RESEARCH ARTICLE OPEN ACCESS

Neural Machine Translation System of Dravidian Languages

T. Madhavi Kumari
Associate Professor, Dept. of ECE, JNTUH CEH

Dr. A. Vinaya Babu
Professor, Dept. of CSE, JNTUH CEH

ABSTRACT

Neural machine translation (NMT) is a recent and effective technique which led to remarkable improvements in comparison of conventional machine translation techniques. The ability of deep neural networks to learn a sensible representation of words is one of the major reasons for this improvement. Proposed NMT model developed for the Telegu language contains encoder-decoder with attention mechanism. In India, almost all the languages are originated from their ancestral language - Sanskrit. They are having inevitable similarities including lexical and named entity similarity. Translating into Indic languages is always being a challenging task. In this work, neural machine translation system (NMT) is presented for efficiently translate Indic languages like Telegu and Hindi that together covers more than 58.49 percentage of total speakers in the country. The performance of NMT model is compared in terms of automatic evaluation matrices such as BLEU, perplexity and TER matrix. The comparison of our network with Google translate is also presented where it outperformed with a margin of 6 BLEU score on English-Gujarati translation.

Keywords:- Deep Neural Network, Machine Translation, NLP, NMT, Telgu-English

I. INTRODUCTION

India is a highly diverse multilingual country in the world. In India, people of different regions use their own regional speaking language, which makes India a country having world's second highest number of languages. Human spoken languages in India belong to several language families. Two main of those families are typically known as Indo- Aryan languages having 78.05 percentage Indian speakers [1] and Dravidian languages having 19.64 percentage Indian speakers. Communication and information exchange among people is necessary for sharing knowledge, feelings, opinions, facts, and thoughts. Variation of English is used globally for human communication. The content available on the Internet is exceptionally dominated by English. Only 20 percent of the world population speaks in English, while in India it is only 0.02 [2]. It is not possible to have a human translator in the country having this much language diversity. In order to bridge this vast language gap we need effective and accurate computational approaches, which require minimum human intervention. This task can be effectively done using machine translation [3].

Machine Translation (MT) is described as a task of computationally translate human spoken or natural language text or speech from one language to another with minimum human intervention. Machine translation aims to generate translations which have the same meaning as a source sentence and grammatically correct in the target language. Initial work on MT started in early 1950s [4], and has advanced rapidly since the 1990s due to the availability of more computational capacity and training data. Then after, number of approaches has been proposed to achieve more and more accurate machine translation as, Rule-based translation, Knowledge-based translation, Corpus-based translation, Hybrid translation, and Statistical machine translation (SMT). All the approaches have their own merits and demerits. Among these, SMT which is a subcategory of Corpus based translation, is widely used as it is able to produce better results compared to other previously available techniques [5].

The usage of the neural networks in machine translation become popular in recent years around the globe and the novel technique of machine translation with the usage of neural network is known as Neural Machine Translation or NMT. In recent years, many works has been carried out on NMT. Little has been done on Indian languages as well [6]. We found the NMT approach on Indic languages is still a challenging task, especially on bilingual machine translation. NMT is a recently formulated method for automatic translation with the help of deep neural networks. NMT has already shown promising results in translations of several language pairs [7]. Unlike SMT, which requires separately trained sub-components for translation, NMT uses a single large neural network for training. This structure comprised of encoder and decoder networks where the encoder consumes the input sentences to produce a vector representation, and the decoder takes this vector and outputs the target language words [8]. In this work, we have developed a system that uses neural model based on Attention mechanism. Our proposed attention based NMT model is tested with evaluation matrices as BLEU, perplexity and TER.

The structure of the paper is as follows: in the next section, related work in the field of machine translation in the English–Indian languages is discussed. The methodology behind NMT and Dataset preparation is described in Section 3. The details about the experiments and the analysis of the results are discussed in Section 4, and finally the paper concludes in Section 5.

II. RELATED WORKS

Hindi, Malayalam, Tamil, and Telegu are among the most widely used languages in India [9]. Also, Hindi and English are the official languages in India. Telegu is the official state language for Telangana and Andhra Pradesh states. Most of the government documents are available in English and Telegu, and they should be translated into other widely spoken language like English to reach the common people. Thus, translation from Telegu into English language has very much importance in our country. Much research works have been done on machine translation from Indian languages to English, mostly focusing on ruled-based methods due to the unavailability of good parallel corpora. In spite of the unavailability of sufficient parallel corpora, significant works were done using statistical as well as hybrid approaches to translate text from Indian languages to English. In this section, we will discuss the notable works done in machine translation [10].

Anglabharti [11] is one of the oldest machine translation systems in India. It was a multilingual translation system that translates English sentences into Indian languages. This work followed a pseudo-target approach in which the source sentence was converted into an intermediate representation, and the target sentences were generated from this intermediate representation using a text generator. Angla-Hindi [12] is an extension to Angla-Bharti. In addition to the techniques used in Ref. [13],Angla-Hindi used an example-based approach to translate frequently occurring noun and verb phrases. Anusaaraka [14, 15] was another initiative put forward by the Indian Institute of Technology, Kanpur, which used the principles of Paninian Grammar for translation. Anusaaraka had two modules: the first module does the language-based analysis of the text, whereas the second module performs the statistical analysis.

MANTRA-Rajyasabha [16] developed by CDAC-Pune, India is a machine-aided translation tool primarily designed to translate government documents from English–Indian languages and vice versa. In this approach, both source and target language grammars were represented using the lexicalized tree-adjoining grammar (LTAG). CDAC-Mumbai, India developed a machine translation system called MaTra [17], which relied on a transfer-based approach for translation. Dave et al. [18] proposed a machine translation approach based on interlingua. The method was designed in such a way that information extracted from sentences in source language are converted into a universal networking language (UNL), and target language sentences are generated from this UNL representation.

IBM India Research Lab, New Delhi, developed an SMT system [19] that used IBM models 1, 2, and 3. Other notable works in English–Hindi translation were done by Jadavpur University, IIT Delhi [20], and most of them followed the example-based statistical approach. Some rule-based approaches toward the English–Malayalam automatic translations were proposed by Refs. [21] and [22] and Sunil et al. [23]. Apart from rule-based approaches, various studies were done on SMT in the English–Malayalam translation. In Ref. [24], Rahul et al. proposed a rule-based approach to modify the SMT output. The source language syntax was changed according to the syntax of the target language using some transformation rules and also using a morph analyzer for separating the suffixes from the root words in both languages. Anand Kumar et al. [25] proposed a factored approach for the English–Tamil translation, which used POS-tagged information and morphological information extracted from both languages to assist in the translation. Sridhar et al. [26] proposed a methodology, which used universal networking language as an intermediate representation in the English–Tamil translation.

III. METHODOLOGY

A. Data Preparation- Parallel Corpus

Machine translation can be stated as the process of translating source language into target language considering the grammatical structure of the source language. This approach of machine translation was based on generating insights from large amount of available parallel corpuses. Example based Machine Translation was first proposed in 1981, but was developed from about 1990 onwards [27]. The core idea is to reuse existing translations for generating a new translation. It is important to collect sentences from various domains while preparing a parallel corpus so that frequently used words in all those domains can be added to the vocabulary. This will further reduce the possibility of occurring out of vocabulary words when the system will be tested. On account of this reason, parallel corpora were prepared by collecting sentences from sources such as websites where bilingual texts are available (e.g. vikaspedia.in), story books, new websites, film subtitles, Bhagavat Gita Bible, Quran, and freely available encyclopedias. The coverage of these sentences spread across various fields like film, sports, politics, short stories, agriculture, religion, health, education, and language. Sentences that were collected from the above-mentioned sources, especially online resources, contained a lot of impurities. The objective of cleaning was to remove foreign characters and incomplete sentences. The sentences that did not convey the meaning properly was omitted and foreign characters were removed using regular expression. In addition to online resources, parallel sentences were also collected from bilingual books [28].

Performance of the four NMT systems was tested using a data set containing 562 English sentences. This data set comprises sentences from different domains such as health, tourism, and entertainment. This test data set was prepared separately, and it did not include any sentence from the training data set.

B. Neural Machine Translation System

The fundamental idea behind an NMT system is to predict a sequence of words $A = (a_1, . . . , a_t)$ in the target language for a given source language sentence $B = (b_1, . . . , b_s)$. This conditional probability distribution is modeled using the RNN-based encoder–decoder architecture. The encoder takes the variable length source language sentence and converts it into a fixed length vector. This fixed vector length vector (sentence embedding) contains the meaning of the input sentence. The decoder then takes this sentence embedding as the input and starts predicting the output words by taking the context of each word into consideration. Mathematically, this thought can be represented as,

$$\text{Log R (A/B)} = \sum_{k=1}^{t} \log R\left(\frac{y_k}{y_{k-1,\ldots\ldots}y_1}, b, c\right)$$

Where *R (A/B)* is the probability of obtaining a target language word *y* for a given source language word *b*, and *c* is the context of that particular word.

i. Encoder

Basically, an encoder transforms a sentence into a vector form, which represents the meaning of that sentence. Initially, word representations for both source and target language words are obtained. This word embedding is then fed into the encoder–decoder network. The encoder network transforms these word representations into a sentence embedding. This task is performed using a bi-directional recurrent neural network which contains two RNNs for computing rightward and leftward hidden state sequences. This allows us to capture both rightward and leftward contexts of each word.

$$\rightarrow h_i = f \rightarrow h_{i-1}, W_e^{xi}) \qquad (2)$$
$$\leftarrow hi = f\,(\rightarrow hi{+}1, W_e^{xi}) \qquad (3)$$

Where $\rightarrow h_i$ and $\leftarrow h_i$ represents the rightward and leftward hidden state sequences, W_e is the word embedding matrix, *xi* is the input (source) word, and *f* is a nonlinear function.

Now, each source language word, *xi* can be represented using both rightward and leftward hidden state information, i.e. $h_i = (\rightarrow h_i, \leftarrow h_i)$. This helps us to obtain more information about each word by incorporating the details of the surrounding words. This information is then fed into the input layer of the decoder. For example, while translating an English sentence "you always work", it is first represented using a word-embedding mechanism and then fed into the encoder as input. The encoder is built with the LSTM/Bi- RNN networks and a zero vector as the starting state. The last hidden layer of the encoder network generates a vector, which carries the meaning of the input sentence. NMT system (encoder–decoder architecture) is given in Figure 1.

ii. Decoder

The decoder is responsible for predicting target language words by taking the sentence embedding obtained at the encoder, previously predicted target words, and the context of each word. From equation (1),

$P\ (b_k|b_{k-1}, \ldots b_1, x, c)$ is computed as,

$$P\ (b_k|b_{k-1}, \ldots b_1, x, c) = g\ (b_{k-1}, f\ (s_{k-1}, b_{k-1}, c_k), c_k) \qquad (4)$$

Where g is a nonlinear function, b_{k-1} is the previously predicted target word, c_k is the context of each word, and s_{k-1} is the decoder hidden state at time $k - 1$.

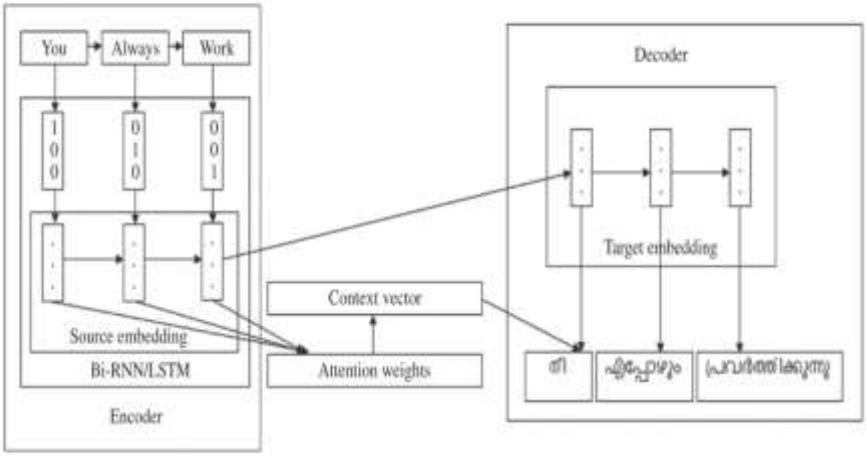

Fig. 1. Architecture of NMT system

In the decoder, the network is initialized to the last layer of the encoder network because the decoder requires access to the input sentence. Therefore, the hidden layer at the source word "work" is shared with the decoder network, which then acts as the initial state to the decoder. Now, the decoder starts generating target words based on the hidden state information obtained from the encoder. This hidden state vector acts as the context vector, which embeds the knowledge of the source sentence. This vector may not be able to capture all the contexts associated with the source language sentence when the sentence is long. Attention mechanism is used to tackle such problems. In attention mechanism, attention weights are generated by comparing each target hidden state with the source hidden states. Based on the attention vector, a context vector is computed. Then, the attention vector is derived by combining the context vector and the attention weights. This attention vector is then fed as the input to the decoder.

IV. RESULTS AND DISCUSSIONS

The models were trained with a corpus described in this paper and bilingual sentences obtained from the TDIL database for training in the NMT system. Before feeding into the network, sentences were tokenized using the tokenizer module given in the Moses tool kit. The system was trained with different parameters such as the number of hidden layer units (250 and 550) and the neural network algorithm (LSTM and Bi-RNN) of which the set of parameters.

Evaluation of machine translation output is extremely important as it is the key factor that determines the quality of the translation, required level of post-editing, etc. Generally, MT outputs are evaluated both automatically and manually. One of the most common automatic evaluation metric is BLEU. In this paper, we used the BLEU score to evaluate the translation. The BLEU score was computed for each of the eight architectures and are listed in Table 1. The LSTM and Bi-RNN algorithms were used to design the encoder network and trained the model with 250 and 550 hidden layer units in each of the hidden layers. It is evident from Table 1 that the Bi-RNN algorithm with 550 hidden layer units produces good translations compared to other models. Therefore, we can infer that, an increase in the hidden layer units will improve the translation. Similarly, the use of the Bi-RNN algorithm along with attention mechanism can capture the dependencies in the text well compared to the LSTM with attention mechanism.

Table 1: BLEU Scores of Translation

Language pair	Algorithm	No. of hidden layers	BLEU score
English to Telugu	LSTM	250	21.59
		550	25.69
	Bi-RNN	250	22.34
		550	26.84

V. CONCLUSION

Machine translation from English to Indian languages is always a difficult task due to the unavailability of a good quality corpus and morphological richness in the Indian languages. For an NMT system to produce better translations, the size of the corpus should be huge. In addition to that, the parallel sentences should convey similar meanings, and the sentences should cover different domains. Modeling the system with such a corpus can assure good translations while testing the model. This corpus is our contribution to the machine translation research community. Apart from the size and coverage of the corpus, the length of the sentences also plays a significant role in determining the quality of translation. The length should not be too short or too long because deep learning architectures cannot extract very long dependencies present in the sentences. As morphological richness in the English and Indian languages are in two extremes of the spectrum, adding linguistic features along with the sentences can improve the translation.

REFERENCES

[1] Shah P, Bakrola V. Neural Machine Translation System of Indic Languages-An Attention based Approach. In2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP) 2019 Feb 25 (pp. 1-5). IEEE.

[2] Premjith B, Kumar MA, Soman KP. Neural Machine Translation System for English to Indian Language Translation Using MTIL Parallel Corpus. Journal of Intelligent Systems. 2019 Jul 26;28(3):387-98.

[3] Suryakanthi T, Sharma K. Discourse translation from English to Telugu. InProceedings of the third international symposium on women in computing and informatics 2015 Aug 10 (pp. 222-227).

[4] Reddy MV, Hanumanthappa M. Indic language machine translation tool: English to kannada/telugu. InMultimedia Processing, Communication and Computing Applications 2013 (pp. 35-49). Springer, New Delhi.

[5] Lingam K, Lakshmi ER, Theja LR. Rule-based machine translation from English to Telugu with emphasis on prepositions. In2014 First International Conference on Networks & Soft Computing (ICNSC2014) 2014 Aug 19 (pp. 183-187). IEEE.

[6] Vaswani A, Zhao Y, Fossum V, Chiang D. Decoding with large-scale neural language models improves translation. InProceedings of the 2013 conference on empirical methods in natural language processing 2013 Oct (pp. 1387-1392).

[7] Koehn P, Knowles R. Six challenges for neural machine translation. arXiv preprint arXiv:1706.03872. 2017 Jun 12.

[8] Abiyev RH, Arslan M, Idoko JB. Sign Language Translation Using Deep Convolutional Neural Networks. KSII Transactions on Internet & Information Systems. 2020 Feb 1;14(2).

[9] Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. arXiv preprint arXiv:1409.3215. 2014 Sep 10.

[10] Rajbhasha.nic.in Oflcial languages in India, Available from:http://rajbhasha.nic.in/en/ constitutional -provisions.Accessed 10 June, 2017.

[11] Sinha RM, Sivaraman K, Agrawal A, Jain R, Srivastava R, Jain A. ANGLABHARTI: a multilingual machine aided translation project on translation from English to Indian languages. In1995 IEEE International Conference on Systems, Man and Cybernetics. Intelligent Systems for the 21st Century 1995 Oct 22 (Vol. 2, pp. 1609-1614). IEEE.

[12] Sinha RM, Jain A. AnglaHindi: an English to Hindi machine-aided translation system. MT Summit IX, New Orleans, USA. 2003 Sep 23;494:497.

[13] A. Bharati, V. Chaitanya, A. P. Kulkarni and R. Sangal, Anusaaraka: machine translation in Stages,arXiv preprintcs/0306130(2003).

[14] S. Chaudhury, A. Rao and D. M Sharma, Anusaaraka: an expert system based machine translation system, in:NaturalLanguage Processing and Knowledge Engineering (NLP-KE), 2010 International Conference on, pp. 1–6, IEEE, Beijing,2010.

[15] CDAC-MANTRA, Available from: https://www.cdacindia.com/html/aai/mantra.asp. Accessed 31 May, 2017

[16] CST-MATTRA, http://www.ncst.ernet.in/matra/. Accessed 31 May, 2017.

[17] S. Dave, J. Parikh and P. Bhattacharyya, Interlingua-based English–Hindi machine translation and language divergence,Mach. Transl.16(2001), 251–304.

[18] R. Udupa and T. A. Faruquie, An English-Hindi statistical machine translation system, in:International Conference onNatural Language Processing, pp. 254–262, Springer, Hainan Island, China, 2004.

[19] S. B. Sitender, Survey of Indian machine translation systems,IJCST3(2012), 47.

[20] R. Harshawardhan, Rule based machine translation system for English to Malayalam language,Diss. de mestrado.Coimbatore: Amrita School of Engineering (2011).

[21] R. Rajan, R. Sivan, R. Ravindran and K. P. Soman, Rule based machine translation from English to Malayalam, in:Advancesin Computing, Control, and Telecommunication Technologies, 2009. ACT'09. International Conference on, pp. 439–441,IEEE, Trivandrum, India, 2009.

[22] R. Sunil, N. Manohar, V. Jayan and K. G. Sulochana, Development of Malayalam text generator for translation from English,in:India Conference (INDICON), 2011 Annual IEEE, pp. 1–6, IEEE, Hyderabad, 2011.

[23] C. Rahul, K. Dinunath, R. Ravindran and K. P. Soman, Rule based reordering and morphological processing for English-Malayalam statistical machine translation, in:Advances in Computing, Control, and Telecommunication Technologies,2009. ACT'09. International Conference on, pp. 458–460, IEEE, Trivandrum, India, 2009.

[24] M. Anand Kumar, V. Dhanalakshmi, K. P. Soman and S. Rajendran, Factored statistical machine translation system forEnglish to Tamil language,Pertanika J. Soc. Sci. Hum.22(2014), 1045–1061.

[25] R. Sridhar, P. Sethuraman and K. Krishnakumar, English to Tamil machine translation system using universal networkinglanguage,S¯adhan¯a41(2016), 607–620.

[26] Reddy Pc, Babu As. A Novel Approach To Analysis District Level Long Scale Seasonal Forecasting Of Monsoon Rainfall In Andhra Pradesh And Telangana. International Journal Of Advanced Research In Computer Science. 2017 Nov 1;8(9).

[27] Reddy PC, Babu AS. Usage of co-event pattern mining with optimal fuzzy rule-based classifier for effective web page retrieval. International Journal of Engineering & Technology. 2018;7(3.29):275-9.

[28] Sucharitha Y, Vijayalata Y, Prasad VK. Analysis of Early Detection of Emerging Patterns from Social Media Networks: A Data Mining Techniques Perspective. InSoft Computing and Signal Processing 2019 (pp. 15-25). Springer, Singapore.

[29] Shaker Reddy PC, Sureshbabu A. An Enhanced Multiple Linear Regression Model for Seasonal Rainfall Prediction. International Journal of Sensors Wireless Communications and Control. 2020 Aug 1;10(4):473-83.