INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 7, Issue 4 - V7I4-1825)
Available online at: https://www.ijariit.com

Improved corpus base English to Hindi language translation sequence-based deep learning approach

Manmeet Kaur
manmeetvirk328@gmail.com
Punjabi University, Patiala, Punjab

Charanjiv Singh Saroa
charanjiv_saroa@yahoo.com
Punjabi University, Patiala, Punjab

ABSTRACT

While the NMT system operates conventional techniques such as rule-based machine translation and statistical machine translation, manual human translation still falls short. Our two NMT systems, RNN sequence-to-sequence and transformer-based models, are used in this paper for English-to-Hindi translation, and are compared to the current MT output for BLEU score. It outperforms current performance systems. However, a thorough review of the translations projected shows that in instances when an unknown word is recognised, blank lines emerge in the output and the source phrase is translated in a number of ways, our NMT systems need to be improved. In addition, the finding of the effect of the bi-gram model on the Hindi translation and relation between comparable Indian languages provides a new research route for direct translation between couples of similar languages. It may be possible to circumvent the limitation of available parallel data in low-resource languages by using linguistic similarities to get accurate results. With English to Hindi, an LSTM-based care mechanism enhances the MT output of the GRU-based NMT system. We also evaluated MT output performance in the Indian language, Hindi, using the BLEU-1, BLEU-2, and BLEU3 scores. For an Indian language like Hindi, it has been pointed out that it is not sufficient to assess on the basis of the BLEU1 score, as in prior research. In any configuration of NMT systems, the average BLEU score obtained is close to the matching bi-gram BLEU score.

Keywords- *Translation Hindi, Deep Learning, Score, Machine Learning*

1. INTRODUCTION

MT can be used as a great tool and when it is best to rely on "human" translators, then there is an insider's view of the difference. Machine translation systems are such applications or online services that use machine-learning techniques to translate into large amounts of text and in their supported languages. The service translates a "source" text from a language into a different "target" language. Although the concept is relatively simple to use machine translation techniques and interfaces, science and technologies are extremely complex behind it, and especially deep learning (artificial intelligence), large data, linguistics, cloud computing, and web API. Translation of text by a computer that does not have any human involvement. In the 1950s, the Pioneer, Machine Translation can be referred to as automatic translation, automatic or instant translation [1,5,46,47].

1.1 How Machine Translation works?

Generic MT mostly is referring to platforms such as Google Translate, Bing, Yandex, and Naver. These platforms provide MTs for advertising to millions of people. Companies can buy generic MTs for batch pre-translation and can connect to their system via APIs. Customizable MT refers to MT software that contains a basic component and can be trained to improve vocabulary accuracy in a chosen domain (medical, legal, IP, or company's own preferred terminology). For example, the WIPO specialist MT engine has translated the patent more accurately than the normalized MT engine, and the solution of eBay can understand and present hundreds of compressions used in electronic commerce. Adaptive MT suggests translators as they type in their CAT-tools, and learn from their inputs continuously in real-time. It is believed that in 2016 by the Lilt and by SDL, the adaptive MT translator is believed to be making significant improvements in productivity and can challenge future translation memory technology. There are more than 100 providers of MT technologies. Some of them are strictly MT developers, other translation firms and IT veterans [46].

1.2 Statistical VS Rule-Based Machine Translation

Statistical machine translation uses a statistical translation model whose parameters come from the analysis of monolingual and bilingual corporation. Creating a statistical translation model is a quick process, but technology relies heavily on the existing

multilingual corporation. For a specific domain, at least 2 million words and even more common is necessary for the general language. Theoretically, it is possible to reach quality limits, but most companies do not have such a large amount of existing multilingual corporation to make the necessary translation models. In addition, the statistical machine translation CPU is intensive and requires a comprehensive hardware configuration to run the translation model for the average performance level. Rule-based MT provides good quality of domain and nature is approximate. The dictionary-based customization guarantee guarantees quality and compliance with corporate vocabulary. But there may be a lack of expectation of the flow candidates in the translation results. In terms of investment, the adaptation cycle necessary to reach quality limits can be long and costly. Performance is also high on standard hardware [46,47, 52].

1.3 Neural Machine Translation

Neural Machine Translation is a machine translation approach that applies a large artificial neural network toward predicting the likelihood of a sequence of words, often in the form of whole sentences. Unlike statistical machine translation, which consumes more memory and time, neural machine translation, NMT, trains its parts end-to-end to maximize performance. NMT systems are quickly moving to the forefront of machine translation, recently outcompeting traditional forms of translation systems [9.10,11,12,13].

Continuous improvements in translations are important. However, performance improvements have plateaued with SMT technology since mid-2010. Taking advantage of the scale and power of Microsoft's AI supercomputers, especially the Microsoft Cognitive Toolkit, Microsoft Translator now provides neural networks (LSTM) based translation that enables a new decade of improved translation quality. These neural network models are available for all spoken languages through a text API using the Microsoft Speech and using the 'normal' category id. Neural network translations are fundamentally different from traditional SMT [13,26]. The following animation shows different phases neural network translations to translate a sentence. Due to this approach, the translation will take in the context of the complete sentence, versus only a few words sliding windows that use SMT technology will produce more fluid and human translated translations. Based on neural-network training, each word represents its unique characteristics within a special language pair (such as English and Chinese) with 500-dimension vector. Depending on the language pairs used for training, the nervous network itself will define what the dimension should be. They can encode simple concepts like gender (feminine, masculine, neutral), humility level (slang, casual, written, formal, etc.), the type of word (verb, noun, etc.), but any other non-obvious Features such as training data are taken from [20,28,29,36,46,].

1.4 How does Neural Machine Translation work?

As referenced above, unlike traditional methods of machine translation that involve separately engineered components, NMT works cohesively to maximize its performance. Additionally, NMT employs the use of vector representations for words and internal state. This means that words are transcribed into a vector defined by a unique magnitude and direction. Compared to phrase-based models, this framework is much simpler. Rather than separate component like the language model and translation model, NMT uses a single sequence model that produces one word at a time [21,22,31].

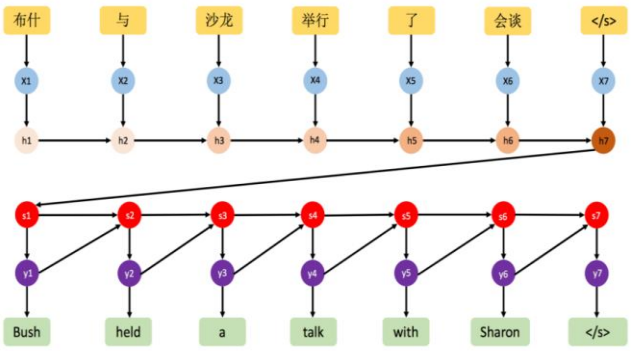

Figure 1: NMT Working [47]

The NMT uses a bidirectional recurrent neural network, also called an encoder, to process a source sentence into vectors for a second recurrent neural network, called the decoder, to predict words in the target language. This process, while differing from phrase-based models in method, prove to be comparable in speed and accuracy.

2. RELATED WORK

ing Zhai et al. in [2] have proposed several typologies to characterize the different translation processes. However, to the best of our knowledge, there has not been effort to automatically classify these fine-grained translation processes. Recently, an English-French parallel corpus of TED Talks has been manually annotated with translation process categories, along with established annotation guidelines. Based on these annotated examples, we propose an automatic classification of translation processes at sub sentential level. Experimental results show that the designers can distinguish non-literal translation from literal translation with an accuracy of 87.09%, and 55.20% for classifying among five non-literal translation processes. This work demonstrates that it is possible to automatically classify translation processes. Even with a small number of annotated examples, our experiments show the directions that we can follow in future work. One of the long-term objectives is leveraging this automatic classification to better control paraphrase extraction from bilingual parallel corpora.

Ankush Garg and Mayank Agarwal [5] proposed numerous methods in the past which either aim at improving the quality of the translations generated by them, or study the robustness of these systems by measuring their performance on many different

languages. In this literature review, discuss statistical approaches (in particular word-based and phrase-based) and neural approaches which have gained widespread prominence owing to their state-of-the-art results across multiple major languages.

Yuming Zhai et al. in [6] present a categorization of translation relations and then the designers annotate a parallel multilingual (English, French, Chinese) corpus of oral presentations, the TED Talks, with these relations. The long-term objective will be to automatically detect these relations in order to integrate them as important characteristics for the search of monolingual segments in relation of equivalence (paraphrases) or of entailment. The annotated corpus resulting from our work will be made available to the community.

Vu Cong Duy Hoang et al. in [9] present iterative back-translation, a method for generating increasingly better synthetic parallel data from monolingual data to train neural machine translation systems. The proposed method is very simple yet effective and highly applicable in practice. They demonstrate improvements in neural machine translation quality in both high and low resourced scenarios, including the best reported BLEU scores for the WMT 2017 hindi↔English tasks.

Myle Ott et al. in [10] shows that reduced precision and large batch training can speedup training by nearly 5x on a single 8-GPU machine with careful tuning and implementation. On WMT'14 English-German translation, we match the accuracy of Vaswani et al. (2017) in under 5 hours when training on 8 GPUs and then obtain a new state of the art of 29.3 BLEU after training for 85 minutes on 128 GPUs. The further improve these results to 29.8 BLEU by training on the much larger Paracrawl dataset.

Chen Mai Xu et al. in [11] tease apart the new architectures and their accompanying techniques in two ways. First, the designers identify several key modeling and training techniques, and apply them to the RNN architecture, yielding a new RNMT+ model that outperforms all of the three fundamental architectures on the benchmark WMT'14 English to French and English to German tasks. Second, the designers analyze the properties of each fundamental seq2seq architecture and devise new hybrid architectures intended to combine their strengths. The hybrid models obtain further improvements, outperforming the RNMT+ model on both benchmark datasets.

Hao Xiong et al. in [12] propose Multi-channel Encoder (MCE), which enhances encoding components with different levels of composition. More specifically, in addition to the hidden state of encoding RNN, MCE takes 1) the original word embedding for raw encoding with no composition, and 2) a particular design of external memory in Neural Turing Machine NTM) for more complex composition, while all three encoding strategies are properly blended during decoding. Empirical study on Chinese-English translation shows that our model can improve by 6.52 BLEU points upon a strong open source NMT system: DL4MT1.

Zhen Yang et al. in [13] proposed unsupervised neural machine translation (NMT) is a recently proposed approach for machine translation which aims to train the model without using any labeled data. The models proposed for unsupervised NMT often use only one shared encoder to map the pairs of sentences from different languages to a shared-latent space, which is weak in keeping the unique and internal characteristics of each language, such as the style, terminology, and sentence structure. To address this issue, the designers introduce an extension by utilizing two independent encoders but sharing some partial weights which are responsible for extracting high-level representations of the input sentences. Besides, two different generative adversarial networks (GANs), namely the local GAN and global GAN, are proposed to enhance the cross-language translation. With this new approach, we achieve significant improvements on English-German, English-French and Chinese-to-English translation tasks

3. THE PROPOSED METHOD

3.1 Proposed Methodology

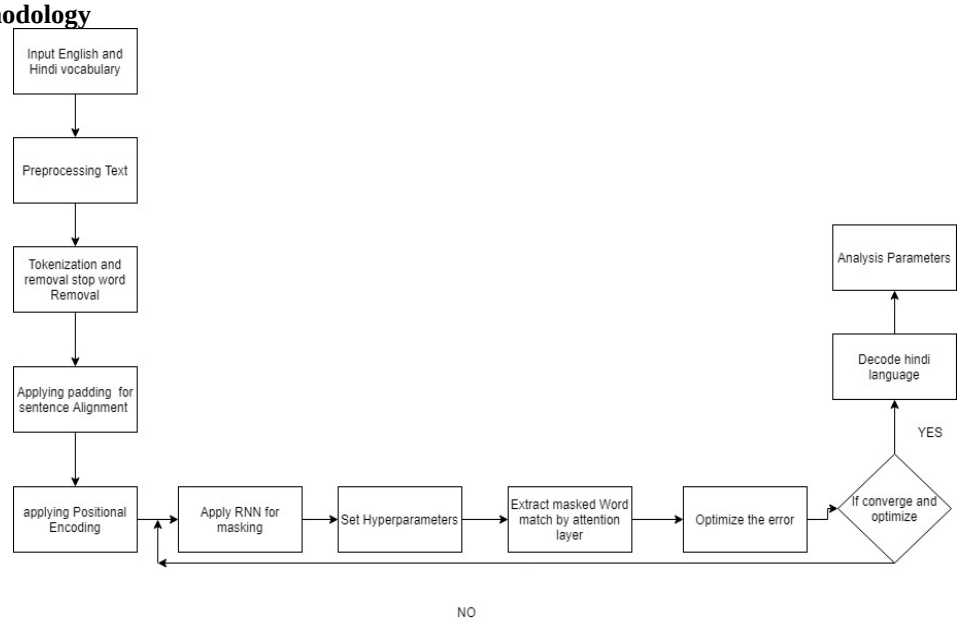

Figure 2: Proposed Flowchart

3.2 Proposed methodology: Flowchart

Step1: Input English and Hindi corpus for pre-processing the text.

Step2: Tokenization and padding the sentence alignment.
Step3: Apply Encoding by RNN approach
Step4: tuning the parameters by Adam optimization.
Step5: If the optimize then decode to English to Hindi
Step6: Analysis BLEU Score

3.3 Convolutional Neural Network

"A CNN model is made up of structural components. This triangular structure may be used to construct many phases.

- The convolutional layer is a crucial component of the CNN; it is the glue that holds the structure together. For the convolutional procedure, a kernel of size mn is swept over the input data, ensuring local connection and weight sharing".
- System-in-pairs: during the convolutional process, a filter examines the input matrices of the system. Each stage, the kernel filter's position in the matrix is shifted by a certain amount. By default, stride persists to a single value. If the stride is wrong, the boundary detail is lost in the model. This issue was addressed by adding more rows and columns to the matrices, so that they begin with all zeros. Zero-padding is the process of adding additional rows and columns to the results that contain no data.

4. RESULT ANALYSIS

4.1 Result Analysis

Performance Evaluation

BLEU compares the n-gram of the candidate's translation to the n-gram of the reference translation to calculate the number of matches. These matches do not rely on the position. The more exact the machine translation matches between the candidate and the reference translation.

$$\text{Brevity Penalty} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}$$

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

BP- brevity penalty
N: No. of n-grams, we usually use unigram, bigram, 3-gram, 4-gram
w_n: Weight for each modified precision, by default N is 4, w_n is 1/4=0.25
P_n: Modified precision
The BLEU measurement ranges from 0 to 1. The machine translation gets a score of one when it is identical to one of the reference translations. As a consequence, not even a human translator gets a score of 1.

Table 4.1 Translation proposed approach parameters

```
input_1 (InputLayer)            (None, None)         0
__________________________________________________________________________________________________
input_2 (InputLayer)            (None, None)         0
__________________________________________________________________________________________________
embedding_1 (Embedding)         (None, None, 300)    4209000     input_1[0][0]
__________________________________________________________________________________________________
embedding_2 (Embedding)         (None, None, 300)    5262300     input_2[0][0]
__________________________________________________________________________________________________
lstm_1 (LSTM)                   [(None, 300), (None, 721200      embedding_1[0][0]
__________________________________________________________________________________________________
lstm_2 (LSTM)                   [(None, None, 300),  721200      embedding_2[0][0]
                                                                 lstm_1[0][1]
                                                                 lstm_1[0][2]
__________________________________________________________________________________________________
dense_1 (Dense)                 (None, None, 17541)  5279841     lstm_2[0][0]
==================================================================================================
Total params: 16,193,541
Trainable params: 16,193,541
Non-trainable params: 0
__________________________________________________________________________________________________
```

```
In [133]:
k+=1
(input_seq, actual_output), _ = next(train_gen)
decoded_sentence = decode_sequence(input_seq)
print('Input Hindi sentence:', X_train[k:k+1].values[0])
print('Actual English Translation:', y_train[k:k+1].values[0][6:-4])
print('Predicted English Translation:', decoded_sentence[:-4])
```

```
Input Hindi sentence: आपने कई बार ये सुना होगा
Actual English Translation:  youve heard that saying
Predicted English Translation:  youve heard that saying
```

Figure 2: Proposed model predicted and actual translation (example-1)

In [134]:
```
k+=1
(input_seq, actual_output), _ = next(train_gen)
decoded_sentence = decode_sequence(input_seq)
print('Input Hindi sentence:', X_train[k:k+1].values[0])
print('Actual English Translation:', y_train[k:k+1].values[0][6:-4])
print('Predicted English Translation:', decoded_sentence[:-4])
```

```
Input Hindi sentence: मेसेज भेजते समय
Actual English Translation:  while textmessaging
Predicted English Translation:  while textmessaging
```

Figure 3: Proposed model predicted and actual translation(example-2)

In [135]:
```
k+=1
(input_seq, actual_output), _ = next(train_gen)
decoded_sentence = decode_sequence(input_seq)
print('Input Hindi sentence:', X_train[k:k+1].values[0])
print('Actual English Translation:', y_train[k:k+1].values[0][6:-4])
print('Predicted English Translation:', decoded_sentence[:-4])
```

```
Input Hindi sentence: और यह सब क्या मानव स्वभाव के बारे में हमें क्या बताता है
Actual English Translation:  and what all of this can tell us about human nature
Predicted English Translation:  and what all of this can tell us about human le
```

Figure 4: Proposed model predicted and actual translation(example-3)

In [136]:
```
k+=1
(input_seq, actual_output), _ = next(train_gen)
decoded_sentence = decode_sequence(input_seq)
print('Input Hindi sentence:', X_train[k:k+1].values[0])
print('Actual English Translation:', y_train[k:k+1].values[0][6:-4])
print('Predicted English Translation:', decoded_sentence[:-4])
```

```
Input Hindi sentence: चूहा एक पूरा जीव है आखिर में
Actual English Translation:  the rat is an entire organism after all
Predicted English Translation:  the rat is an entire organism after all
```

In [137]:
```
k+=1
(input_seq, actual_output), _ = next(train_gen)
decoded_sentence = decode_sequence(input_seq)
print('Input Hindi sentence:', X_train[k:k+1].values[0])
print('Actual English Translation:', y_train[k:k+1].values[0][6:-4])
print('Predicted English Translation:', decoded_sentence[:-4])
```

```
Input Hindi sentence: केवल बडे खेतों में इस्तेमाल के लिये ही बनायी गयी थी।
Actual English Translation:  and they were constructed for fields that were too large
Predicted English Translation:  and they were constructed for fields that were
```

Figure 5: Proposed model predicted and actual translation (example-4 and 5)

Table 2: Proposed and existing approach Bleu score training and testing scores

Approaches	Training-BLEU	Testing-BLEU
CNN-1gram	80	53.12
CNN-2gram	81.43	54.12
CNN-3gram	80.12	43.12
CNN-4gram	78.12	23.12
Proposed	98.12	92.12

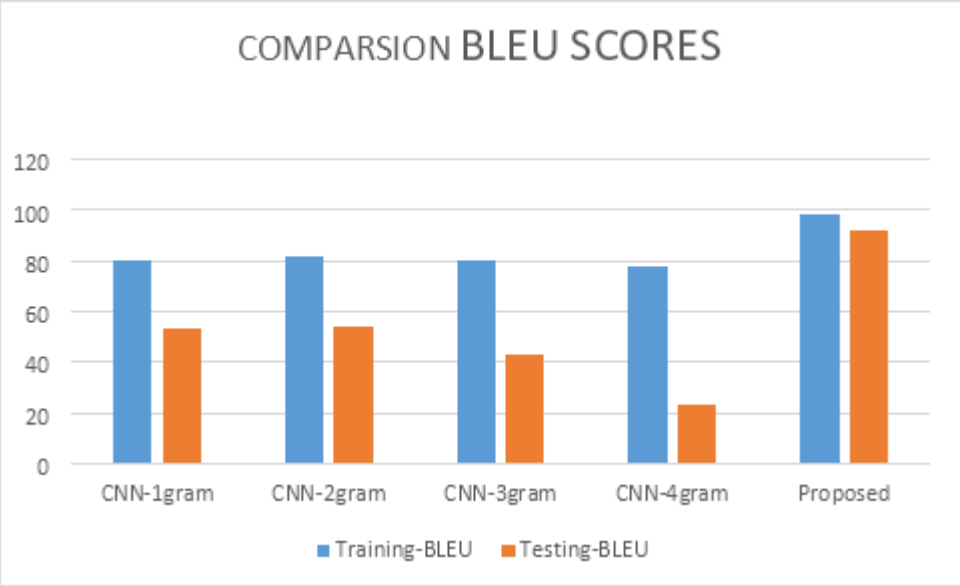

Figure 6:Training and Testing of BLEU

5. CONCLUSION

MT acts as a channel for cross-language communication in natural language processing (NLP). MT addresses problems of linguistic ambiguity by translating between two languages automatically while preserving its meaning. MT systems developed using a corpus-based approach based on a rule, that eliminated the need for language expertise, an endless list of NLP tasks such as the recognition of a named entity, language tagging parts, chunking, Word Sense disambiguation and the interlingua-based MT linguistic variation problem[1]. Corpus-based machine translation (MT) systems are usually classified as Example-based machine translation (EBMT), SMT, and natural language machine translation (NMT). The usage of EBMT is extremely limited. The encoder decoder RNN used conditional gated recurrent units (GRU) with an attention mechanism to obtain a BLEU score of 12.23 in English and Hindi. In [12], an LSTM with an attention mechanism was used by a sequence-to-sequence RNN to perform the same Translation and a BLEU score of 23.25, which is the top score in MTIL-2017 1. Neither of the NMT systems predicting translation has nevertheless completed an in-depth study of the effect of different ngram BLEU scores

6. REFERENCES

[1] Wei, X., Yu, H., Hu, Y., Zhang, Y., Weng, R. and Luo, W. Multiscale collaborative deep models for neural machine translation. *arXiv preprint arXiv:2004.14021*, 2020.

[2] Zhai, Y., Safari, P., Illouz, G., Allauzen, A. and Vilnat, A. Towards Recognizing Phrase Translation Processes: Experiments on English-French. *arXiv preprint arXiv:1904.12213*, 2019.

[3] Takushima, H., Tamura, A., Ninomiya, T. and Nakayama, H. Multimodal Neural Machine Translation Using CNN and Transformer Encoder. In *Proceedings of the 20th International Conference on Computational Linguistics and Intelligent Text Processing (CICLING 2019)*.

[4] Wu, L., Wang, Y., Xia, Y., Tian, F., Gao, F., Qin, T., Lai, J. and Liu, T.Y., Depth growing for neural machine translation. *arXiv preprint arXiv:1907.01968*, 2019.

[5] Garg, A. and Agarwal, M. Machine translation: A literature review. *arXiv preprint arXiv:1901.01122*, 2018.

[6] Zhai, Y., Max, A. and Vilnat, A. Construction of a multilingual corpus annotated with translation relations. In *Proceedings of the first workshop on linguistic resources for natural language processing*, pp. 102-111, 2018.

[7] Vyas, Y., Niu, X. and Carpuat, M. Identifying semantic divergences in parallel text without annotations. *arXiv preprint arXiv:1803.11112*, 2018.

[8] M. Q. Pham, J. Crego, J. Senellart, and F. Yvon. Fixing translation divergences in parallel corpora for neural MT in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 2967–2973, 2018.

[9] Hoang, V.C.D., Koehn, P., Haffari, G. and Cohn, T. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pp. 18-24, 2018.

[10] Ott, M., Edunov, S., Grangier, D. and Auli, M., 2018. Scaling neural machine translation. *arXiv preprint arXiv:1806.00187*.

[11] Chen, M.X., Firat, O., Bapna, A., Johnson, M., Macherey, W., Foster, G., Jones, L., Parmar, N., Schuster, M., Chen, Z. and Wu, Y. The best of both worlds: Combining recent advances in neural machine translation. *arXiv preprint arXiv:1804.09849*, 2018.

[12] Xiong, H., He, Z., Hu, X. and Wu, H. Multi-channel encoder for neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32, No. 1, 2018.

[13] Yang, Z., Chen, W., Wang, F. and Xu, B., Unsupervised neural machine translation with weight sharing. *arXiv preprint arXiv:1804.09057*, 2018.

[14] Sharma, A., Banerjee, P.S., Sharma, A. and Yadav, A. A French to English Language Translator Using Recurrent Neural Network with Attention Mechanism. In *International Conference on Nanoelectronics, Circuits and Communication Systems*, pp. 437-451, 2018, Springer, Singapore.

[15] Singh, S.P., Kumar, A., Darbari, H., Singh, L., Rastogi, A. and Jain, S. Machine translation using deep learning: An overview. In *2017 international conference on computer, communications and electronics (comptelix)*, pp. 162-167, 2017, IEEE.

[16] Deng, D. and Xue, N. Translation divergences in chinese–english machine translation: An empirical investigation. *Computational Linguistics*, *43*(3), pp.521-565, 2017.

[17] Bojanowski, P., Grave, E., Joulin, A. and Mikolov, T. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, pp.135-146, 2017.

[18] Yang, Z., Chen, W., Wang, F. and Xu, B. Improving neural machine translation with conditional sequence generative adversarial nets. *arXiv preprint arXiv:1703.04887*, 2017.

