International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Global Attention on BiLSTMs with BPE for English to Telugu CLIR

B N V Narasimha Raju[1], K. V. V. Satyanarayana[2], M. S. V. S. Bhadri Raju[3]

Submitted: 25/10/2023 **Revised:** 12/12/2023 **Accepted:** 25/12/2023

Abstract: An effective Cross-Lingual Information Retrieval (CLIR), will heavily rely on the accurate translation of queries and this is typically accomplished through Neural Machine Translation (NMT). NMT serves as a widely utilized method for translating queries from one language to another. In the present work, NMT is used to translate a query in English to the Indian language Telugu. For performing translation, NMT requires a parallel corpus. However the English-Telugu parallel corpora are resource-poor, so it may not be possible to supply the required amount of parallel corpus. The NMT will struggle to handle problems like Out Of Vocabulary (OOV) in resource-poor languages. The Byte Pair Encoding (BPE) mechanism will be helpful in solving OOV problems in resource-poor languages. In BPE, it segments the rare words into subword units and tries to translate the subword units. In NMT, the efficiency of translation still has issues in handling Named Entity Recognition (NER). The NER problems can be fulfilled using Bidirectional Long Short-Term Memories (BiLSTMs). The BiLSTMs will be helpful for training the system in the forward and backward directions for the dataset, which helps in recognizing the named entities. These NMT mechanisms will be sufficient for handling sentences without having long-range dependencies, but they will face issues while handling long-range dependencies in the sentences. Global Attention is useful to address these challenges, which is an integration between the encoder and decoder in NMT. This global attention mechanism proves beneficial in enhancing the translation quality, particularly for source sentences with long-range dependencies. In NMT, the Bilingual Evaluation Understudy (BLEU) scores and other parameters have shown that the efficiency in translating the source sentences is higher for global Attention on BiLSTMS with BPE than in regular models.

Keywords*: Attention, Global Attention, Cross-language IR, Bidirectional LSTMs, Byte pair encoding, Preprocessing, NMT*

1. Introduction

CLIR involves the task of identifying pertinent information within a collection of documents in a language other than the one in which the user is searching. Therefore, when a user needs data in multiple languages, CLIR becomes a valuable option for retrieving such data. The CLIR is mainly facing issues in translations for regional languages in countries like India and so on. Machine Translation (MT) is widely used for making such translations. In India, the CLIR is very useful because many people who are not conversant with English are showing interest in the data for regional languages. KPMG analysis in 2017 stated that there will be an increase in the Indian language internet users to 234 million by 2021 and in 2019 KPMG analysis stated that another 300 million Indian language internet users will be added to this by 2025 and by 2030 importance will be raised even more. Indeed, MT is useful for bridging the language barrier in CLIR. By translating the queries using MT, the users can be able to access the vast trove of content otherwise they are unable to access them. So, MT plays a major role in making CLIR an effective way of accessing information in other languages.

Earlier MT is performed by using direct translation which helps in translating the sentence from one language to another by using dictionaries. Later on, they shifted to corpus-based translations due to better performance than direct translations. In the domain of corpus-based translations, the performance of NMT is superior to that of Statistical Machine Translation (SMT). NMT leverages the fusion of neural networks with machine learning models and requires a substantial repository of English-Telugu languages parallel corpora to generate target sentences [1], [2]. The Telugu language is one of the most used Indian languages in the states like Andhra Pradesh and Telangana. In this paper, we have used global attention on BiLSTMs along with BPE for NMT which has better accuracy in generating target sentences than regular NMT models like BiLSTMs, Long Short-Term Memories (LSTMs) and so on.

The main phases in global attention on BiLSTMs with BPE are preprocessing, BPE, Encoding using BiLSTMs, global attention, and decoding using LSTMs. The scarcity of the parallel corpus for English-Telugu languages poses a sufficient challenge for MT systems. The limited availability of datasets necessitates collection of parallel corpora manually or using some automated tools but both mechanisms may introduce inconsistencies and noise into the parallel corpus. The Telugu language is morphologically rich and collecting the datasets manually will further complicate the situation by introducing replications and variations in word forms. All these issues will negatively impact the performance of NMT systems. In this,

[1] Research Scholar, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram-522502, India
[2] Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram-522502, India
[3] Professor, Department of Computer Science and Engineering, S R K R Engineering College, Bhimavaram-534204, India
** Corresponding Author Email: buddaraju.narasimharaju@gmail.com*

preprocessing plays a crucial role in addressing the issues regarding inconsistencies, noises, and replications in the parallel corpus. The preprocessing will be employed to enhance the performance of NMT.

NMT systems are mainly suffering from problems like OOV [3], NER and handling long-range dependencies. All these problems will lead the NMT systems towards poor performance [4]. The negative impact on the performance is also due to the scarcity of English-Telugu parallel corpora and high-frequency words in parallel corpus will lead to OOV problems. If frequent words are present in the source sentence, then the NMT systems translation will be good otherwise it will lead to poor performance. These are common problems for this kind of English-Telugu parallel corpora. BPE is a subword segmentation technique that addresses the OOV problem in NMT. BPE will iteratively merge the most frequent pairs of characters in a corpus. This creates a set of subwords that can be used to represent both known and unknown words. Now, try to translate subwords that are unknown words [5], [6]. BPE is shown to be effective in reducing the OOV rate in machine translation. The translations of the OOV words will be better by using BPE because it allows the NMT systems to learn the meaning of subwords, which can be used to translate OOV words.

In NMT, one of the crucial challenges is NER which helps in identifying the named entities. Named entities can be people, organizations and so on. NER is useful for MT, question answering and so on. BiLSTMs are well suited for NER tasks because they are able to learn the context information in both forward and backward directions. In NER, this is very important because the meaning of a word may depend on any word present in the source sentence. A unidirectional LSTM will learn the context in the forward direction only. So BiLSTMs are more effective than unidirectional LSTMs for NER tasks.

Handling the source sentences having long-range dependencies is one of the main issues in NMT. In regular NMT systems having an encoder-decoder will generate a fixed-length vector for the essential information in the source sentence, but it is challenging for longer sentences. In a standard encoder-decoder setup, performance declines with the increasing length of the source sentence. To address this concern, we've incorporated global attention alongside the encoder-decoder. Global attention will translate a word by searching the various positions in the source sentence to identify the crucial information in the sentence. The global attention model will forecast the translation for a word by using the context vector of the source sentences and the target words that were generated previously in the translation.

2. Related Work

Earlier SMT was used for making translations for the source sentences but later on, NMT came into existence which has better translation quality than SMT. Mai Oudah et al. proposed [7] a technique that is a combination of NMT and SMT. NMT is performing better than SMT but it is suffering from handling the short sentences. This problem was addressed by using this technique but still it has tokenization issues. The tokenization issues can be solved during preprocessing. Preprocessing is a crucial step in NMT as it significantly improves the quality of translation.

NMT has demonstrated superior translation quality compared to other methods such as SMT. However, NMT's performance can deteriorate when the number of unknown words increases, this is also known as OOV problems. To handle the OOV problem B N V Narasimha Raju et al. have utilized [8] a BPE mechanism that segments the unknown words and translates them. The LSTM in NMT will be relying on the leftward context of the source sentence to generate the next word in the translation. This approach is effective and sufficient if the meaning of a word translation depends on the leftward context but it will be insufficient if it depends on the rightward context. Mike Schuster et al. have suggested [9] a model called a bidirectional recurrent neural network, that has performed better when compared to a regular recurrent neural network.

NMT generates encouraging outcomes, as indicated by the findings of Sutskever et al. [10], who reported that the NMT generates a fixed-length vector to the input by employing a multi-layered LSTM. This vector is used for the creation of the target sequence by employing the deep LSTM network. Significantly it has achieved enhanced LSTM performance by reversing the source sentence word order. This reversal has created multiple short-term dependencies between input and output sentences, thereby it has streamlined the optimization.

NMT has multiple limitations when dealing with larger vocabulary sentences. The complexity in the training and decoding phases will be increased with the increase in target words. Sébastien Jean et al. [11] have introduced a method based on importance sampling, enabling the utilization of an extensive target vocabulary without escalating training complexity. Efficient decoding is achievable even with a significantly expanded target vocabulary, by selecting a smaller representative subset from the entire target vocabulary. The results indicate that models trained using this approach have better performance and even surpass baseline models with a smaller vocabulary. Many kinds of mechanisms have been applied in NMT to handle large sentences effectively, but they have faced several issues during translations.

3. Global Attention on BiLSTMs with BPE

Global attention on BiLSTMs with BPE consists of preprocessing, BPE, encoding (using BiLTMs), global attention, and decoding (using LSTM) as shown in Figure 1. NMT is heavily reliant on parallel data and for better translations a sizable amount of data is necessary. However, parallel corpora for resource-poor languages like English-Telugu frequently contain noise, inconsistencies, and replications. These issues in the parallel corpus can have an impact on the performance of NMT systems. One effective approach to address these issues is through preprocessing. In the preprocessing phase, the initial step is to convert all characters to lowercase and then undesirable characters are removed. Now duplicate entries within the parallel corpora are eliminated. To eliminate such entries the parallel corpus is converted into Unicode format, making it easier to identify and remove any duplicate entries in the data.

3.1. Byte Pair Encoding

In order to train the global attention on BiLSTMs, it will be heavily reliant on a parallel corpus consisting of languages like English-Telugu. It is crucial to show the special importance on the availability of the English-Telugu parallel corpus because it is limited due to resource constraints. In such cases, the sentences in the corpus may contain frequently used words, potentially leading to OOV issues [12] - [14]. To address this issue a BPE mechanism [6], [8] is utilized which serves as a valuable data compression technique and effectively merges frequently occurring byte pairs and this significantly mitigates OOV problems. The BPE mechanism is beneficial for segmenting words and starts with symbol vocabularies that are initially populated with character vocabularies. Delimiters are used to mark the end of character sequences and will play a vital role in reassembling the original tokens. The most frequent symbol pairs are replaced with an n-gram character. By merging frequent n-grams, a single symbol is created. The algorithm for BPE is given in Figure 2.

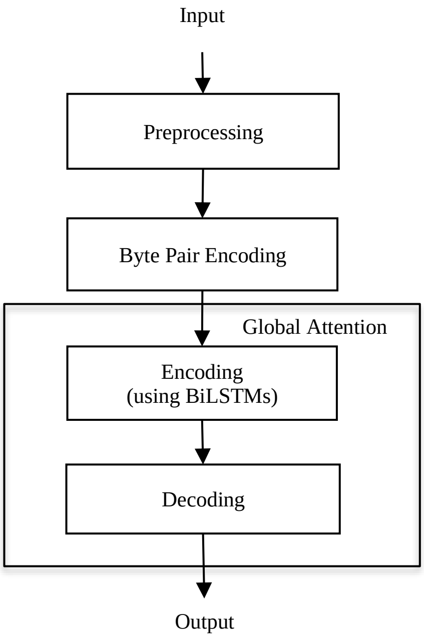

Fig. 1. Global Attention on BiLSTMs with BPE

```
BPE Algorithm
Input: $C$ is the string collection and the target vocabulary
    size is $v$.
BPE ($C$, $v$)
    • $W$ is the set of unique characters and $W \in C$
    • Repeat when $|W| < v$
        ○ $m$, $n$ are frequent bigrams and $m, n \in C$
        ○ $l$ is $m + n$
        ○ $W$ is $W + [l]$
        ○ Each instance of $m$, $n$ in $C$ is replaced with
          $l$
    • Return $W$.
```

Fig. 2. BPE Algorithm

A crucial step in this process involves counting the occurrences of symbol pairs within the corpus. The most frequent symbol pairs are replaced with an n-gram character. It's worth noting that BPE maintains vocabularies of the same size at both the initial and final stages of this process. The BPE algorithm is applicable to both the source and target vocabularies which leads to a reduction in the overall vocabulary size, making it more compact.

3.2. Bidirectional LSTMs

In NMT, after the application of BPE, the processed input is passed to the encoder. Applying global attention to BiLSTMs requires the encoder to be operated in a bidirectional but the decoder is operated in a unidirectional manner. In standard NMT, predicting the next token will rely on the preceding tokens in the sentence. In this scenario, a unidirectional LSTM as an encoder is adequate. However, in other tasks, the prediction of the next token may depend on both the leftward and rightward tokens in the sentence. In such cases, instead of unidirectional LSTM as an encoder, the BiLSTMs would be effective [15]. In BiLSTMs, the encoder consists of two separate unidirectional LSTM layers that are connected to process the same input in opposite directions. In the initial layer, the LSTM input is $x_1, x_2, ..., x_n$ and in the next layer, the LSTM input will be in reverse i.e. $x_n, x_{n-1}, ..., x_1$.

LSTM evaluates $\overrightarrow{h_t}$ to left reference of sentence of each word t. The $\overrightarrow{h_t}$ is given in equation (1)-(4).

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{qf}q_{t-1} + b_f) \quad (1)$$

$$q_t = (1 - f_t) \odot q_{t-1} + f_t \odot tanh(W_{xq}x_t + W_{hq}h_{t-1} + b_q) \quad (2)$$

$$p_t = \sigma(W_{xp}x_t + W_{hp}h_{t-1} + W_{qp}q_{t-1} + b_p) \quad (3)$$

$$h_t = p_t \odot tanh(q_t) \quad (4)$$

where σ is nothing but a function called sigmoid which performs element-wise operation and $\odot$ performs element-wise multiplication. LSTM evaluates an $\overrightarrow{h_t}$ to the left reference of the sentence and with right reference $\overleftarrow{h_t}$ is achieved by reading same sequence in the reverse direction. The representation of words in this model is a combination of both the representations and $h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}]$. In this, $h_t$will be helpful for finding the named entities.

3.3. Global Attention on BiLSTMs

We employ a stacked LSTM architecture [16] for our NMT systems, as depicted in Figure 3. Attention mechanisms [17], [18] hold a pivotal role in NMT. The concept of global attention [19], empowers the model to prioritize all tokens within the input sentence, regardless of their distance from the current token. This concept is applicable to sequence-to-sequence models, encompassing text summarization and MT. In these models, a series of hidden states are generated by the encoder for the input sentence. Subsequently, at each time step the decoder generates one token for constructing the output sentence. At each time step, hidden states of the encoder and present hidden states are utilized by the decoder, to show focus on the input sequence for producing the subsequent output token.

To each hidden state in the encoder, compute the weight to attain global attention. These calculated weights will be helpful in the formation of a context vector. This context vector holds significant importance in the decoder's role of generating the next output token. Here is an illustrative example of how global attention is employed in machine translation.

- The source sentence is passed to the encoder to generate a stream of hidden states.
- The generation of the target sentence will be initiated by the decoder and it produces only one word at a time.
- The decoder calculates the context vector at each time step by focusing on the hidden states of the encoder.
- The information in all the segments of the source sentence is encapsulated in a context vector and it aids in producing the output by decoder.

The target sequence is generated by the decoder until it encounters a token called end-of-sentence. Global attention is designed to learn and capture long-range relationships within the input sequence, greatly improving its efficiency while handling lengthy sequences. It stands as an important attention mechanism, which is capable of enhancing the generation of outputs in sequence-to-sequence architectures.

The global attention will consider the hidden states of the encoder while doing the calculation of the context vector, represented as c_v. An alignment vector is also utilized in this model and it is denoted as a_v, expressed in equation (5). The alignment vector is varied in length and on the source side, it corresponds to the number of time steps. The alignment vector is a comparison of each of the source hidden states $\overline{h_s}$ and with current target hidden state h_v.

$$a_v(s) = align(h_v, \overline{h_s}) = \frac{\exp\left(score(h_v, \overline{h_s})\right)}{\sum_{s'} \exp\left(score(h_v, \overline{h_{s'}})\right)} \quad (5)$$

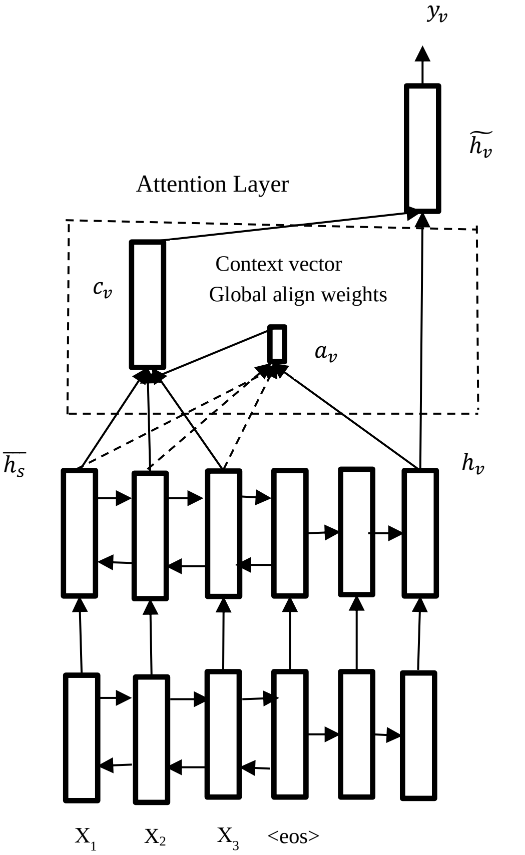

Fig. 3. Global Attention on BiLSTMs

In this context, the score is a function based on content, and it is computed using the following equation (6).

$$score(h_v, \overline{h_s}) = h_v^T W_a \overline{h_s} \quad (6)$$

The alignment vector a_v is used as weights and for the evaluation of context vector c_t, utilize the source hidden states $\overline{h_s} = (\overline{h_1}, \overline{h_2}, \ldots, \overline{h_r})$ as in equation (7)

$$C_v = \sum_{s=1}^{r} a_v(s)\overline{h_s} \quad (7)$$

In the decoding phase, for each time step v, the process starts with the utilization of the topmost layer's hidden state h_v from a stacked LSTM. The effective inclusion of pertinent source information requires the creation of a context vector c_v, and for projecting the enhancements in generating the current target word y_v. This is achievable by concatenating layer which is a combination of source context vector c_v and the target hidden state h_v. The fusion of information from these two vectors leads to the formation of an attentional hidden state, as demonstrated in equation (8).

$$\widetilde{h_v} = \tanh\,(W_c[c_v;\ h_v]) \quad (8)$$

The attentional vector, denoted as $\widetilde{h_v}$, is processed through a softmax layer to generate the predictive distribution, as specified in equation (9).

$$P(y_v/y_{<v}, x)\ =\ softmax(W_s\widetilde{h_v}) \quad (9)$$

During the generation of output sequences, the suggestion of unknown replacements [20], leads to a significant enhancement in Bilingual Evaluation Understudy (BLEU) scores. This finding confirms the effectiveness of our attentional models in learning valuable alignments for words that were not encountered previously.

Model evaluation metrics encompass a range of measures, including the BLEU score, cross-entropy, accuracy, and perplexity. The BLEU score gauges the quality of model predictions. The model employed for computing the loss function is Cross-entropy, The degree of correct classifications is quantified by accuracy, and perplexity assesses the precision of sample predictions made by the probability model.

4. Results and Discussions

The English-Telugu parallel corpus does not consist of any duplications, as these would entail multiple translations for a single source sentence. By eliminating all replicated sentences from the corpus, the system can produce accurate translations of the source text. The presence of duplications in the corpus can be a source of confusion during training, preventing the model's ability to acquire new features and leading to overfitting, which in turn diminishes translation performance. When both the train and test datasets share identical sentences, the resulting translations exhibit a high level of quality. However, when the model is assessed on unseen sentences, the quality of translations may decline. Therefore, addressing these issues within the parallel corpus can yield translations of superior quality.

Preprocessing is employed to eliminate inconsistent and noisy data, as well as to remove any duplications present in the parallel corpus. To evaluate performance, a comparison is made between NMT models using Global Attention on BiLSTMs with BPE and Bidirectional LSTMs with BPE. Both models are fed with the English-Telugu parallel corpus. These models are configured with two encoding layers and decoding layers. In both models, the LSTM layer is set to a size of 500, and they employ an Adam optimizer with a learning rate of 0.01. The decay rates for these models and dropout rates are fixed at 0.5 and 0.3, respectively. A total of 35,000 training steps were conducted. NMT performance is evaluated using these techniques on English-Telugu parallel corpora.

NMT performance is assessed by comparing techniques, specifically, Global Attention on BiLSTMs with BPE and BiLSTMs with BPE. Both these models are evaluated based on various parameters, including training and validation accuracies, training and validation cross-entropies, and training and validation perplexities. The specific parameter values tested within the systems are detailed in Table 1. Across all these parameters, it's evident that NMT with Global Attention on BiLSTMs with BPE consistently demonstrates superior performance.

Table 1. Parameters Comparison to Global Attention on BiLSTMs with BPE

Parameters		**Global Attention on BiLSTMs with BPE**	**BiLSTMs with BPE**
Training	Accuracy	97.83	97.25
	Perplexity	1.08	1.09
	Cross-Entropy	0.07	0.09
Validation	Accuracy	58.82	50.24
	Perplexity	79.68	107.9
	Cross-Entropy	4.37	4.68

The training accuracy comparison between global attention on BiLSTMs with BPE and BiLSTMs with BPE is depicted in Figure 4. Specifically, the training accuracy for global attention on BiLSTMs with BPE is 97.83, while for BiLSTMs with BPE, it stands at 97.25. In this context, a higher training accuracy indicates a stronger performing model, clearly establishing global attention on BiLSTMs with BPE as the superior performer in this parameter.

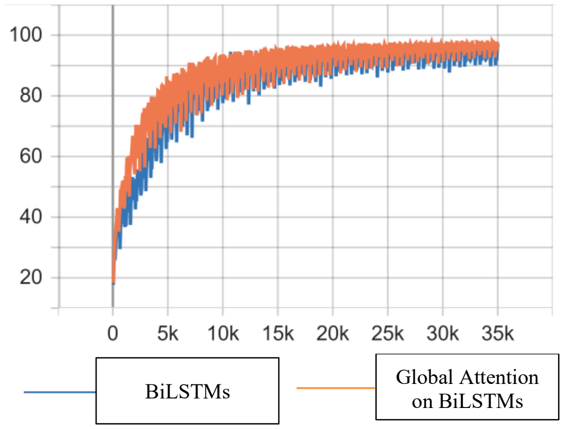

Fig. 4. Training Accuracy

The training perplexity comparison between global attention on BiLSTMs with BPE and BiLSTMs with BPE is depicted in Figure 5. Specifically, the training perplexity for global attention on BiLSTMs with BPE is 1.08, while for BiLSTMs with BPE, it stands at 1.09. In this context, a lower training perplexity indicates a stronger performing model, clearly establishing global attention on BiLSTMs with BPE as the superior performer in this parameter.

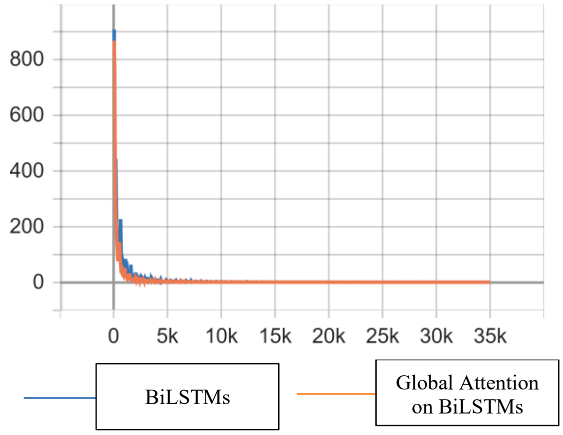

Fig. 5. Training Perplexity

The training cross-entropy comparison between global attention on BiLSTMs with BPE and BiLSTMs with BPE is depicted in Figure 6. Specifically, the training cross-entropy for global attention on BiLSTMs with BPE is 0.07, while for BiLSTMs with BPE, it stands at 0.09. In this context, a lower training cross-entropy indicates a stronger performing model, clearly establishing global attention on BiLSTMs with BPE as the superior performer in this parameter.

The validation accuracy comparison between global attention on BiLSTMs with BPE and BiLSTMs with BPE is depicted in Figure 7. Specifically, the validation accuracy for global attention on BiLSTMs with BPE is 58.82, while for BiLSTMs with BPE, it stands at 50.24. In this context, a higher validation accuracy indicates a stronger performing model, clearly establishing global attention on BiLSTMs with BPE as the superior performer in this parameter.

The validation perplexity comparison between global attention on BiLSTMs with BPE and BiLSTMs with BPE is depicted in Figure 8. Specifically, the validation perplexity for global attention on BiLSTMs with BPE is 79.68, while for BiLSTMs with BPE, it stands at 107.9. In this context, a lower validation perplexity indicates a stronger performing model, clearly establishing global attention on BiLSTMs with BPE as the superior performer in this parameter.

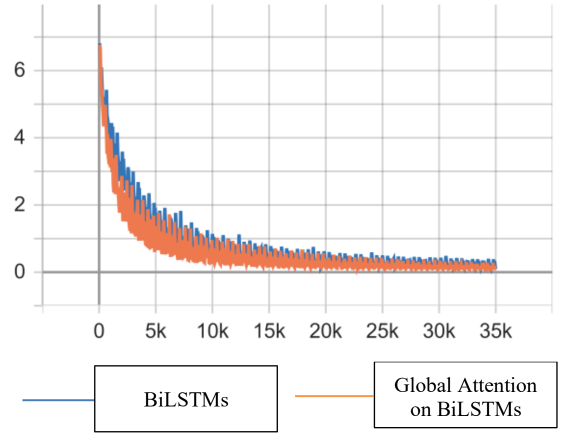

Fig. 6. Training Cross-Entropy

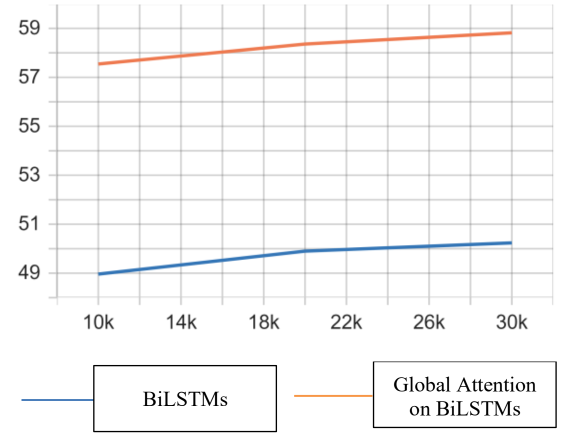

Fig. 7. Validation Accuracy

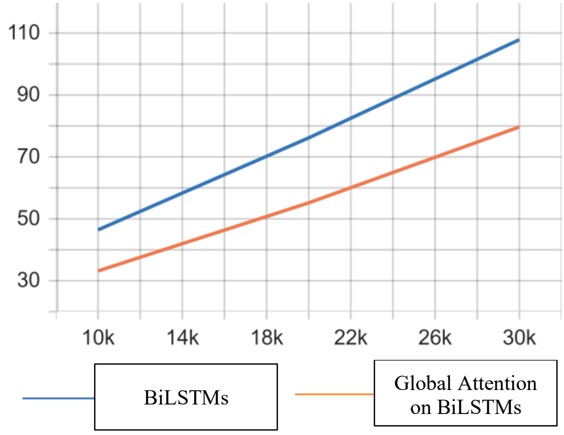

Fig. 8. Validation Perplexity

The validation cross-entropy comparison between global attention on BiLSTMs with BPE and BiLSTMs with BPE is depicted in Figure 9. Specifically, the validation cross-entropy for global attention on BiLSTMs with BPE is measured at 4.37, while for BiLSTMs with BPE, it stands at 4.68. In this context, a lower validation cross-entropy indicates a stronger performing model, clearly establishing global attention on BiLSTMs with BPE as the superior performer in this parameter.

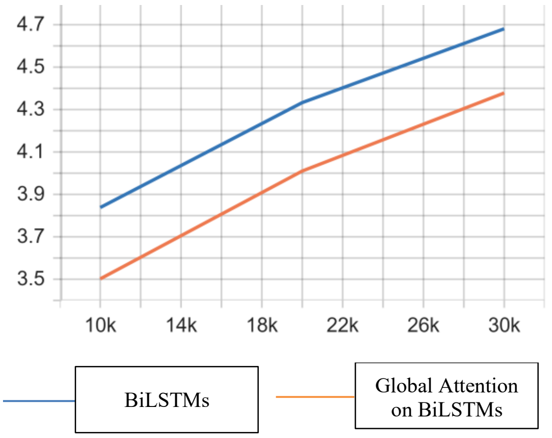

Fig. 9. Validation Cross-Entropy

The evaluation metric used for both models is the BLEU score, and the corresponding BLEU values are presented in Table 2. The BLEU score, which serves as the basis for comparing the performance between global attention on BiLSTMs with BPE and BiLSTMs with BPE, clearly demonstrates that NMT employing global attention on BiLSTMs with BPE produces more precise translations. The preprocessing and BPE models play a vital role in enhancing the parallel corpus quality and addressing OOV issues. This, in turn, significantly contributes to the NMT system's ability for generating more accurate translations.

Table 2. Performance in Comparison using BLEU Score

Model	BLEU Score
Global Attention on BiLSTMs with BPE	30.45
BiLSTMs with BPE	19.57

The comparison of BLEU scores for global attention on BiLSTMs with BPE and BiLSTMs with BPE is depicted in Figure 10. Specifically, the BLEU score for global attention on BiLSTMs with BPE is 30.45, whereas for BiLSTMs with BPE, it is 19.57. A higher BLEU score signifies a higher model, clearly establishing global attention on BiLSTMs with BPE as the stronger performer in this context.

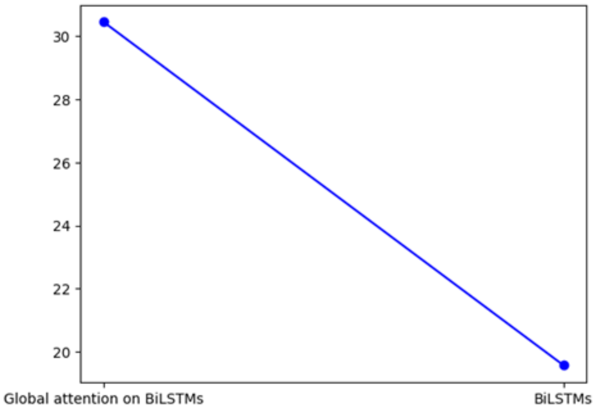

Fig. 10. BLEU Score

5. Conclusion

In NMT, enhancing translation quality involves eliminating replicates, noisy data, and inconsistencies. Additionally, addressing OOV problems, such as unknown words, is achieved through the use of BPE. In this study, the English-Telugu parallel corpus is utilized during the preprocessing stage, and the outcomes are then fed into the encoding phase. The English-Telugu language pairs serve for testing the model, and the evaluation metrics include BLEU scores, accuracy, cross-entropy, and perplexity to assess translation quality. The performance comparison between global attention on BiLSTMs with BPE and BiLSTMs with BPE demonstrates that global attention on BiLSTMs with BPE achieves higher accuracy and translations. BPE is effective in mitigating OOV issues, ultimately leading to improved translations, especially in resource-poor languages. Moreover, Bidirectional LSTMs within NMT excel in Named Entity Recognition (NER) compared to Unidirectional LSTMs. Bidirectional LSTMs incorporate both leftward and rightward context in sentence analysis, whereas Unidirectional LSTMs rely solely on leftward context. Consequently, Bidirectional LSTMs demonstrate improved translation accuracy. During the decoding phase, LSTM layers are used, and Global attention is applied to both the encoding and decoding phases to manage long-range dependencies within the sentences. The BiLSTMs will be helpful for enhancing the translation quality for shorter sentences but to handle longer sentences the global attention mechanism is useful and it can further enhance the translation efficiency. During the translation process, an unknown word replacement technique is employed, which contributes to an improved translation quality compared to the standard mechanism. The BLUE score for the global attention on BiLSTMs with BPE is 30.45 and without global attention is 19.57. Thus, NMT utilizing global attention on BiLSTMs with BPE emerges as the preferred choice for the English-Telugu corpus. Consequently, global attention on BiLSTMs with BPE proves instrumental in elevating translation accuracy in Cross-Language Information Retrieval.

References

[1] Karunesh Kumar Arora, Shyam S. Agrawal, “Pre-Processing of English-Hindi Corpus for Statistical Machine Translation,” Computación y Sistemas, pp. 725-737, 2017.

[2] Richard Kimera, Daniela N. Rim, Heeyoul Choi, “Building a Parallel Corpus and Training Translation Models Between Luganda and English,” Journal of KIISE, Vol. 49, No. 11, pp. 1009-1016, 2022.

[3] G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. (references)

[4] Thi-Vinh Ngo, Thanh-Le Ha, Phuong-Thai Nguyen, and Le-Minh Nguyen. “Overcoming the Rare Word Problem for low-resource language pairs in Neural Machine Translation,” In Proceedings of the 6th Workshop on Asian Translation, pp. 207–214, 2019.

[5] Mengjiao Zhang and Jia Xu, “Byte-based Multilingual NMT for Endangered Languages,” In Proceedings of the 29th International Conference on Computational Linguistics, pp. 4407–4417, 2022.

[6] Rico Sennrich, Barry Haddow and Alexandra Birch, “Neural Machine Translation of Rare Words with Subword Units,”In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, August 7-12, pp. 1715-1725, 2016, DOI: 10.18653/v1/P16-1162.

[7] Mai Oudah, Amjad Almahairi and Nizar Habash, “The Impact of Preprocessing on Arabic-English Statistical and Neural Machine Translation,”ArXiv.org, Aug. 19-23, pp. 214-221, 2019.

[8] B. N. V. Narasimha Raju, M. S. V. S. Bhadri Raju, K. V. V. Satyanarayana, “Effective preprocessing based neural machine translation for English to Telugu cross-language information retrieval,” IAES International Journal of Artificial Intelligence (IJ-AI), pp. 306-315, Vol. 10, No. 2, June 2021, DOI: 10.11591/ijai.v10.i2.pp306-315.

[9] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," IEEE Transactions on Signal Processing, vol. 45, no. 11, pp. 2673-2681, Nov. 1997.

[10] Sutskever, I., Vinyals, O., and Le, Q, “Sequence to sequence learning with neural networks,” Proceedings of the 27th International Conference on Neural Information Processing Systems,Vol. 2, pp. 3104–3112, 2014.

[11] Sébastien, J., Kyunghyun, C., Memisevic, R., and Bengio, Y, “On using very large target vocabulary for neural machine translation,” In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, pp. 1-10, 2015.

[12] Ali Araabi, Christof Monz, and Vlad Niculae, “How Effective is Byte Pair Encoding for Out-Of-Vocabulary Words in Neural Machine Translation?,” In Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track), pp. 117–130, 2022.

[13] Aloka Fernando and Surangika Ranathunga, “Data Augmentation to Address Out of VocabularyProblem in Low Resource Sinhala English Neural Machine Translation,” In Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation, pp. 61–70, 2021.

[14] Longtu Zhang and Mamoru Komachi, “Neural Machine Translation of Logographic Language Using Sub-character Level Information,” In Proceedings of the Third Conference on Machine Translation: Research Papers, pp. 17–25, 2018.

[15] Martin Sundermeyer, Tamer Alkhouli, Joern Wuebker, and Hermann Ney, “Translation Modeling with Bidirectional Recurrent Neural Networks,” Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 14–25, October 25-29, 2014, DOI: 10.3115/v1/D14-1003.

[16] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le, “Sequence to sequence learning with neural networks,” In Proceedings of the 27th International Conference on Neural Information Processing Systems, Vol. 2, pp. 3104–3112, 2014.

[17] Hamidreza Ghader and Christof Monz, “What does Attention in Neural Machine Translation Pay Attention to?,” In Proceedings of the Eighth International Joint Conference on Natural Language Processing, Vol. 1, pp. 30–39, 2017.

[18] Bahdanau, Dzmitry & Cho, Kyunghyun & Bengio, Y, “Neural Machine Translation by Jointly Learning to Align and Translate,” ArXiv. 1409, 2014.

[19] Thang Luong, Hieu Pham, and Christopher D. Manning, “Effective Approaches to Attention-based Neural Machine Translation,” In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 1412–1421, 2015.

[20] Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio,” On Using Very Large Target Vocabulary for Neural Machine Translation,” In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Vol. 1, pp. 1–10, 2015.