CME LIVE OPEN ACCESS

Engineering and Applied Sciences Journal

Research Article

Enhancing English to Amharic Machine Translation with Prior Knowledge Integration: Leveraging Syntactic Structures of the Source Language

Muluken Hussen Asebel[1*], Shimelis Getu Assefa[2] and Mesfin Abebe Haile[1]

[1]*Adama Science and Technology University, Adama, Ethiopia.*

[2]*University of Denver, Katherine A. Ruffatto Hall, 1999 East Evans Avenue Denver, CO 80208-1700, USA.*

Corresponding Author: Muluken Hussen Asebel, Adama Science and Technology University, Adama, Ethiopia.

Received: 2024 Nov 07 **Accepted:** 2024 Nov 27 **Published:** 2024 Dec 31

Abstract

Machine translation has made significant progress in automating the conversion of human languages via computational methods. However, achieving human-level performance remains challenging, particularly for languages such as Amharic. This paper aims to bridge this gap by integrating prior knowledge, particularly the syntactic structure of the source language, into graph neural networks for English-to-Amharic machine translation. Our objective is to systematically evaluate the effectiveness of integrating source language syntactic information into GNNs to improve English to Amharic machine translation quality. We conduct a thorough review of the relevant literature and describe the preprocessing steps for both existing and newly collected parallel corpora used in training. Our approach involves preprocessing data and discussing the proposed Graph2Seq models. The experimental results demonstrate a notable 4.56% increase in the bilingual evaluation understudy (BLEU) score compared with the baseline score, indicating a significant improvement in translation quality. Moreover, our models exhibit a 1.98% enhancement in the BLEU score over previous attempts, highlighting the value of integrating syntactic information into graph neural networks. Through meticulous experimentation and analysis, we illustrate the efficacy of incorporating source language syntax into GNNs for enhancing English-to-Amharic machine translation. This study advances machine translation systems, particularly for low-resource languages, and lays the foundation for future research in integrating syntactic knowledge across diverse linguistic tasks and languages.

Keywords: Graph Neural Networks, English, Amharic Language, Syntactic of Source Language, Prior Knowledge, Machine Translation, BLEU Score

1. Introduction

Machine translation, the automated conversion of human languages via computational methods, has been the subject of extensive research and development in recent years. While significant progress has been made in improving translation quality and efficiency, the field continues to face challenges in achieving human-level performance. Researchers have proposed leveraging previous translation knowledge, such as syntactic language structures, to reinforce neural machine translation (NMT) systems [1,2]. Despite these advancements, current NMT systems predominantly rely on sequential encoder-decoder architectures, which often overlook the explicit consideration of syntax or the hierarchical nature of language [3-5]. This underutilization of syntactic information in NMTs can be attributed to the difficulty of effectively integrating structured linguistic information into neural encoders, especially recurrent neural networks (RNNs).

Acknowledging the cautionary insights of researchers such as is essential, as they highlight the ongoing challenge of attaining human-like translation quality [6]. Despite advancements in machine learning and natural language processing, machines still struggle to emulate the nuanced understanding and context sensitivity inherent in human translation. On the other hand, scholars have emphasized the importance of knowledge-based approaches in machine translation [7,8]. Emphasized the need to integrate domain-specific knowledge and linguistic rules into translation systems to improve their accuracy and robustness. Similarly, the research of delves into the fusion of linguistic and worldly knowledge in meaning representation, advocating for a holistic approach to translation that goes beyond mere language processing [7]. Collectively, these studies underscore the pivotal role of integrating prior knowledge in machine translation. By providing syntactic information from the source language to encoder translation models, we can better understand and generate accurate translations. This holistic approach not only enhances translation quality but also lays the groundwork for advancing the capabilities of machine translation systems toward achieving human-like performance.

Citation: Asebel, M. H., Assefa, S. G., Haile, M. A. (2025). Enhancing English to Amharic Machine Translation with Prior Knowledge Integration: Leveraging Syntactic Structures of the Source Language. *Eng Appl Sci J, 1*(1), 1-9.

Attention-based NMT systems, proposed and use latent feature vectors to represent the words of the source sentence in the encoder [4,9,10]. These vectors are used for translation generation. Our objective is to automatically incorporate information about the syntactic relationships of the source language into the encoder, potentially improving the quality of the translation. Since the vectors correspond to words, it is logical to utilize a dependency syntax tree. Syntax dependency trees, shown in Figure 1, illustrate the syntactic relationships between words. In the example, "I" is the subject of the predicate "booked," and "ticket" is its object.

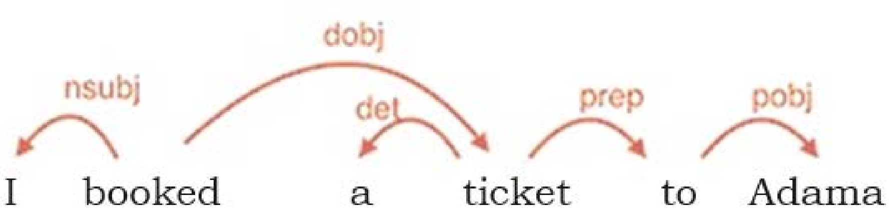

Figure 1: Syntax Dependency Tree for the Example Sentence: I Booked a Ticket to Adama

We employ graph neural networks (GNNs) to create syntax-aware feature representations for words. This approach leverages the structural information within syntactic graphs to enhance the understanding of word relationships and dependencies, resulting in more accurate and contextually relevant feature representations [4]. By using GNNs, we can effectively capture and incorporate the intricate syntactic patterns present in the language, leading to improved performance of the machine translation model. Research has underscored the effectiveness of integrating source language syntax as prior knowledge in graph neural networks (GNNs), resulting in notable improvements in natural language processing (NLP) overall and particularly in machine translation tasks [3-5]. These studies collectively highlight the significant potential of integrating source language syntax as prior knowledge in GNNs to increase performance levels in machine translation tasks.

Scholars have explored various methodologies for English-to-Amharic machine translation. For example, researchers like [11-16] have employed statistical machine translation. And have also worked on phone-based statistical machine translation, while focused on phrase-based statistical approaches. Additionally, combined context-based machine translation (CBMT) with recurrent neural networks and investigated the impact of normalized Amharic phonemes using a transformer model [17,18]. However, to the best of our knowledge, no studies have attempted to integrate source language syntax and employ an attention-based graph-to-sequence methodology for English-to-Amharic machine translation.

The objective of this research is to explore the integration of prior knowledge, specifically the syntactic structure of the source language, into graph neural networks (GNNs) for English-to-Amharic machine translation. By incorporating syntactic information during the translation process, the aim is to increase the accuracy and fluency of machine-translated text, thereby enhancing the performance of English-to-Amharic translation systems. We based our research on the findings of who utilized the transformer model and further refined it by fine-tuning it with the pretrained M2M100 48 M, which served as a baseline translation system [18]. This approach offers a benchmark against which we can assess the performance of our translation model. By leveraging their methodology, we can systematically evaluate the effectiveness of the proposed modifications or enhancements.

Section II presents an in-depth review of the relevant literature, followed by Section III, which elaborates on both the existing parallel corpus and the recently acquired corpus from a distinct domain and delves into the syntactical encoder. The procedural steps involved in preprocessing both corpora are outlined in Section IV. Section V delves into the discussion of the proposed Graph2Seq models, while Section VI presents the outcomes of the experiments. Finally, Section VII concludes the paper, providing reflections on future research directions.

1.1. Related Work

In this section, we explore the complexities of machine translation specifically tailored for the Amharic language. We shed light on the diverse approaches and challenges within this domain and the limited research dedicated to low-resource languages such as Amharic, largely due to the scarcity of parallel data [18]. Several approaches have been explored for translating Amharic to English, with differing levels of success. Achieved a BLEU score of 35.32% via the statistical machine translation (SMT) method, which was further elevated to 37.53% by incorporating phonemic transcription in 2020, conducted an evaluation of Amharic machine translation (MT) systems to assess their quality [14-20]. While the study revealed the potential of these systems, it also highlighted their relatively low BLEU scores. From an alternative perspective, these studies collectively underscore the potential for enhancing Amharic-English translation, particularly through the utilization of SMT and phonemic transcription methodologies.

Examined the effectiveness of employing a combination of context-based machine translation (CBMT) and recurrent neural network machine translation (RNNMT) for English–Amharic translation [17]. Their study revealed that this hybrid approach outperformed simple neural machine translation (NMT) and delved into the impact of dictionaries on translation quality. The findings showed that combining CBMT and RNNMT yielded enhanced translation results for English–Amharic translation, particularly with larger datasets such as the New Testament Bible. Additionally, the accuracy of the dictionary utilized by CBMT significantly influences the overall performance. Asserted that normalizing Amharic homophone characters can significantly augment the performance of Amharic-English machine translation in both directions [18]. Their study focused on a comprehensive Amharic-English parallel sentence dataset and examined the impact of Amharic homophone normalization on machine translation performance, with the objective of improving Amharic-English translation. The findings suggest that normalizing Amharic homophone characters leads to improved machine translation performance. Notably, the M2M-100 model outperforms the transformer-based models, and homophone normalization further enhances the performance of the NMT system.

In the domain of neural machine translation (NMT), the integration of prior knowledge emerges as a critical factor in augmenting translation quality. It plays a pivotal role in empowering systems to effectively manage lexical and syntactic ambiguity [21]. Additionally, the incorporation of similarity-aware NMT, which identifies promising sentences and leverages translation memory, has significantly reduced the workload of human translators [22]. Furthermore, recent investigations into the significance of context in NMT underscore the value of considering a broader unstructured context to enhance translation quality, as emphasized in studies such as [23]. In recent years, there has been a notable surge in interest regarding the application of graph neural networks (GNNs) in natural language processing (NLP), driven by their potential to capture the complex relationships and structures inherent in linguistic data. Scholars such as have made substantial contributions to this growing field through their comprehensive surveys, which offer detailed insights into the diverse subdomains of the NLP impacted by GNNs [24,25]. Their surveys meticulously categorize various areas of NLP research, ranging from sentiment analysis to machine translation, and explore how GNNs are employed within each domain. Moreover, they provide an overview of benchmark datasets and commonly used evaluation metrics tailored to GNN models, facilitating a deeper understanding of their performance and capabilities.

In particular, enriched the discourse by introducing a taxonomy of GNNs for NLP, delineating the research landscape into three main areas: graph construction, graph representation learning, and graph-based encoder-decoder models [23]. This framework offers a structured approach to understanding the diverse methodologies and techniques employed in GNN-based NLP research. Collectively, these research efforts highlight the growing importance of GNNs in NLP and underscore the need for continued research efforts to address the remaining challenges, such as scalability, interpretability, and generalization across different linguistic tasks and languages. Graph convolutional networks (GCNs) have emerged as a prominent methodology, leveraging predicted syntactic dependency trees of source sentences to generate word representations, or hidden states of the encoder, that incorporate syntactic neighborhood information. GCNs seamlessly integrate as layers into standard encoders, such as bidirectional RNNs or convolutional neural networks. An evaluation conducted through English–German and English–Czech translation experiments revealed substantial improvements over syntax-agnostic versions across all the setups. This approach underscores the potential of incorporating syntax-aware features to augment the capabilities of neural machine translation models [4].

2. Methodology

2.1. Parallel Dataset Preparation

The dataset used for training and evaluating English-to-Amharic machine translation is a parallel corpus created by [18]. This corpus is larger than previous ones and is freely available for research purposes. It was used to train neural machine translation models. Importantly, the neural machine translation models, especially those using subword units, achieved the highest BLEU scores, demonstrating their exceptional performance in this task [13]. The following table illustrates that compiled the most extensive parallel corpus for English to Amharic language pairs. In addition to the datasets outlined in Table 1, also played a role in MT research by developing a novel parallel corpus [11,18]. This corpus consists of 33,955 sentence pairs sourced from various news platforms, including the Ethiopian Press Agency, Fana Broadcasting Corporate, and Walt Information Center. As the data are drawn from diverse sources, they encompass a wide range of domains, such as religious texts, politics, economics, sports, and news.

Data source	# Sentence pairs	Accessible
Am-En ELRA-W0074	13,347	Yes
Biadgligne, Y., & Smaïli, K. (2021)	225,304	Yes
Horn MT[2]	2,030	Yes
Am-En MT corpus[3]	53,312	Yes
(Gezmu et al., 2022)	145,364	Yes
Abate et al. (2018)	40,726	Yes
Lison & Tiedemann (2016)	562,141	Yes
Tracey & Strassel (2020)	60,884	No
Admasethiopia[4]	153	Yes

MT Evaluation Dataset[5]	2,914	Yes
Destaw Belay et al. (2022)	33,955	Yes
Total	1,140,130	Yes
Unique sentence pairs	888,837	Yes

Table 1: Available Amharic and English Parallel Data

The dataset consists of approximately 1.1 million parallel sentences, with approximately 888,837 being distinct. This variation occurs because of duplicate sentences in the source materials. Notably, this collection of unique parallel sentences is the most extensive compilation achieved to date [18]. We have extensively used this dataset for our experimental endeavors.

2.2. Graph Neural Networks

Graph neural networks (GNNs) are a class of neural networks designed to perform machine learning tasks on graph-structured data. Unlike traditional neural networks that operate on vector spaces, GNNs can capture the relationships and dependencies between entities represented as nodes and edges in a graph. GNNs play a crucial role in the Graph2Seq architecture, which is specifically designed to handle structured data represented as graphs and generate sequential outputs. Graph2Seq uses an encoder–decoder framework that excels in machine translation. The GNN functions as the encoder, transforming the input graph composed of nodes and edges into node embeddings that capture both the structural and feature information of the graph. By aggregating information from neighboring nodes and iteratively updating node representations, the GNN effectively captures the topology and features of the graph.

The node embeddings generated by the GNN contain abundant information, representing the syntactic or structural characteristics of the graph. For tasks such as machine translation, these embeddings can encode the syntactic structure of a sentence, capturing dependencies and relationships between words. GNNs provide contextual embeddings for each node, taking into account the entire graph structure, which aids in capturing long-range dependencies and complex relationships. These contextual embeddings are vital for accurately generating sequences that are contextually relevant during the decoding phase.

2.3. Syntactic Encoder

In our proposed system, we seamlessly incorporate source language syntax into the translation process through the use of graph neural networks (GNNs). These GNNs are applied to the predicted syntactic dependency trees of the source sentences. Through this approach, the representations of words are made sensitive to their syntactic neighborhoods. This means that the relationships between words in a sentence, as represented by the syntactic dependency tree, are taken into account when encoding the source sentence. The GNNs within our system operate by taking word representations as inputs and producing word representations as outputs. They can be incorporated as layers into standard encoders, such as those on top of bidirectional RNNs or convolutional neural networks. This integration of syntax into the encoder allows the encoder to have access to rich syntactic information. Consequently, the encoder gains the flexibility to discern which aspects of syntax are beneficial to the machine translation task without imposing rigid constraints on their interaction. This adaptable approach ensures that the translation process remains dynamic and responsive to the intricacies of language structure. Overall, leveraging GNNs to incorporate syntactic information into the encoder enriches the translation process by harnessing the structural relationships among words in the source language sentences. This informed approach has the potential to improve translation quality by ensuring a deeper understanding of the underlying linguistic context.

2.4. The Proposed Graph2Seq Machine Translation Model

The overall structure of the sequence-to-sequence model (encoder-decoder), which is commonly used, is shown in Figure 1. This research aims to incorporate prior knowledge on the encoder side. Despite their versatility and capacity for expressive output, Seq2Seq models are constrained by a significant limitation: their applicability is limited to tasks with input data presented solely as sequences. However, sequences represent merely the fundamental form of structured data, while many critical problems demand a more sophisticated structure. For example, graphs, with their capacity to encapsulate intricate pairwise relationships within the data, are indispensable for addressing complex challenges [26]. Therefore, the proposed method uses a GNN, which enables us to embed prior knowledge, such as the syntax of the source language.

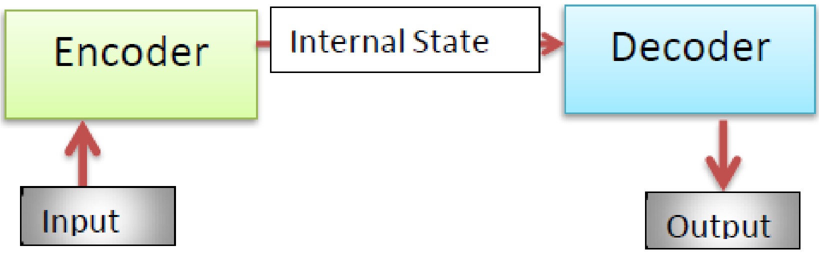

Figure 2: Encoder-Decoder Model

In our study, the encoder we employ utilizes a syntactic dependency tree of the source language, which undergoes processing via graph neural networks (GNNs). The construction of our encoder involves a series of steps aimed at maximizing the utilization of this graph-based representation. First, we represent the input sentence as a graph structure. We then incorporate two layers to facilitate the learning of node representations, leveraging this graph representation. These node representations serve as the basis for generating the attention-based context vector, which is then passed to the decoder. Notably, our architecture employs the standard transformer decoder. By focusing exclusively on the states of textual nodes, our approach empowers the decoder to dynamically leverage contextual information, thereby enhancing its translation capabilities [27].

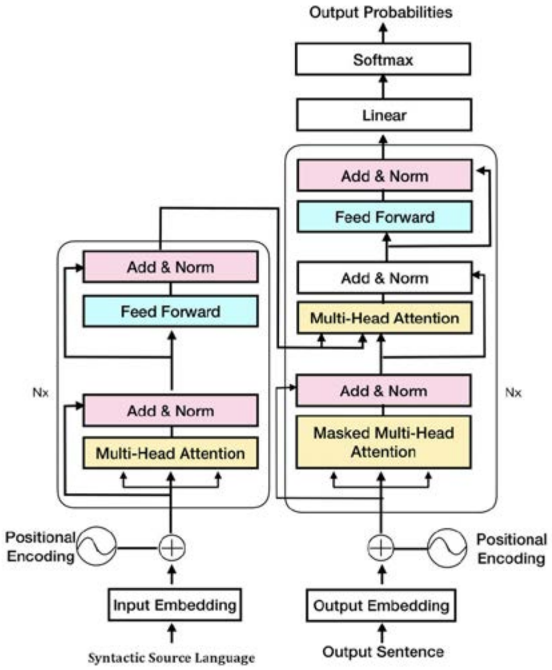

Figure 3: Transformer Encoder-Decoder Architecture; the Proposed English-to-Amharic Machine Translation Architecture

Figure 2 shows the main workflow of our study. The architecture consists of two main components: the encoder and the decoder. The encoder is responsible for processing the source language, which, in this case, is English, through multiple layers before passing it to the decoder. Conversely, the decoder receives input from the encoder and generates the target language, which in our context is Amharic. The preprocessed graph-based data undergo processing in an embedding layer before being fed into the stacked fusion layers. The first layer in the stack of fusion layers is the multihead self-attention layer. In this layer, self-attention mechanisms are used to generate contextual representations for each node, combining messages from neighboring nodes.

Formally, the contextual representations $\mathrm{C}_{\mathrm{x}}^{(\mathrm{l})}$ of all textual nodes are calculated as follows

$$\mathrm{C}_{\mathrm{x}}^{(\mathrm{l})} = \mathrm{MultiHead}\left(\mathrm{H}_{\mathrm{x}}^{\mathrm{l}-1}, \mathrm{H}_{\mathrm{x}}^{\mathrm{l}-1}, \mathrm{H}_{\mathrm{x}}^{\mathrm{l}-1}\right), \qquad (1)$$

where Multi Head (Q, K, V) is a multihead self-attention function that takes a query matrix Q, a key matrix K, and a value matrix V as inputs.

We also adopt positionwise feed forward networks $FFN(M_x^l)$ to generate textual node states $H_x^{(l)}$

$$H_x^{(l)} = FFN(M_x^l), \qquad (2)$$

where $M_x^{(l)} = \{M_{xi}^{(x)}\}$ denotes the above updated representations of all textual nodes.

On the side of the decoder, we employ a layer similar to the transformer decoder layer. In the method of Ld identical layers are stacked, where each layer l is made up of three sublayers, to create target-side concealed states [27]. To integrate the target and source-side contexts, the first two sublayers are masked self-attention and encoder-decoder attention:

$$E^l = MultiHead(S^{l-1}, S^{l-1}, S^{l-1}), \qquad (3)$$

$$T^l = MultiHead\left(E^l, H_x^{(Le)}, H_x^{(Le)}\right), \qquad (4)$$

where S^{l-1} denotes the target-side hidden states in the l-1-th layer. In particular, $S^{(0)}$ are the embeddings of the input target words. Then, a positionwise fully connected forward neural network is used to produce $S^{(l)}$ as follows:

$$S^{(l)} = FFN(S^{(l)}) \qquad (5)$$

Finally, the probability distribution of generating the target sentence is defined by using a Softmax layer, which takes the hidden states in the top layer as input:

$$P(Y\ |X) = \prod_t Softmax(WS_t^{Ld} + b) \qquad (6)$$

where X is the input sentence, Y is the target sentence (i.e., the Amharic sentence in our case), and W and b are the parameters of the Softmax layer.

3. Experimental Setup and Results

3.1. Experimental

To validate the effectiveness of our proposed system, we conducted experiments employing the attention-based Graph2Seq model [26,28]. This model, renowned for its ability to capture syntactic dependency trees within graph structures, served as a robust framework for our evaluation. We employed Google Colab to train our English-to-Amharic translation models. This entailed partitioning parallel sentences for English and Amharic into three distinct sets: 80% for training, 10% for validation, and 10% for testing purposes. We use the Adam optimizer with a learning rate of 0.001 [29]. The batch size is set to 32, and the hidden size is set to 128. We apply dropout with a probability of 0.1 between layers. We train for 50 epochs. Assessing model performance relies on the bilingual evaluation under study (BLEU) metric [30]. Ranging from 0 to 1, the BLEU score gauges the resemblance between the translated output and the reference. A score of 1 signifies a flawless match, whereas 0 denotes no matching words. In addition to the English-Amharic dataset, we evaluated the proposed model using English-Tigrinya dataset. Both Tigrinya and Amharic are Ge'ez-scripted Semitic languages that are low-resource and share considerable morphological and lexical similarities [31]. For this experiment, we used a parallel dataset of 340K English-Tigrinya sentence pairs.

Transformer: we used the Open NMT framework in conjunction with TensorFlow deep learning to train Transformer sequence-to-sequence models specifically for English to Amharic NMT. The training process was conducted from scratch [32]. To tokenize the text, we employed Byte Pair Encoding a subword tokenization method [33]. Byte Pair Encoding acts as a data compression algorithm that replaces the most frequently occurring pair of consecutive bytes with a byte that does not appear in the data. The model was trained using various. parameters, including 512 hidden units, 6 layers, a learning rate of 0.0001, a maximum step of 50K, a batch size of 32, and the Adam optimizer. Pre-trained model: To develop our bi-directional English-to-Amharic NMT system, we utilized the multilingual Facebook M2M-100 pre-trained model with 418M parameters [34]. For fine-tuning, the training and validation were conducted with a maximum source and target length of 128 per device. We used a batch size of 4 and trained for 4 epochs.

Parameters, including 512 hidden units, 6 layers, a learning rate of 0.0001, a maximum step of 50K, a batch size of 32, and the Adam optimizer. Pre-trained model: To develop our bi-directional English-to-Amharic NMT system, we utilized the multilingual Facebook M2M-100 pre-trained model with 418M parameters [34]. For fine-tuning, the training and validation were conducted with a maximum source and target length of 128 per device. We used a batch size of 4 and trained for 4 epochs. parameters, including 512 hidden units, 6 layers, a learning rate of 0.0001, a maximum step of 50K, a batch size of 32, and the Adam optimizer.

3.2. Pre-Trained Model

To develop our bi-directional English-to-Amharic NMT system, we utilized the multilingual Facebook M2M-100 pre-trained model with 418M parameters [34]. For fine-tuning, the training and validation were conducted with a maximum source and target length of 128 per device. We used a batch size of 4 and trained for 4 epochs.

4. Results and Discussion

In our experimental setup, we utilized a methodically normalized dataset, ensuring consistency and reliability across our analyses. Leveraging the syntactic source language, we use Graph2seq transformer models, capitalizing on their efficacy in capturing complex syntactic dependency tree patterns. To establish a robust benchmark, we engaged in meticulous fine-tuning on M2M100 418 M[9], a state-of-the-art pretrained language model tailored for English to Amharic translation. This step provided a solid foundation (baseline) for our subsequent evaluations.

Furthermore, we use the attention-based GNN2Seq model and Google Syntax Net, enriching the encoder with a syntactic dependency tree while maintaining the integrity of the decoder from the standard transformer architecture. This innovative approach allowed us to explore the nuanced influence of integrating syntax dependency into the graph neural networks of the encoder, revealing new insights into the interplay between syntactic structures and sequence generation. The following table shows the results.

Table 2 presents the results of three different translation models evaluated on the task of translating from English to Amharic and English to Tigrinya, with the performance metric being the BLEU score, which is a common metric used to evaluate the quality of machine-translated text.

Model	Direction	Result (BLEU Score)
Transformer	(English→Tigrinya)	12.69
M2M100 418 M (baseline)	(English→Tigrinya)	15.59
GNN2Seq (with syntactic integration)	(English→Tigrinya)	22.32
Transformer	(English→Amharic)	13.06
M2M100 418 M	(English→Amharic)	32.74
GNN2Seq (with syntactic integration)	(English→Amharic)	37.3

Table 2: Experimental Results for the BLEU Score

4.1. English→Tigrinya

we evaluated the performance of various machine translation models in translating from English to Tigrinya. The models considered include a standard Transformer model, the M2M100 418M model as a baseline, and the GNN2Seq model with syntactic integration. The transformer model achieved a BLEU score of 12.69, indicating that while it can perform the translation task, its accuracy and fluency are relatively low compared to more other models. In contrast, the M2M100 418M model, serving as the baseline, yielded a BLEU score of 15.59, which suggests a significant improvement over the Transformer model due to its more sophisticated architecture and larger training data. Moreover, the GNN2Seq model, which incorporates syntactic features into the translation process, achieved the highest BLEU score of 22.32, demonstrating a substantial improvement over both the Transformer and the baseline M2M100 418M models and highlighting the effectiveness of integrating syntactic information in enhancing translation quality. The comparative analysis of machine translation models for the English-to-Tigrinya language pair reveals that incorporating syntactic features into the translation process substantially enhances performance. The GNN2Seq model with syntactic integration outperforms both the Transformer model and the M2M100 418M model, achieving a BLEU score of 22.32. In comparing our results with the baseline, we observe a significant increase of 6.73% in the BLEU score, indicating a substantial improvement in translation quality. This improvement underscores the importance of leveraging linguistic structure in low-resource language translation tasks.

4.2. English→Amharic

The Seq2Seq (Transformer) model achieved a BLEU score of 13.06 while the pre-trained model achieved a BLEU score of 32.74, whereas the GNN2Seq model, with syntax integration, outperformed and achieved a higher BLEU score of 35.3. The BLEU score is a measure of how closely the generated translation matches human-generated reference translations. A higher BLEU score indicates better translation quality, with scores above 30 generally considered to be indicative of relatively good translation performance. The improvement in the BLEU score from the Seq2Seq model to the GNN2Seq model suggests that incorporating syntax dependency tree into the graph neural networks of the encoder, as in the GNN2Seq model, leads to enhanced translation quality. This finding indicates that leveraging syntactic information during the translation process can improve the accuracy and fluency of the translated text. Additionally, the difference in BLEU scores between the two models provides valuable insight into the effectiveness of integrating syntactic information into neural machine translation models.

Table 3 summarizes various studies on machine translation between English and Amharic or related languages, comparing the datasets used, the methodologies applied, and the resulting BLEU scores. The methods range from traditional statistical machine translation (SMT) and phrase-based SMT to more advanced neural machine translation (NMT) techniques and the use of pre-trained models like M2M100. The table also highlights the size of the datasets used in each study, showing a wide variation from as few as 1,915 sentence pairs to over a million, as well as the BLEU scores that measure the quality of the translations produced.

Authors	# Dataset used	Method(s)	BLEU score
Biadgligne & Smaïli (2022)	225,304	Neural machine translation	32.44
Abate et al. (2018)	40,726	Statistical machine translation	13.31
Teshome & Besacier (2012)	18,432	Phrase-based statistical machine translation	35.32
Ashengo et al. (2021)	8,603	Combination of context-based MT (CBMT) with RNN	11.34
Hadgu et al. (2020b)	1915	Google translate	9.6
Destaw Belay et al. (2022)	1,140,130	M2M100 418 M fine-tuning pre-trained model	32.74
Our work	1,140,130	Attention-based Graph2seq	37.3

Table 3: Previous Studies on English-Amharic Machine Translation Have Been Assessed in Terms of Dataset Size, Method(s) Used, and the BLEU Score Achieved

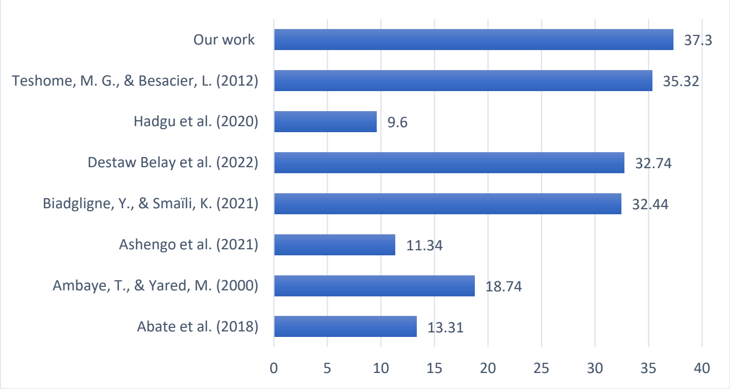

Figure 4: Reporting of Our Work Result with Other Previous Works of English-to-Amharic Machine Translation

In comparing our results with the baseline, we observe a significant increase of 4.56% in the BLEU score, indicating a substantial improvement in translation quality. Additionally, as shown in Figure 3, our models demonstrate a noticeable improvement of 1.98% in the BLEU score compared with previous attempts. This remarkable enhancement can be largely attributed to the careful incorporation of syntactic nuances from the source language into the graph neural networks. By integrating such intricate linguistic structures, our approach enhances the model's comprehension of sentence syntax and semantics, thereby enabling more precise and fluent translations.

5. Conclusion and Future Work

In conclusion, this study highlights the importance of integrating prior knowledge, specifically source language syntax, into GNN machine translation systems. The incorporation of syntax-aware features in GNN-based models shows promise for enhancing translation quality, particularly for low-resource languages such as Amharic and Tigrinya. We have shown consistent BLEU score improvements for challenging English–Amharic and English-Tigrigna language pairs. Future research should further explore the potential of syntactic integration in improving translation performance in the translation direction across diverse linguistic tasks and languages, contributing to the advancement of machine translation toward human-like performance [35-37].

Acknowledgement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Reference

1. Chen, K., Wang, R., Utiyama, M., & Sumita, E. (2021). Integrating prior translation knowledge into neural machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30,* 330-339.
2. Yang, Y., Li, X., Jiang, T., Kong, J., Ma, B., Zhou, X., & Wang, L. (2017, November). Improving adversarial neural machine translation with prior knowledge. In *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)* (pp. 1373-1377). IEEE.
3. Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv*:1409.0473.
4. Bastings, J., Titov, I., Aziz, W., Marcheggiani, D., & Sima'an, K. (2017). Graph convolutional encoders for syntax-aware neural machine translation. *arXiv preprint arXiv*:1704.04675.
5. Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems,* 27.
6. Isabelle, P., & Foster, G. (2005). Machine translation: overview.
7. Mahesh, K., & Nirenburg, S. (1996). Meaning representation for knowledge sharing in practical machine translation. *Proe. the FLAIRS-96 Track on Information Interchange.*
8. Nirenburg, S. (1989). Knowledge-based machine translation. *Machine Translation, 4*(1), 5-24.
9. Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv*:1409.0473.
10. Luong, M. T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv*:1508.04025.
11. Biadgligne, Y., & Smaïli, K. (2022, March). Offline corpus augmentation for english-amharic machine translation. In *2022 5th International Conference on Information and Computer Technologies* (ICICT) (pp. 128-135). IEEE.
12. Abate, S. T., Melese, M., Tachbelie, M. Y., Meshesha, M., Atinafu, S., Mulugeta, W., ... & Shifaw, S. (2018, August). Parallel corpora for bi-directional statistical machine translation for seven ethiopian language pairs. In *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing* (pp. 83-90).
13. Gezmu, A. M., Nürnberger, A., & Bati, T. B. (2021). Extended parallel corpus for Amharic-English machine translation. *arXiv preprint arXiv*:2104.03543.
14. Teshome, M. G., & Besacier, L. (2012, May). Preliminary experiments on English-Amharic statistical machine translation. In *SLTU* (pp. 36-41).
15. Teshome, M. G., Besacier, L., Taye, G., & Teferi, D. (2015, September). Phoneme-based English-Amharic statistical machine translation*. In AFRICON 2015* (pp. 1-5). IEEE.
16. Teshome, M. G., Besacier, L., Taye, G., & Teferi, D. (2015b). Phoneme-based English-Amharic statistical machine translation. AFRICON 2015, 1–5.
17. Ashengo, Y. A., Aga, R. T., & Abebe, S. L. (2021). Context based machine translation with recurrent neural network for English–Amharic translation. *Machine translation, 35*(1), 19-36.
18. Destaw Belay, T., Lambebo Tonja, A., Kolesnikova, O., Muhie Yimam, S., Ayele, A. A., Bogale Haile, S., ... & Gelbukh, A. (2022). The effect of normalization for bi-directional amharic-english neural machine translation. *arXiv e-prints, arXiv*-2210.
19. Hadgu, A. T., Beaudoin, A., & Aregawi, A. (2020). Evaluating amharic machine translation. *arXiv preprint arXiv*:2003.14386.
20. Hadgu, A. T., Beaudoin, A., & Aregawi, A. (2020). Evaluating amharic machine translation. *arXiv preprint arXiv*:2003.14386.
21. Moussallem, D., Wauer, M., & Ngomo, A. C. N. (2018). Machine translation using semantic web technologies: A survey. *Journal of Web Semantics, 51*, 1-19.
22. Zhang, T., Huang, H., Feng, C., & Wei, X. (2020). Similarity-aware neural machine translation: reducing human translator efforts by leveraging high-potential sentences with translation memory. *Neural Computing and Applications, 32*(23), 17623-17635.
23. Popescu-Belis, A. (2019). Context in neural machine translation: A review of models and evaluations. *arXiv preprint arXiv:*1901.09115.
24. Wu, L., Chen, Y., Shen, K., Guo, X., Gao, H., Li, S., ... & Long, B. (2023). Graph neural networks for natural language processing: A survey. *Foundations and Trends® in Machine Learning, 16*(2), 119-328.
25. Liu, X., Su, Y., & Xu, B. (2021, December). The application of graph neural network in natural language processing

and computer vision. In *2021 3rd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)* (pp. 708-714). IEEE.

26. Xu, K., Wu, L., Wang, Z., Feng, Y., Witbrock, M., & Sheinin, V. (2018). Graph2seq: Graph to sequence learning with attention-based neural networks. *arXiv preprint arXiv*:1804.00823.
27. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems, 30.*
28. Itoh, T. D., Kubo, T., & Ikeda, K. (2022). Compositionality-Aware Graph2Seq Learning. *arXiv preprint arXiv*:2201.12178.
29. Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.
30. Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311-318).
31. Feleke, T. L. (2017, April). The similarity and mutual intelligibility between Amharic and Tigrigna varieties. In *Proceedings of the fourth workshop on nlp for similar languages, varieties and dialects (vardial)* (pp. 47-54).
32. Klein, G., Kim, Y., Deng, Y., Senellart, J., & Rush, A. M. (2017). Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv*:1701.02810.
33. Gage, P. (1994). A new algorithm for data compression. *The C Users Journal, 12*(2), 23-38.
34. Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., ... & Joulin, A. (2021). Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research, 22*(107), 1-48.
35. Asebel, M. H., Assefa, S. G., & Haile, M. A. (2024). Enhancing English to Amharic machine translation with prior knowledge integration: Leveraging syntactic structures of the source language.
36. Lison, P., & Tiedemann, J. (2016). Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles.
37. Tracey, J., & Strassel, S. (2020, May). Basic language resources for 31 languages (plus English): The LORELEI representative and incident language packs. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)* (pp. 277-284).

Citation: Asebel, M. H., Assefa, S. G., Haile, M. A. (2025). Enhancing English to Amharic Machine Translation with Prior Knowledge Integration: Leveraging Syntactic Structures of the Source Language. *Eng Appl Sci J, 1*(1), 1-9.