In LSTM (Fig. 4(a) i, f, o is input, forget and output gates respectively. c and c denote the memory cell and new memory content. In GRU (Fig. 4(b) r and z are reset and update gates and h and h are the activation and candidate activation.[9][10].

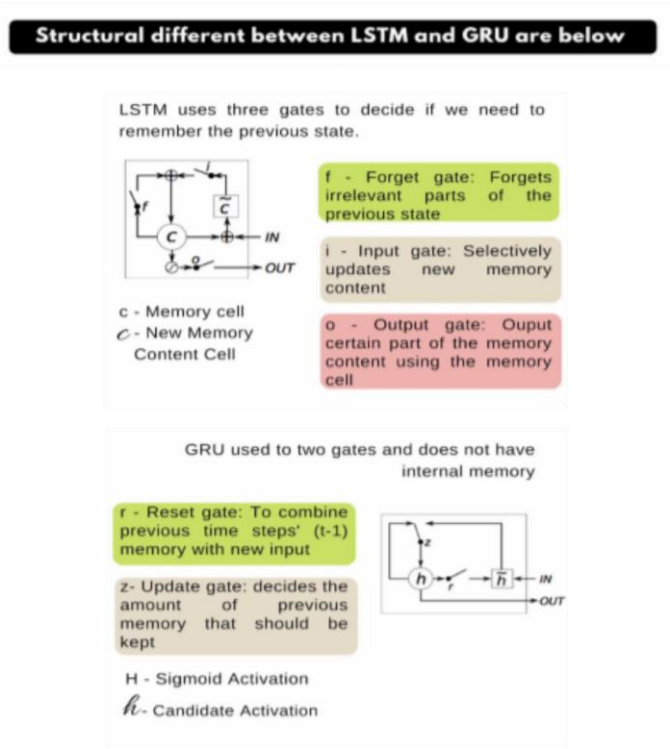

Fig-4. (a) LSTM [9], (b) GRU [10]

In the WMT19 paper [11], the methodology starts with the translation of Gujarati (News Domain) into English and later the system translates it to the regional language. This adds another layer to the process and complicates more. Not to forget the nuances lost during translation. The system used here is a multilingual setting where first a high resource language (English) is trained on the system using the Hindi-English corpus and then the low resource language (Gujarati). As the future works of this paper suggest, creating a one-to-many system for multilingual low resource language. Alternatively, we can also work on the translation of low resource language to a high resource language without a leverage parallel corpus.

3. SYSTEM ARCHITECHTURE

The NMT architecture is comprised of a encoder to decoder layered model. It is based on Gated Recurrent Unit NMT model as shown in Fig. 5. inspired by Yonghui[12].

Tokenization and embedding are the two constituent tasks during pre-processing. Tokenizer separates the words in the input sentence and give each word an integer value. The values given are based on the frequency of the occurrence of the words. The higher the integer the lower the frequency. For e.g., if 'house' is repeated more than 'home' than 'house' will have low integer value than 'home' during translation.

The language taken into consideration for this task is the source language (Marathi) and the sentences are targeted separately. Since source language (Marathi) and destination language (Gujarati) have different word-list also known as vocabularies we will look into both languages separately. The first layer of the model is the embedding layer where the semantic meaning of the words is taken. The layer then converts each value into vectors. Additionally, two markers are added to identify the beginning and end of the sentence. These markers are independent to the words in both Marathi and Gujarati corpus.

The model is trained after applying the above pre-processing steps. The neural network then summarizes the sentences each into a thought vector using word2vec encoding. The weights of the neural network are assigned in the manner that the repeated sentence should have same thought vector. Also, when decoded, the thought vector will decode the corresponding similar sentence for the target - Gujarati Language.

4. EXPERIMENTAL SETUP

The proposed approach of the task: Marathi-Gujarati. translation is thoroughly evaluated. Bilingual parallel corpora were used for India's PM narration of Radio - Mann ki Baat address to the nation [13]. The dataset consists of 7000 parallel sentences which we can update as more episodes get added to it. Also, to achieve better results, we have added this on 3000 parallel sentences obtained from web scraping.

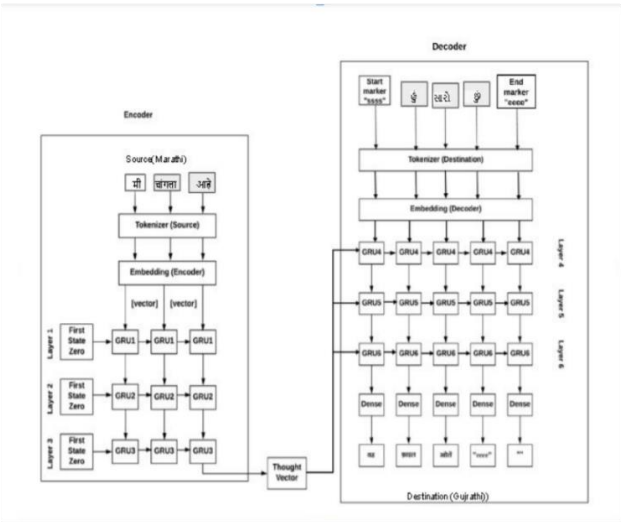

Fig-5. Flowchart of the architecture

The validation of the training is done on 0.21 of the training data. The vocabulary size taken into consideration is close to 1k unique words which can be modified based on the size of the corpora as mentioned in the standards followed by Google Neural Machine Translation [12]. Keras tokenizer [14] is used for the pre-processing of the dataset. The tokenizer is applied on both the source and the target language i.e., Marathi as well as Gujarati. datasets separately due to the different vocabularies. The loss function used was a custom-defined sparse cross-entropy function rather than a built-in python. BLEU score evaluation metric is used to test or grade the model.

5. OBSERVATIONS AND RESULT

The NMT model results are shown in the following table where the BLEU score is calculated using NLTK library of python [15].

Table -1: BLEU SCORE USING NLTK

Sr. No.	BLEU Score		
	Description	*Individual Score*	*Cumulative Score*
1.	Overall Score	0.7529	0.752959
2.	1-gram Score	0.321429	0.321429
3.	2-gram Score	-	0.566947
4.	3-gram Score	-	0.687603
5.	4-gram Score	-	0.752959

6. CONCLUSION

The agenda was to create a translation model using RNN for human language. Although the model is not exactly trained to understand human languages or the meaning of any of the words in any language; it's rather an advanced approximation function that returns the nearest similar value based on the frequency of the most used words for the training data set for both the languages.

One of the conclusions drawn focuses on the training of the model. It states, even though the model is fairly accurate, the training of the model can be improved with a better corpus. As future work, we propose using the NMT-GRU method for a one-to-many model for the Indian regional languages. This eliminates the need for a separately trained model for every pair of parallel languages. However, the model will require a multilingual parallel corpus which is a tedious task for Indian regional languages.

REFERENCES

[1] http://www.mt-archive.info/

[2] Hutchins, W.J., 1995. “Machine translation: A brief history. In Concise history of the language sciences (pp. 431-445). Pergamon.

[3] Dzmitry Bahdanau, KyungHyun Cho, Yoshua Bengio, “Neural Machine Translation by jointly learning to align and translate”, ICLR, 2015, pp. 1-15.

[4] S. Saini and V. Sahula, "Neural Machine Translation for English to Hindi," Fourth International Conference on Information Retrieval and Knowledge Management (CAMP), Kota Kinabalu, 2018, pp. 1-6.

[5] K. Cho, B.Merrienboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation, CoRR, 2014, Vol. abs/1406.1078

[6] Sutskever, Ilya and Vinyals, Oriol and Le, Quoc V, Sequence to Sequence Learning with Neural Networks, Advances in Neural Information Processing Systems 27, 2014, pp. 31043112

[7] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, Yoshua Bengio, ”Empirical Eval- uation of Gated Recurrent Neural Networks on Sequence Modeling”, CoRR, 2014, vol. abs/1412.3555

[8] Karthik Revanuru, Kaushik Tarulpaty, Shrisha Rao, “Neural Machine Translation of Indian Languages”, Compute ’17, November 16–18, 2017, Bhopal, India.

[9] S. Hochreiter and J. Schmidhuber, ”Long short-term memory”, Neural computation, 1997, vol. 9, issue. 8, pp.1735-1780

[10] Cho, Kyunghyun, Van MerriA¨den boer, Bart, Gulcehre, Caglar, Bougares, Fethi, Schwenk, Holger, and Bengio, Yoshua, ”Learning phrase representations using rnn encoder-decoder for statistical machine translation”, CoRR, 2014, vol. abs/1406.1078

[11] Cho, Kyunghyun, Van MerriA¨den boer, Bart, Gulcehre, Caglar, Bougares, Fethi, Schwenk, Holger, and Bengio, Yoshua, ”Learning phrase representations using rnn encoder-decoder for statistical machine translation”, CoRR, 2014, vol. abs/1406.1078

[12] Vikrant Goyal, Dipti Misra Sharma, “The IIIT-H Gujarati-English Machine Translation system for WMT19”,

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers (Day 1) pages 191–195 Florence, Italy, August 1-2, 2019

[13] Yonghui Wu et al, "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation", CoRR, 2016, vol. abs/1609.08144

[14] http://preon.iiit.ac.in/~jerin/bhasha/

[15] Papineni, Kishore and Roukos, Salim and Ward, Todd and Zhu, Wei-Jing, "BLEU: A Method for Automatic Evaluation of Machine Translation", Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 2002, pp. 311- 318

[16] https://www.nltk.org/_modules/nltk/translate/bleu_score.html