International Conference on Recent Trends in Data Science and its Applications
DOI: rp-9788770040723.087

Machine Translation from German to English and English to German

NarayanaDarapaneni
Director – PGDCSAIGreat Learning/Northwestern University, Illinois,
USA
darapaneni@gmail.com

ReibhuSant
Student - PGDCSAI
Great Learning,
Bangalore, India
reibhusant1@gmail.com

HemaAllarapu
Student - PGDCSAI
Great Learning,
Bangalore, India
allarapuhema@gmail.com

Anwesh Reddy Paduri
Senior Data Scientist
Great Learning,
Hyderabad, India
anwesh@greatlearning.in

PunamYadav
Student - PGDCSAI
Great Learning,
Bangalore, India
py.dpsi@gmail.com

AnkushPatil
Student - PGDCSAI
Great Learning,
Bangalore, India
ptlankush4u@gmail.com

Abstract—Language is the foremost method of human communication which consists of grammar and vocabulary. Language majorly consists of words which are structured in a meaningful way. The natural human language is expressed through speech or writing. A computer is a binary machine i.e., it understands codes which are strings of 0s and 1s aka binary digits (“bits”). Binary Language is also known as Machine Language. A computer does not understand the natural human language. In an endeavour to bring the computers and the humans closer, by allowing a computer to analyse the sentence spoken by a user aka input speech recognition and process what does it mean, Natural Language Processing (NLP) was devised. NLP combines the fields of Artificial Intelligence (AI) and analyticalsyntax so that humans and computers can communicate with each other effortlessly. Today, mobile devices have applications such as Alexa, Siri, among others, that make it easy to do tasks like add activities to a calendar or search for a contact using voice command. All these applications use in some form or another natural language processing and machine learning.

Keywords—LSTM, RNN, Bidirectional, Embedding, English, German, Machine Translation

I. INTRODUCTION

Language is the foremost method of human communication, which incorporates both grammar and vocabulary. On the other hand, a computer understands Machine Language, which are numeric codes in the form of algorithms. Machine Language allows a computer to directly execute operations. These codes or algorithms are strings/combinations of 0s and 1s aka bits. A computer cannot understand the human language in its natural form.

To overcome the gap between the machines and humans, Natural Language Processing (NLP) was devised. It enables a computer to process the input speech recognition and analyse what the user stated. NLP combines the fields of computational linguistics and Artificial Intelligence (AI) so that both computers and humans can communicate effectively.Today, applications such as Alexa, Siri, among others, in the smartphones have made it easy to do tasks like add activities to a calendar or search for a contact by voice.

To be able to talk with mankind, a computer program should understand grammatical syntax, word meanings or semantics, correct use of tenses aka morphology and conversational pragmatics. Numerous regulations which are to be taken care of are staggering and explains the reasons of failure of previous attempts at NLP. Instead of taking a rule-based approach, of lately NLP has changed its method to a pattern learning based computer programming. Here we have attempted to develop a Machine Translation model that can translate sentences from the given source material into other language, without human intervention. We have taken sentences of German language as database which is to be converted into English language and vice-versa. The idea behind this project is to allow the two parties to communicate and exchange ideas from different countries.

II. LITERATURE REVIEW

For translation of text or speech among different set of languages, Machine Translation (MT) was devised. It uses software which substitutes word in one language with an equivalent word in the target language. Results produced with this approach were not satisfactory, as at times an appropriate word match might not be available in the target language. Also, a single word, depending upon the context, may have multiple meanings.

To overcome this issue, modern machine translation took a holistic approach. Here an attempt has been made to check whether there is meaningful translation of the entire sentence from one language to another rather than simple mechanical substitution of words. It analyses the influence of all the text elements and words on one another.

Machine Translation models are found to give the best results in domain or occupation specific test samples, as it reduces the scope of allowable substitutions.

A. *Machine translation origin*

The origins of machine translation can be found in the works of Al-Kindi, a ninth-century Arabic cryptographer who developed methodology for systemic language translation, which included cryptanalysis, frequency analysis, probabilityand statistics, all used in modern machine translation.

On the 7th January 1954 the IBM-Georgetown experiment was conducted in which an IBM machine translated 60 sentences from Russian to English for the first time in the history. To advancing research in the field of machine translation, researchers continued to get together by making forums like Association for Machine Translation

DOI: rp-9788770040723.087

and Computational Linguistics was formed in 1962. The French Textile Institute used Machine Translators to translate abstractsvarious languages like French, English, German and Spanish in the 1970s. Brigham Young University began a project to translate Mormon texts using automated translation in 1971.

In the research done by M.D.Okpor[11] it is evident that more innovations were included as the time passed. In 2012, Google made a pathbreakingannouncement that Google Translatetranslates roughly enough words to fill 1 million books in one day.

B. Machine translation methods and techniques

i) Statistical machine translation

Statistical machine translation or SMT is a machine translationmodelwhere language translation is done using the statistical probability of the words, which are extracted from the analysis of multi-language texts. Warren Weaver in 1949, first thought of statistical translation models,including the idea of applying Claude Shannon's information theory to the model. Statistical machine translation produces translations using statistical methodsas probability, mean, mode, variance, standard deviation et ceterabased on multi-language text collections. When such compilations are available, good results can be achieved translating similar texts.However, such collections arerare for many language pairs. Though, newer approaches into Statistical Machine translation such as METIS II and PRESEMT use minimal collection sizes. Theyfocus on derivations of semantic structure with the help of pattern recognition that prediction of words. With some additional development, these techniques allow for SMT models to operate off a monolingual text compilation. SMT's biggest flawis thatit depends on large amounts of parallel texts to learn. It faces problems with linguistic-rich languages and is oftenunable to correct errors with single words.

C. Neural machine translation

Deep learning-based approach to MT, neural machine translation has made great progress in machine translations overtaking the SMT as the primary commercial way of machine translation. In this model, the word sequence modelling is first accomplished using a recurrent neural network (RNN). A bidirectional recurrent neural network, also known as an encoder, is used by the neural network to encode a source sentence whereas for second RNNa decoder,is used to predict words in the target language. RNNs, however, face difficulties in encoding long inputs into a single vector as shown by D.Bahdanau [14]. Itwas compensated by a mechanismthat allowed the decoder to focus on different parts of the input while generating wordsfor the output. NMT has become a widely used technique for machine translation, and other related NLP tasks such as dialogue, parsing, and summarization.To address the multi-word expression translations like idiomatic phrases, and low-frequency words, language-focused linguistic features have been used in state-of-the-art neural machine translation (NMT) models as mentioned by F.Meng[12]. Through his research C. Kobus[7] has showed that the neural machine translation (NMT) as a methodology for machine translation has led to astounding improvements in machine translation, particularly in terms of human evaluation, compared to statistical machine translation (SMT) systems.

III. Materials and Methods

A. Sample Data

We have data from News commentary Ion the same topic in German and English. We will also use Europarl V7 and Common crawl in German and English. We will use this data to train two models:

1) English to German
2) German to English

These two models will be incapsulated in a layer where the user can select which model to use. We would use different NMT methods such as RNN, LSTM, Bi-directional RNN, Bi-directional LSTM, Bi-directional RNN with embedding and Bi-directional LSTM with embedding. This is done so as to evaluate the performance of various Neural Network algorithms to find the algorithm best suited for the task. We will define the Neural networks in functions so we can initialize two objects. One object shall handle English to German translation and the other will handle German to English translation. Ouraim is to make an application that is an End-to-End translator which has the most optimum performance.

B. Exploratory Data Analysis

We have 3 sets of files in English and German each. Each data set contains about 2 million sentences.Sentences from each dataset are cut into 300 sentences each, to get a smaller training set which has 900 English sentences and 900 German sentences. This now is usedto train the neural networks. This is done as to limit the size of the dataset and ensure it can work on a machine with a limited computing capability. The different datasets are combined into respective English and German datasets to provide a consolidated dataset. We first read these files into lists to clean the data and perform tokenization. In general pre-processing, we make all the data in lower case, remove the spaces at front and back of the dataset and remove the punctuation marks. Further pre-processing is required to make sure that each index of the list contains a sentence. In the second pre-processing, we tokenize all the words of the sentence and get a tokenized sentence and a tokenizer. The tokenizer is a list of all the words of the language with their corresponding id number. This helps to give a numerical value to words and helps in putting the file in translator. We define two tokenizer an English Tokenizer and a German tokenizer.

DOI: rp-9788770040723.087

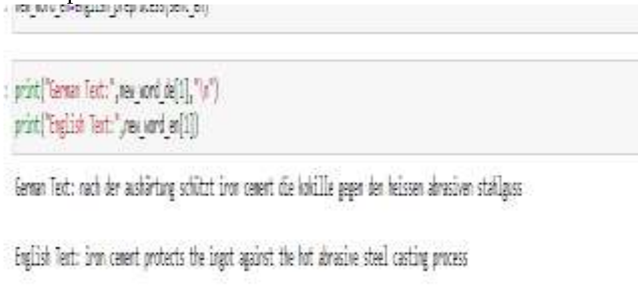

Fig. 1. Data after Pre-processing

C. Base Model and Architecture.

We have used RNN models, LSTM models, Bidirectional RNN models, Bidirectional LSTM models, Bidirectional RNN Models with Embeddings and Bidirectional LSTM Models with Embeddings. We made the models as functions that can be used to define separate models for both English to German and German to English translation using the same model definition function. We trained the model to 20 epochs, for each model the optimizer is Adam with a learning rate as 0.1.Sparse Categorical Cross-entropy has been used as loss function. The models do not have a lot of layers. This is done to ensure that the models can even work on the limited computing power. We are using a GRU layer and Time Distribution layer in RNN. An Embedding, LSTM and Time Distribution layer in LSTM. A bidirectional layer, GRU layer and Time Distribution layer in Bidirectional RNN. A bidirectional layer, LSTM layer, Repeat Vector layer, LSTM layer and Time Distribution layer in Bidirectional LSTM. An Embedding Layer, bidirectional layer, GRU layer and Time Distribution layer in Bidirectional RNN. An Embedding Layer, bidirectional layer, LSTM layer, Repeat Vector layer, LSTM layer, three dense layers with ReLU activation and one dense layer with Sigmoid Activation layer and Time Distribution layer in Bidirectional LSTM.

Fig. 2. RNN Model

Fig. 3. LSTM Model

Fig. 4. Bidirectional RNN

Fig. 5. Bidirectional LSTM

Fig. 6. Bidirectional RNN with Embeddings

DOI: rp-9788770040723.087

Fig. 7.Bidirecteional LSTM with Embeddings

IV. Results

The LSTM model was the most accurate model. However, it suffered from overfitting. The rest of the models struggled to learn from the given datasets and were unable to give any predictions. To improve the model, we need a larger dataset with more computing power, which is possible with the use of cloud computing, to increase the model accuracy. The model needs to be deployed on Cloud services to use the optimum infrastructure available for translation and be available on demand. Cloud services can help the model to be accessible anytime. Using the cloud infrastructure will increase the speed of translation as well as the speed of processing training data. Cloud infrastructure also helps in multiple users using the model at any given point of time. We can then use this data to further increase the accuracy of the translation. Analysing the user base, we can optimise the model to be more available in regions where the model is used the most.

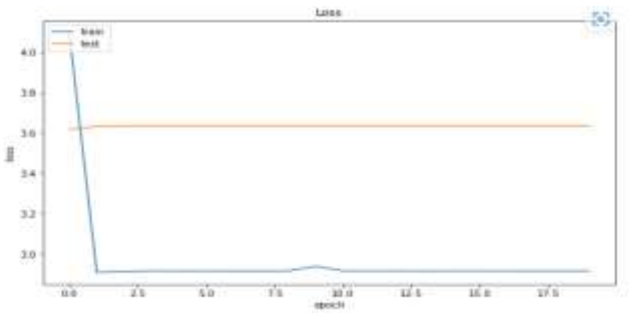

Fig. 8. Graph for RNN model

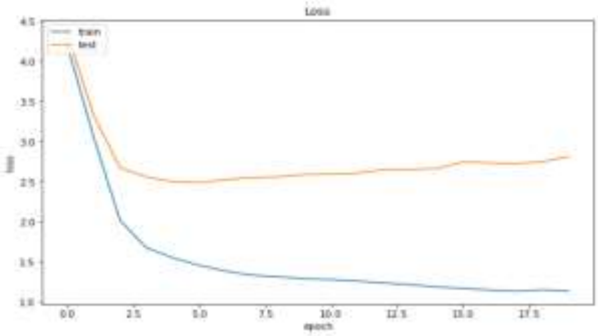

Fig. 9: Graph for LSTM Model

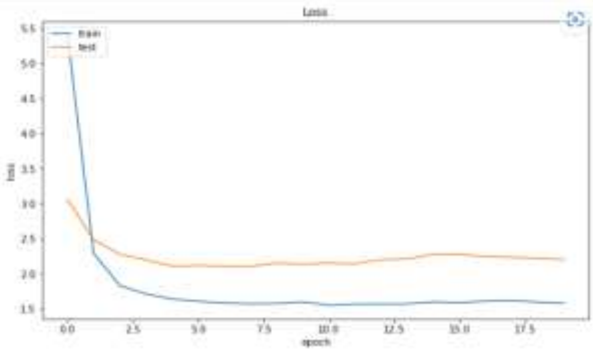

Fig. 10. Bidirectional RNN

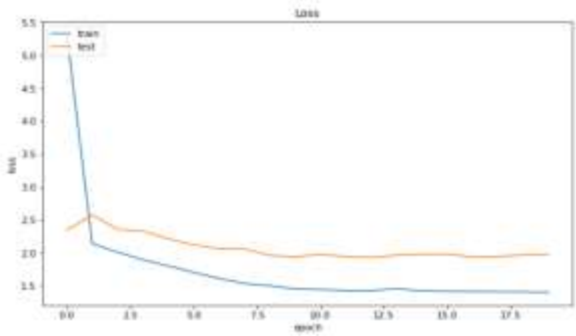

Fig. 11: Bidirectional LSTM

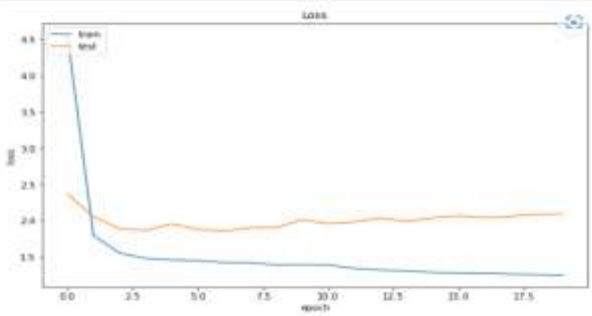

Fig.12. Bidirectional RNN with Embeddings

DOI: rp-9788770040723.087

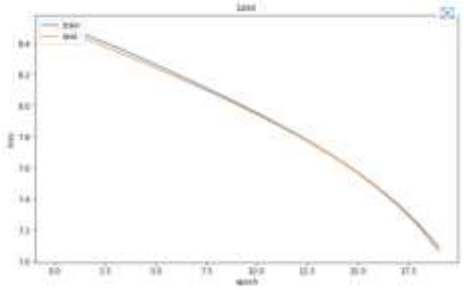

Fig.13: Bidirectional LSTM with Embeddings

V. DISCUSSION AND CONCLUSIONS

The above results illustrate the need for bigger datasets and as a result increase in computing power. This could be done on the cloud where the computing power is available on demand. The increased computing power will allow us to feed more data into the model leading to better learning and accurate predictions of the model. Using cloud infrastructure, we can also make the application be accessible anytime and anywhere. This results in multiple users accessing the application at a time. Increase in user base can then be further used to feed more data to the model and further increasing accuracy.

Also, from the available models, the LSTM model is the most preferred although it suffers from overfitting.

REFERENCES

[1] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush, "OpenNMT: Open-Source Toolkit for Neural Machine Translation," arXiv [cs.CL], 2017.

[2] F. Stahlberg, "Neural Machine Translation: A Review," J. Artif. Intell. Res., vol. 69, pp. 343–418, 2020.

[3] H. Wang, H. Wu, Z. He, L. Huang, and K. W. Church, "Progress in machine translation," Engineering (Beijing), 2021.

[4] S. Maruf, F. Saleh, and G. Haffari, "A survey on document-level neural machine translation: Methods and evaluation," ACM Comput. Surv., vol. 54, no. 2, pp. 1–36, 2022.

[5] M. R. Costa-Jussà and M. Farrús, "Statistical machine translation enhancements through linguistic levels: A survey,"ACMComput. Surv., vol. 46, no. 3, pp. 1–28, 2014.

[6] M.T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," arXiv [cs.CL], 2015

[7] C. Kobus, J. Crego, and J. Senellart, "Domain Control for Neural Machine Translation," arXiv [cs.CL], 2016.

[8] Y. Bar-Hillel, "The present status of automatic translation of languages," in Advances in Computers Volume 1, vol. 1, F. L. Alt, Ed. Elsevier, 1960, pp. 91–163.

[9] J. F. Dooley, "Cryptology before 1500 – A bit of magic," in History of Cryptography and Cryptanalysis, Cham: Springer International Publishing, 2018, pp. 13–23.

[10] W. J. Hutchins, "Machine Translation: A Brief History," in Concise History of the Language Sciences, E. F. K. Koerner and R. E. Asher, Eds. Elsevier, 1995, pp. 431–445.

[11] M. D. Okpor, "Machine Translation Approaches: Issues and Challenges," International Journal of Computer Science Issues, 2014.

[12] F. Meng, Z. Lu, H. Li, and Q. Liu, "Interactive Attention for Neural Machine Translation," arXiv [cs.CL], 2016.

[13] P. Koehn, "Europarl: A Parallel Corpus for Statistical Machine Translation," in Proceedings of Machine Translation Summit X: Papers, Sept 13-15 2005, pp. 79–86.

[14] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," arXiv [cs.CL], 2014.

[15] Z. Tu, Z. Lu, Y. Liu, X. Liu, and H. Li, "Modelling Coverage for Neural Machine Translation," arXiv [cs.CL], 2016.

[16] Gomathy, V., Janarthanan, K., Al-Turjman, F., Sitharthan, R., Rajesh, M., Vengatesan, K., &Reshma, T. P. (2021). Investigating the spread of coronavirus disease via edge-AI and air pollution correlation. ACM Transactions on Internet Technology, 21(4), 1-10.

[17] J. T. Reinert, Ed., "In-Q-Tel: The Central Intelligence Agency as Venture Capitalist." 2013.

[18] Gallafent and Alex, "Machine Translation for the Military." AnViOpenSource Knowledge Trust, 2011.

[19] Jackson and William, "GCN – Air force wants to build a universal translator." Government Computer News (GCN), 2003.

[20] S. Dupond, Ed., "A thorough review on the current advance of neural network structures." Annual Reviews in Control, 2019.

[21] M. Miljanovic, "Comparative analysis of recurrent and finite impulse response neural networks in time series prediction." Indian Journal of Computer Science and Engineering, 2012.

[22] S. Fernández, A. Graves, and J. Schmidhuber, "An application of recurrent neural networks to discriminative keyword spotting," in Lecture Notes in Computer Science, Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 220–229.

[23] N. ulSehr Zia, M. F. Naeem, S. M. K. Raza, M. M. Khan, A. Ul-Hasan, and F. Shafait, "A convolutional recursive deep architecture for unconstrained Urdu handwriting recognition," Neural Comput. Appl., vol. 34, no. 2, pp. 1635–1648, 2022.

[24] J. C. Heck and F. M. Salem, "Simplified minimal gated unit variations for recurrent neural networks," in 2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS), 2017, pp. 1593–1596.

[25] Rajesh, M., &Sitharthan, R. (2022). Introduction to the special section on cyber-physical system for autonomous process control in industry 5.0. Computers and Electrical Engineering, 104, 108481.

[26] T. K. Moon, "The expectation-maximization algorithm," IEEE Signal Process. Mag., vol. 13, no. 6, pp. 47–60, 1996.

[27] E. Teshome, "Bidirectional English-Amharic machine translation: An experiment using constrained corpus," Addis Ababa University, 2013.

[28] M. X. Chen et al., "The best of both worlds: Combining recent advances in neural Machine Translation," arXiv [cs.CL], 2018.

[29] D. He et al., "Dual Learning for Machine Translation," Advances in Neural Information Processing Systems, vol. 29, 2016.

[30] J. Senellart, P. Dienes, and T. Váradi, "New generation Systran translation system," in Proceedings of Machine Translation Summit VIII, 2001.

[31] Y. Song, J. Wang, Z. Liang, Z. Liu, and T. Jiang, "Utilizing BERT intermediate layers for aspect-based sentiment analysis and natural language inference," arXiv [cs.CL], 2020.