2020 3rd International Conference on Intelligent Sustainable Systems (ICISS) | 978-1-7281-7089-3/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICISS49785.2020.9316016

Dual Translation of International and Indian Regional Language using Recent Machine Translation

N Jayanthi[1]
Computer Science and Engineering
Institute of Aeronautical Engineering
Hyderabad, India
jneelampalli.phd@gmail.com

Ch Suresh Kumar Raju[3]
Computer Science and Engineering
Institute of Aeronautical Engineering
Hyderabad, India
suresh.chittuluri@gmail.com

AluriLakshmi[2]
Information Technology
Institute of Aeronautical Engineering
Hyderabad, India
alurilakshmi@gmail.com

B Swathi[4]
Computer Science and Engineering
Institute of Aeronautical Engineering
Hyderabad, India
swathi.redy03@gmail.com

Abstract—**Neural Machine Translation (NMT) is a modern and powerful approach that has resulted in major improvements compared to traditional techniques of machine translation for translating one language into another.In the world, India is a very multicultural and multilingual country. People of India from various regions use their own regional communication languages, making India stand at the second position in the world to have maximum languages. In India, English is provided as the second extra official language. But the usage of English in India is very less forming a communication gap. To minimize this gap by translating one language into another language is almost impossible for humans. It can be achieved by a machine translation. This paper focuses on translating Indic languages using the translation technique of Neural technology. A Sequence to Sequence model with encoder-decoder attention mechanism of neural machine translation is proposed for Telugu language conversion into English and vice versa.**

Keywords—Machine Translation, Attention Mechanism, LSTM, Encoder, Decoder, Neural Machine Translation.

I. INTRODUCTION

In the midst of individuals, contact and toggling of information are mandatory for exchanging awareness, emotional state, ideas, thoughts and facts. In the world, India is a very multicultural and multilingual country. People of India from various regions use their own regional communication languages, making India stand at the second position in the world to have maximum languages. The two popular languages from the collection are Indo-Aryan languages which are used by Indian speakers[1]. Where 78.05% is used orally and Dravidian languages are used 19.64[2]%.In total, 121 languages and 270 mother tongues exist. At the time of the 2001 presentation, the number of scheduled languages was 22. In the 2011 Census, these 22 languages were retained and the languages of non-scheduled were extended from 99 to 100 [3].In India out of the total population, 96.71 per cent people have their mother tongue as one of the Scheduled Languages while 3.29 per cent people have languages to be different [3]. Apart from these languages with Article 343 of the Constitution of India, English is provided as the second extra official language[1]. Dissimilarities in English are used for human communication worldwide. English is heavily controlled by the content made available on the Internet. 20 % of the world's population communicates in English, while in India it is only 0.02[1]. This form a communication gap between regional languages of India and English. There is a need for operational and precise computational methods to connect this vast language gap, requiring minimal human interference. Using computer perception, this undertaking can be adequately achieved.

Machine Translation (MT) is defined as a technology that can translate human beings communicated standard language into another language with least human involvement. Machine interpretation plans to construct interpretations in the desired language that have a similar meaning as in the source phrase and try for syntactical accuracy. In mid-1950s early work on MT has started [4] which then progressed steadily till 1990s due to the incremental accessibility and knowledge preparation of computational limits. Numerous approaches to achieve more precise mama Chinese interpretation have been proposed at that point, such as "rule-based interpretation, knowledge-based interpretation, corpus-based interpretation, hybrid interpretation, and statistical machine translation (SMT)"[4]. The advantages and faults of all methodologies are their own. Among these, SMT, which is a sub-category of Corpus, is commonly used as it can yield better results compared to other methods already available. As of late around the world, the use of the Neural networks in machine interpretation is well known and the current machine interpretation technique uses a neural system which is known as "Neural Machine Translation or NMT". Various studies have been done on NMT as of late. There is little work done on Indian dialects [4]. The NMT approach to Indic dialects, especially bilingual machine interpretation, is observed as a failure.

In this paper, English-Telugu language pair used for applying our proposed model.A framework has been developed by utilizing the the attention mechanism of our neural model. Our proposed model is NMT based model. Section 2 present diagram of the proposed method followed by the next section that gives the basics of machine interpretation method of a neural system utilizing consideration instrument. Section 4 gives a relative examination of different programmed assessment lattices. Section 5 presents the dual translation method. The last but one 6thSectionexplain the usage and present results of the NMT model, the finish of the paper is presented in Section 7.

II. LITERATURE SURVEY

The method of converting or changing a native language into the desired language without the involvement of human is known as Machine Translation (MT). Converting words of native languages to another language without changing the meaning is said to be successful translation. In the mid-'50s, IBM research laboratory specialists brought about the start of work in the field of computer perception. In 1956, they also provided an important demonstration of the computer interpretation framework[4]. In any event, soon reported by the American government's programmed language planning warning board, the computer interpretation task is impossible to scale because of the asset calculation it needs. Another achievement in computer interpretation came just after 1979 when the clear interpretation system for the interpretation of climate bulletins from English to French [6] [7] was introduced. IIT Kanpur analysts developed the AnglaBharati-I Computer Perception Framework [8] [9] in 1991. With space customization, it was a widely useful interpretation system. It is specifically intended to render Hindi's interpretation of English. The CDAC created a machine interpretation system called MANTRA in 1999 [8], which make use of an exchange-based machine interpretation. The system is planned to take a shot at the English-Telugu, English-Bengali, English-Gujarati, English-Hindi, and information collections. Thereafter, the paradigm is moved to AnglaBharati-II[8][9] using a half-and-a-half-method of computer interpretation in 2004. The AnglaBharati-II works better thanAnglaBharati-I.

III. MACHINE TRANSLATION

The process of translating the native language into the chosen language, without losing the meaning of source language is said to be Machine Translation. By using an immense amount of multiple available corpuses bits of information is generated with techniques of machine translation. In 1981 Example based proposal had been made for Machine Translation. In 1990 It was completely developed[1]. The main purpose of this technique is to convert a native language into new [4].

A. Statistical Machine Translation

This type of models works on probability. In this technique, the source language word is referred to related word(s) in the chosen language. In any case, it needs a wide corpus of accurate translations in both languages.

B. Rule-based Machine Translation

There exist grammar measures for almost all languages that are used by people,it can be illustrated by setting criteria, building natural fluent phrases in the chosen language. In the Rule-based mechanism, by using crucial theory a model is built that is used for translating source language to target language. Enormous different rules with different contexts are required to decrypt Indian languages [5].

C. Machine Translation based on a phrase

A small word collection which has special meaning is known as a phrase. This technique has a phrase table with a translation between the source and the destination language. It has data on how the descriptions of various sentencesare recreated to construct meaningful language sentences. But these sorts of machine translation frameworks couldn't create natural or oral or written regular sentences in a language as it isn't possible to have all blend of various expression[5].

D. Machine Translation based on Neural Network

The recent techniques of the translation done by machines that work on neural-network-based are Neural Machine Translation. This technique uses conditional probability for converting native language to a specified desired language. Figure 1shows sequence to sequence model of NMT. NMT is a technique that requires less information about the pattern of native and target languages. It has overcome conventional MT models in enormous scope interpretation errands, for example, English French and German [10]. Even though there are many existing models for neural organizations the model with attention mechanism is recent of all.

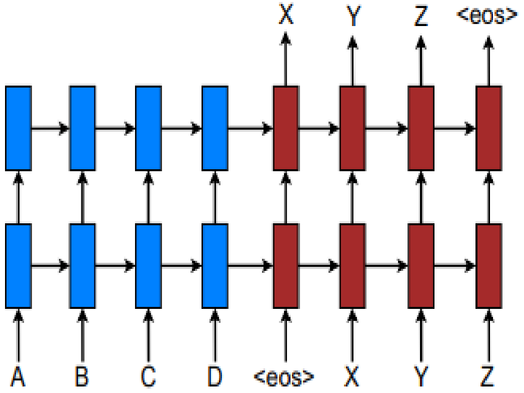


Fig. 1: sequence to sequence model for converting native language into the desired language using [13]

Intermittent models normally factor calculation along with the image places of the info and yield successions. Adjusting the situations according to calculation time, an arrangement of shrouded states ht, component of the past hidden state ht+ 1 and position t as input is obtained[11]. This nature of RNN does not allow parallelization to be applied inside preparing models. If the sequence is long constrains on memory would apply[12]. One of the significant disadvantages of models that takes a shot at arrangement to-succession model is that it can't create words that are once in a while experienced in input corpus. For taking care of this issue, consideration system can be used with conventional grouping to-arrangement model. It permits displaying of conditions regardless of their separation in the info or yield. The idea of "consideration" has picked up fame as of late in preparing of neural organizations, permitting models to study arrangements connecting various modes, e.g., among picture items and operator activities in the dynamic control issue [12].

IV. EVALUATIONMATRICES

This section list different assessment grids. In this paper, the accompanying assessment frameworks are advanced techniques which are planned to be used in the near future .

A. Error rate for Translation

Interpretation blunder rate or TER measures the measure of altering it needs to coordinate the human-produced yield. Itwas intended for assessing the yield of machine translation staying away from the information seriousness of importance based methodologies. This technique gives more significant bits of knowledge when there are countless reference sentences accessible in the dataset. Equation (1) can be used to discover TER of any interpreted sentences.

$$\text{TER} = \frac{number\ of\ edits}{Average\ number\ of\ reference\ word} \quad (1)$$

B. Perplexity Matrix

Perplexity is a proportion of language model execution dependent on normal likelihood. Perplexity can be characterized as the converse likelihood of the sentences accessible in test information, standardized by the number of words in produced sentences. It tends to be determined utilizing equation (2):

$$PP_T(PM) = \frac{1}{(\Pi^{t}_{i=0} PM(w_i \mid w_1 ... w_{i-1}))^{\frac{1}{2}}} \quad (2)$$

C. BLEU

BLEU utilizes the essential ideas of n-gram accuracy to ascertain similitude among the recommendation and sentence that is created. It corresponds profoundly with a human master audit as it utilizes the normal score of all outcome in test dataset instead of giving later effect of every sentence. BLEU score can be processed utilizing the accompanying condition (3):

$$p_n = \frac{\sum_{c \in \{candidates\}} \sum_{ngram \in c} Count_{clip}(ngram)}{\sum_{c' \in \{candidates\}} \sum_{ngram' \in c'} Count(ngram')} \quad (3)$$

V. IMPLEMENTATION

As appeared in Figure 2, it likewise gives setting which will get accommodating for producing more common looking sentences including uncommon words. As of late, attentional NMT models have overwhelmed the field of machine interpretation. Improvements are being made to increase the limit of interpretation execution by proceeding with fresh improvement in NMT structures.

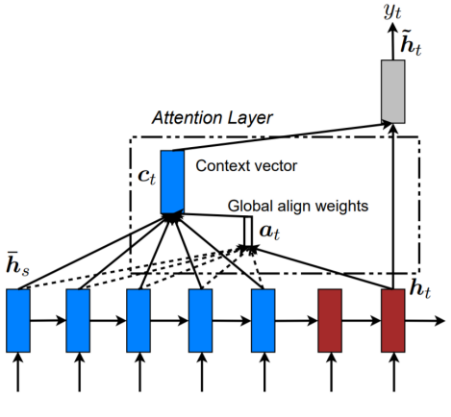


Fig. 2:AttentionMechanism [13]

Figure 3 shows our proposed model. It contains three parts decoder, Encoder and Attention component. The encoder has layers of LSTM formed by using 128 LSTM cells. The encoder will generate a vector which is in encoded form. This vector is given as input to the decoder. The decoder on the other hand is also built up using layers of LSTM which are also formed with 128 LSTM cells. The decoder takes input from the encoder and generates decoded form data which can be used. At whatever point any yield has created the estimation of the shrouded state is contrasted with all information states with infer loads for consideration instrument. In light of consideration loads, setting vector is determined and it is given as an extra contribution to the decoder for creating setting pertinent interpretation dependent on past results.

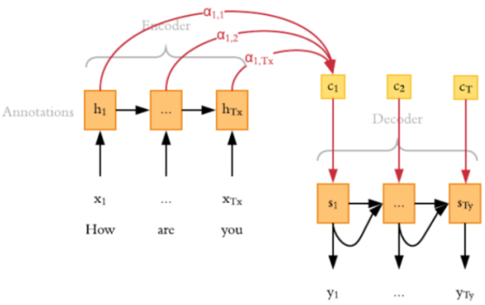


Fig. 3: Dual Conversion System

Datasets

A large amount of information is required for working with neural networks. When neural networks learn to learn, with training, learning is more reliable. For non-Indian languages,a lot of research has been done. Dataset is taken from http://www.manythings.org/anki/

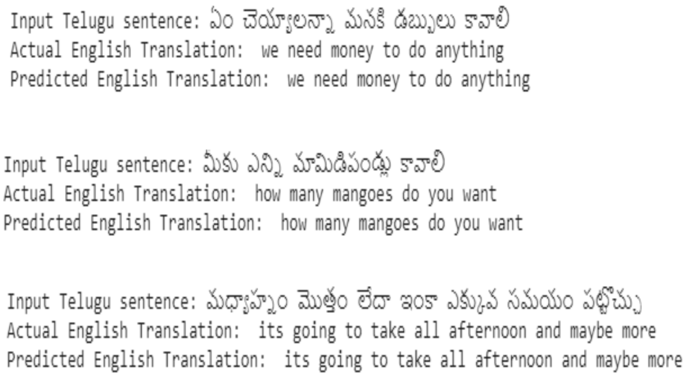
Experiment Setup

The 15he 15GB of primary memory, Google Cloud and 12GB of GPU memory of Tesla K80(2496 CUDA Core) GPU are used. For training deep neural networks and for creating deep neural networks Tensorflowlibrary is used[19].

VI OUTPUT AND DISCUSSION

Output

Our proposed model which is built using neural machine technique was trained using teleng parallel corpus. it was trained with an epoch count of 50. Figure 4 and 5 presents the proposed model results and the accuracy is presented in Figure 6 :

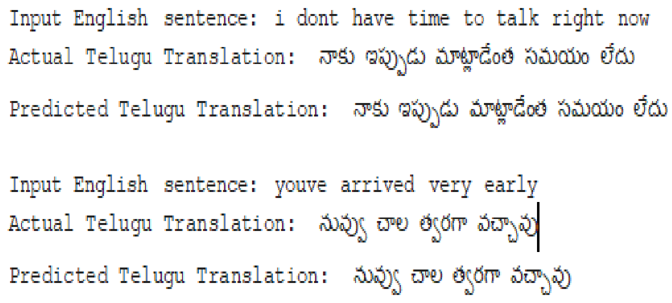


```
Input Telugu sentence: ఏం చెయ్యాలన్నా మనకి డబ్బులు కావాలి
Actual English Translation:  we need money to do anything
Predicted English Translation:  we need money to do anything


Input Telugu sentence: మీకు ఎన్ని మామిడిపండ్లు కావాలి
Actual English Translation:  how many mangoes do you want
Predicted English Translation:  how many mangoes do you want


Input Telugu sentence: మధ్యాహ్నం మొత్తం లేదా ఇంకా ఎక్కువ సమయం పట్టొచ్చు
Actual English Translation:  its going to take all afternoon and maybe more
Predicted English Translation:  its going to take all afternoon and maybe more
```

Fig. 4: Conversion of Telugu to English sentences



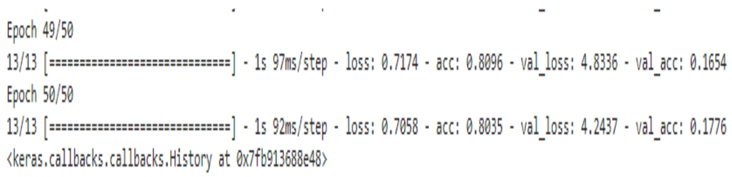
```
Input English sentence: i dont have time to talk right now
Actual Telugu Translation:  నాకు ఇప్పుడు మాట్లాడేంత సమయం లేదు

Predicted Telugu Translation:  నాకు ఇప్పుడు మాట్లాడేంత సమయం లేదు


Input English sentence: youve arrived very early
Actual Telugu Translation:  నువ్వు చాల త్వరగా వచ్చావు

Predicted Telugu Translation:  నువ్వు చాల త్వరగా వచ్చావు
```

Fig.5: Conversion of English to Telugu sentences



```
Epoch 49/50
13/13 [==============================] - 1s 97ms/step - loss: 0.7174 - acc: 0.8096 - val_loss: 4.8336 - val_acc: 0.1654
Epoch 50/50
13/13 [==============================] - 1s 92ms/step - loss: 0.7058 - acc: 0.8035 - val_loss: 4.2437 - val_acc: 0.1776
<keras.callbacks.callbacks.History at 0x7fb913688e48>
```

Fig. 6: Accuracy of the model

VII CONCLUSION

Conventional machine translation methods are quick and effective enough for processing. They have been shown to be important in providing great results with their narrow capability of operation. But they face difficulties in producing corpus or the desired language which is fluent and possible with human inclusion. The difficulties of traditional approaches of machine translations are overcome by Neural machine translation. The recent NMT model like Seq-2-Seq has shown great results in producing the desired language. Although, in some real-time circumstances, especially when the model has to go through rare words, the model reduces its accuracy drastically. This limitation is overcome by the recent NMT models, that generates a context vector by using an

Attention mechanism. Even though the model faces unknown words its accuracy will not be decreased.

REFERENCES

[1] Registrar General & Census Commissioner, “Abstract of speakersstrength of languages and mothertongues 2011,”2011.
[2] https://en.wikipedia.org/wiki/Languages_of_India

[3] https://indianexpress.com/article/india/more-than-19500-mother-tongues-spoken-in-india-census-5241056/

[4] P. Sheridan, “Research in language translation on the ibm type 701,”IBM Technical Newsletter, vol. 9, pp. 5–24, 1955.
[5] P. Shah, V. Bakarola, and S. Pati, “Neural machine translation system forindic languages using deep neural architecture,” in Smart and InnovativeTrends in Next Generation Computing Technologies, P. Bhattacharyya,
H. G. Sastry, V. Marriboyina, and R. Sharma, Eds. Singapore: SpringerSingapore, 2018, pp. 788–795.
[6] V. Lawson, Ed., Practical experience of machine translation. North-Holland Publishing Company, 1982.
[7] J. Durand, P. Bennett, V. Allegranza, F. van Eynde, L. Humphreys,P. Schmidt, and E. Steiner, “The eurotra linguistic specifications: Anoverview,” Machine Translation, vol. 6, no. 2, pp. 103–147, 1991.
[Online]. Available: http://dx.doi.org/10.1007/BF00417680
[8] S. B. Sitender, “Survey of indian machinetranslation systems,” IJCST ,vol. 3, no. 1, 2012.
[9] S. K. Dwivedi and P. P. Sukhadeve, “Machine translation system inindian perspectives,” Journal of computer science, vol. 6, no. 10, p.1111, 2010.
[10] Y. Wu, et.al., “Google’s neural machine translation system: Bridging the gapbetween human and machine translation,” CoRR, vol. abs/1609.08144,2016.
[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez,Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in Advancesin Neural Information Processing Systems, 2017, pp. 5998–6008.
[12] M. Luong, H. Pham, and C. D. Manning, “Effective approaches toattention-based neural machine translation,” CoRR, vol. abs/1508.04025,2015. [Online]. Available: http://arxiv.org/abs/1508.04025.
[13] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequencelearning with neural networks,” in Advances in Neural InformationProcessing Systems 27, Z. Ghahramani, M. Welling, C. Cortes, N. D.Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc.,2014, pp. 3104–3112. [Online]. Available: http://papers.nips.cc/paper/
5346-sequence-to-sequence-learning-with-neural-networks.pdf