DEPENDENCY TREE PARSING USING TRANSFORMER-BASED LANGUAGE REPRESENTATIONS

Nwe Nwe Win

University of Computer Studies,Yangon,
Yangon, Myanmar
Nwenwewin1@ucsy.edu.mm
http://ucsy.edu.mm

Win Pa Pa

Naypyitaw State Polytechnic University,
Naypyitaw, Myanmar
winpapa@nspu.edu.mm
http://ucsy.edu.mm

Nang Kham Htwe

Associate Professor, Polytechnic University, Bamaw,
Bamaw, Myanmar
nangkhamhtwe@ucsy.edu.mm
http://ucsy.edu.mm

Abstract

Dependency parsing in Myanmar face many challenges due to agglutinative morphology, flexible word order, and lack of explicit word boundaries. This paper addresses these challenges by investigating the performance of Myanmar dependency parsing based three pretrained-transformer models: XLM-RoBERTa-Large, XLM-RoBERTa-base-Longformer-4096, and language-specific model, MyanBERTa. The proposed methodology utilizes a parameter-efficient model with adapter layers and a biaffine parsing mechanism by optimizing multi-task learning objective across six linguistic prediction tasks. The experimental results demonstrated that XLM-RoBERTa-Large achieves the highest Labeled Attachment Score (LAS) and Unlabeled Attachment Score (UAS) on both development set and test set. Despite XLM-RoBERTa-base-Longformer-4096 model can extended its input capacity and MyanBERTa is pretrained on a language capacity, their performance drop compared to XLM-RoBERTa-Large. These findings suggest that the scale and architectural strength of models are more critical for high-performance dependency parsing than models with an extended input capacity or those trained on a language-specific corpus.

***Keywords*: dependency parsing; biaffine parsing; labeled attachment score; unlabeled attachment score.**

1. Introduction

Dependency parsing is instrumental uncovering the grammatical relationships between words within a sentence by representing as a directed tree structure in natural language processing. This foundational analysis is critical for a myriad of downstream applications such as machine translation, information extraction, and advanced question answering systems. Dependency parsing in low-resource languages like Myanmar is challenging due to agglutinative morphology, flexible word order, and the lack of explicit word boundaries, which complicate tokenization and syntactic analysis. These linguistic factors and the poor representation of Myanmar in pre-trained language models face challenges in optimizing model performance and lead to degrade parsing accuracy.

The landscape of NLP has been widely transformed by the emergence of transformer-based pre-trained language models [4,10], which provide rich contextualized representations crucial for enhancing parsing capabilities. Among these, RoBERTa [4] has consistently demonstrated superior performance across various linguistic tasks by establishing itself as a powerful baseline. In the realm of dependency parsing, RoBERTa-based models leverage their deep layer-wise encoding and multi-head self-attention mechanisms to effectively capture intricate syntactic dependencies. Therefore, the proposed dependency parsing model is built on top of RoBERTa to model a robust framework for capturing rich contextual representations that are crucial for constructing accurate dependency trees.

This paper focus on modeling and investigation the efficiency of the transformer-based dependency parsing for the Myanmar language. We conduct a comparative analysis of two prominent RoBERTa-based models: XLM-RoBERTa-Large [4] which is a multilingual model renowned for its cross-lingual transfer capabilities, and XLM-RoBERTa-base-Longformer-4096 [2], a variant engineered to process significantly longer input sequences. We first implemented a dependency parsing model based on RoBERTa-Large architecture, which accepts input sequences of up to 512 tokens. This limitation hindered its ability to effectively process longer sentences in the Myanmar language. To overcome this, we developed an extended model capable of handling inputs up to 4096 tokens. Although the XLM-RoBERTa-base-Longformer-4096 extends the input capacity to handle longer sequences, it surprisingly fails to outperform XLM-RoBERTa-Large in parsing accuracy as evidenced by their comparative LAS and UAS scores. Although the model with a 4096-token input can process longer sentences, the 512-token embedding of XLM-RoBERTa-Large performs better in encoding syntactic structures essential for accurate parsing. This suggests that merely increasing the maximum input length through architectural modifications alone may not inherently enhance a model's ability to capture complex syntactic dependencies.

In addition, we also built and analysis the Myanmar dependency parsing based on MyanBERTa[7] which is a RoBERTa-base model specifically pre-trained on Myanmar language data. Our objective is to highlight the impact of model scale, architectural design (standard sequence vs. long-sequence), and the relative effectiveness of language specific pre-training and broad multilingual pre-training in enhancing dependency parsing performance for Myanmar language. However, MyanBERTa generally yields lower LAS and UAS scores compared to RoBERTa-large.

Although MyanBERTa adopts the RoBERTa architecture, its dependency parsing performance is lower than RoBERTa-large due to limitations in both scale and training data. MyanBERTa is implemented with a BERT base-sized configuration (L=12 layers, H=768 hidden size, A=12 attention heads; ~110M parameters) and trained only on Myanmar-specific corpora using a 30,522 byte-level BPE vocabulary. The restricted size and syntactic diversity of these resources limit the model’s ability to generalize across a wide range of linguistic structures. In contrast, RoBERTa-large employs a substantially larger architecture (L=24 layers, H=1024, A=16; ~355M parameters) and is pretrained on massive multilingual corpora without language-specific constraints. This large-scale pretraining not only increases model capacity but also exposes the model to diverse syntactic patterns across languages by enabling effective cross-linguistic transfer. Such transfer is particularly valuable for Myanmar where annotated data is scarce. Consequently, while both models share the same underlying RoBERTa design, RoBERTa-large’s greater parameter capacity and multilingual training allow it to achieve superior dependency parsing performance as demonstrated by higher LAS and UAS scores compared to MyanBERTa.

Our contributions to Myanmar dependency parsing are multi-faceted and highly significant:

- We present the first systematic comparative analysis of XLM-RoBERTa-Large, XLM-RoBERTa-base-Longformer-4096, and MyanBERTa, providing crucial insights into their respective strengths and weaknesses within a low-resource context.
- Our study empirically demonstrates that a larger, robustly pre-trained multilingual model (XLM-RoBERTa-Large) can significantly outperform a smaller, language-specific model (MyanBERTa) for complex tasks like dependency parsing in low-resource settings, challenging conventional assumptions.
- We implement and optimize a parameter-efficient transformer-based model integrating adapter layers and a biaffine parsing mechanism. This model, trained with a sophisticated multi-task learning objective across six distinct prediction tasks (UPOS, XPOS, FEATS, LEMMA, HEAD, DEPREL), fosters balanced learning and robust shared representations for Myanmar's highly interdependent linguistic phenomena.
- Although XLM-RoBERTa-base-Longformer-4096 extends the input capacity for dependency parsing, our findings reveal its practical limitations. This outcome emphasizes that merely increasing sequence length does not inherently enhance performance; rather, core model capacity and pre-training quality remain paramount.

The remainder of this paper is structured as follows: Section 3 and provides background theory and details the model overview. The dataset preparation and the implementation details of our experiments are in section 4 and section 5. Section 6 presents the results and discussion of our findings, and section 7 concludes the paper.

2. Related Works

Major The evolution of dependency parsing can be traced through a progression of increasingly sophisticated methodologies, moving from statistical models to modern deep learning architectures. These statistical methods relied heavily on manual feature engineering to represent lexical, morphological, and positional information. The field saw a paradigm shift with the adoption of neural networks, which automated the process of feature extraction. Among these models, bidirectional long short-term memory (BiLSTM) encoders became popular for predicting head-dependent arcs and labels because they can learn rich contextual representations for each word. This eliminated the need for hand-crafted features and significantly improved performance. The current state-of-the-art methodology builds upon large-scale, pre-trained Transformer-based models as powerful encoders because these models provide highly expressive contextualized word embeddings that capture global sentence-level dependencies. Therefore, our proposed dependency parser model is built upon Roberta, pretrained-transformers with biaffine attention mechanisms. This architecture is the prevailing paradigm for achieving state-of-the-art results due to its ability to simultaneously model both arc and label predictions with high accuracy.

Li et al. [10] proposed a fully end-to-end sequence-to-sequence (seq2seq) dependency parser that directly predicts the relative head position for each word by avoiding reliance on traditional transition systems or graph-based parsing algorithms. This system is composed of BiLSTM encoder and LSTM decoder with attention by allowing for global sentence representation and context-sensitive decoding. To enrich input features, the model combines multiple embeddings: GloVe for semantic content, Node2Vec for syntactic structure, byte-pair encoded subwords, character-level features, and POS tags. A key innovation is the use of beam search with tree constraints to guarantee well-formed dependency trees while preserving decoding efficiency. Additionally, the introduction of sub-root decomposition mitigates long-distance dependency errors by segmenting complex sentences into shorter and more manageable substructures. The experimental results were demonstrated based on English PTB and Chinese benchmarks CTB dataset. In this result, the model achieves 94.11% and 88.78% UAS respectively. This method outperformed previous seq2seq methods and state-of-the-art traditional parsers.

Vietnamese NLP has progressed from traditional models like VnCoreNLP which used feature-based methods for POS tagging, NER, and dependency parsing to neural models leveraging pre-trained language representations. PhoBERT, a BERT-based model trained on Vietnamese data, achieved strong performance when fine-tuned separately for each task. However, maintaining separate models is inefficient for practical use. To address this, Nguyen et al. introduced PhoNLP [12] which is the first multi-task learning model and designed to jointly perform POS tagging, named entity recognition (NER), and dependency parsing for the Vietnamese language. This model performs all three tasks using a shared PhoBERT encoder and task-specific decoders by incorporating soft POS embeddings to improve cross-task performance. A key feature of the PhoNLP model is its hierarchical architecture where the POS tagging layer generates "soft" embeddings that are subsequently used as features for the NER and dependency parsing layers. This design allows the model to leverage POS tag information to improve the performance of the other two tasks.

This paper presents a Korean dependency parsing system using the Deep Biaffine Parser in combination with XLM-RoBERTa-Large embeddings. Korean sentences are tokenized using Stanza and refined with segmentation rules from the Sejong Project 2.0. Word representations are generated via XLM-RoBERTa and combined with POS information before being passed to a BiLSTM encoder.The parser uses biaffine attention mechanisms for both arc and label prediction. Two MLPs project the BiLSTM outputs into head and dependent spaces, enabling the model to compute arc scores and assign syntactic labels. Training and evaluation are conducted on the Sejong 2.0 corpus, a large-scale annotated dataset of Korean. The model achieves strong performance, with 94.17% UAS and 91.33% LAS on combined test sets, demonstrating its effectiveness in capturing Korean syntactic structures.

Hromei et al. [8] proposed an end-to-end approach to dependency parsing using auto-regressive Large Language Models (LLMs), specifically the LLaMA architecture. In contrast to traditional parsers with task-specific architectures, their method treats parsing as a sequence-to-sequence generation problem where the model outputs representation of the dependency tree from raw input sentences. To ensure training efficiency, they employed Q-LoRA for parameter-efficient fine-tuning on limited hardware resources. Experimental results on Italian Universal Dependencies treebanks have demonstrated to competitive performance related to state-of-the-art systems such as UDPipe and UDPipe+. Furthermore, the study showed that incorporating multilingual data during fine-tuning yielded greater improvements than scaling to larger models. Their findings highlight the viability of instruction-tuned LLMs for dependency parsing with promising implications for multilingual and low-resource scenarios.

3. Methodology

This section described about the comprehensive methodology for Myanmar dependency parsing by focusing on an optimized approach for linguistic and the unique characteristics of the low resource Myanmar language. The system implementation is based on a parameter-efficient transformer-based model by integrating adapter layers and a biaffine parsing mechanism to jointly predict various linguistic annotations.

3.1. *Architecture of Dependency Parsing Tree based Transformer Model*

The dependency parsing model is built upon a pre-trained RoBERTa transformer encoder, which serves as the core architecture for extracting rich contextualized representations to acquire effective syntactic parsing of Myanmar sentences. The transformer architecture produces context-aware token embeddings by leveraging multi-head self-attention and deep layer-wise encoding. This context-awareness is invaluable for dependency parsing, because it allows the model to accurately identify syntactic heads and correctly establish the complex dependency relationships between words in a sentence. The workflow of the proposed model for Myanmar dependency parsing tree is shown in Fig. (1)

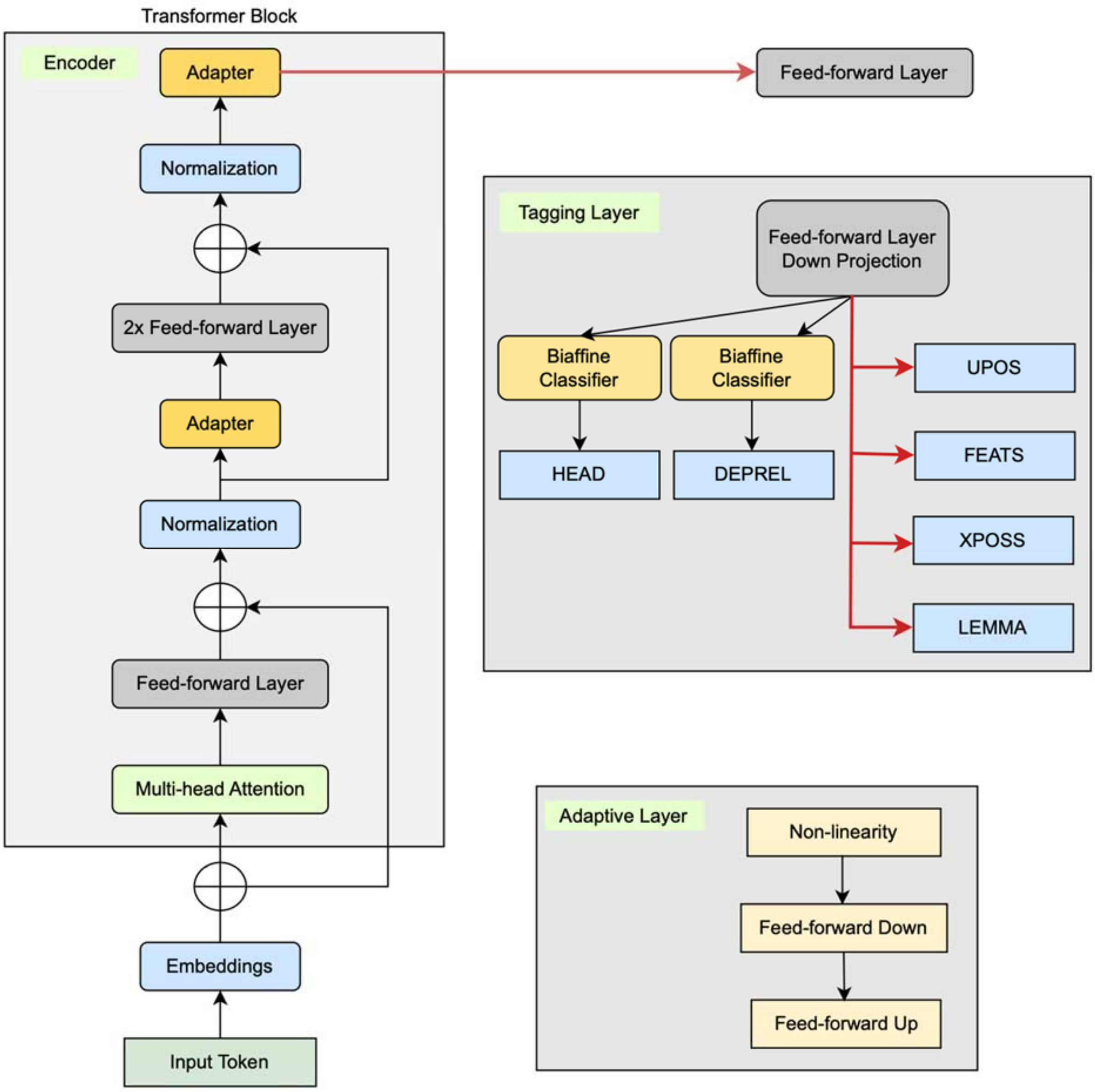

Fig. 1. Framework of Dependency Parsing Tree Based Pretrained-Transfomers

The encoder consists of 12 stacked Transformer blocks. Each input token is first mapped to a numerical representation via an embedding layer and then sequentially processed through the transformer layers. The Transformer Block is the foundational component of the Transformer neural network architecture. It primarily consists of Multi-Head Attention and a Feed-Forward Network, and layer normalization. Multi-Head Attention allows the model to weigh the importance of various parts of the input sequence in parallel to effectively capture long-range dependencies efficiently. Positional encodings are added to input embeddings to convey sequence order. Given the substantial computational cost of training large-scale Transformer models, lightweight adapter modules are incorporated into the RoBERTa-based architecture. These modules are specifically integrated after the Multi-Head Attention layer and the second Feed-Forward Network layer.

Each Adapter module is designed as a compact bottleneck feed-forward network. It comprises a Feed-Forward Down projection layer, a Nonlinearity activation function), and a Feed-Forward UP projection layer. This innovative design allows for effective fine-tuning of the model to the target language and specific parsing task without necessitating updates to the full set of parameters within the pre-trained transformer encoder. It can significantly reduce training overfitting and low computational cost because this selective fine-tuning strategy significantly reduces the number of trainable parameters. In addition, it benefits from the general language knowledge and robust representations already acquired by the pre-trained RoBERTa model without further pretraining.

The Tagging Layer module is a core component of the proposed model and designed to derive diverse linguistic annotations from the contextualized representations of input tokens. Following the encoder block, this layer processes the high-dimensional token embeddings to predict Universal Part-of-Speech (UPOS) tags, Treebank-specific Part-of-Speech (XPOSS) tags, morphological features (FEATS), and lemmas. For these sequence labeling tasks, the layer employs distinct linear projections by mapping the encoder's output directly to the respective tag vocabularies. Furthermore, the Tagging Layer incorporates a sophisticated biaffine classification mechanism that predicts head-dependent relationships (HEAD) and their corresponding dependency relations (DEPREL). This multi-headed design enables the simultaneous prediction of diverse linguistic phenomena and significantly enhances the model's overall linguistic understanding.

The proposed dependency parsing model is rigorously optimized using a multi-task learning objective and defined as the average of individual cross-entropy losses from six prediction tasks: UPOS, XPOSS, FEATS, LEMMA, HEAD, and DEPREL. This is represented by the Eq. (1).

$$L_{total} = \frac{1}{6}\left(L_{head} + L_{deprel} + L_{upos} + L_{xpos} + L_{feats} + L_{lemma\_scripts}\right) \quad (1)$$

The synonyms of these 6 parameters are described in Table 1:

Symbols	Descriptions
L_{head}	Loss from predicting incorrect syntactic heads; penalizes wrong tree structures.
L_{deprel}	Loss from mislabeling dependency relations between tokens and their heads.
L_{upos}	Loss from incorrect universal POS tag assignments.
L_{xpos}	Loss from incorrect language-specific POS tag predictions.
L_{feats}	Loss from errors in predicting morphological features (e.g., gender, case)
$L_{lemma\_scripts}$	Loss from incorrect lemmatization and script normalization.

Table 1. Descriptions of Parameters

This averaging-based optimization strategy promotes balanced learning across all syntactic and morphological tasks, preventing any single objective from dominating the training process. By jointly optimizing multiple linguistic annotations, the model encourages the transformer encoder and adapter layers to develop robust, shared representations that generalize well across tasks. This unified learning paradigm is especially beneficial for morphologically rich and syntactically flexible languages like Myanmar, where part-of-speech tags, morphological features, and syntactic dependencies are deeply interdependent.

4. Dataset Preprocessing

To ensure compatibility with transformer-based architectures, raw syntactic data that is structured in CoNLL-U format is subjected to a series of preprocessing steps designed to preserve token-level dependency structure during subword segmentation. Each sentence is parsed to extract lexical tokens, their corresponding head indices, and dependency relation labels.Subsequently, tokenization is performed using the RobertaTokenizerFast, which adheres to byte-level Byte-Pair Encoding (BPE) as employed in RoBERTa's pretraining regime. Given the possibility of a single token being split into multiple subword units, supervision signals for dependency label classification and head prediction are systematically aligned to the initial subword unit of each token.

In this framework, gold-standard syntactic head indices are recalibrated to reflect subword-aligned token spans. Dependency relation labels are encoded as integer class indices via a static label mapping dictionary. All preprocessed samples are serialized as structured dictionary objects containing four core components: input_ids, attention_mask, head_indices, and deprel_labels. Input sequences are padded to a fixed length to facilitate

e-ISSN : 0976-5166
p-ISSN : 2231-3850

efficient batched training across variable-length sentences. Attention masks are simultaneously computed to differentiate valid token positions from padding artifacts, ensuring proper contextualization during encoding.This strategy ensures that RoBERTa's byte-level encoding does not compromise token-level dependency supervision, thereby enabling accurate modeling of syntactic structure in neural parsing architectures.

In this experiment, dependency parsing based transformer-pretrained models are trained on Myanmar Treebank (PTB) that contains 30000 Myanmar sentences. Myanmar Dependency Tree Corpus dataset is a syntactically annotated Myanmar treebank formatted in CoNLL-U that is widely used for Universal Dependencies (UD) tasks such as dependency parsing. The dataset preparation process involves several key stages: initially, Myanmar-language textual data is collected and preprocessed through sentence segmentation and tokenization. Each token is then annotated with linguistic information including lemmas, universal (UPOS) and language-specific (XPOS) part-of-speech tags, morphological features (like number or tense), and syntactic dependencies indicating the head word and the type of dependency relation (e.g., nsubj, obj, root). In this training process, 20,000 sentences are used as the training dataset, while the remaining 1,0000 sentences are divided between the development set and the test set. Figure 2 shows the sample of Myanmar parsing tree sentences.

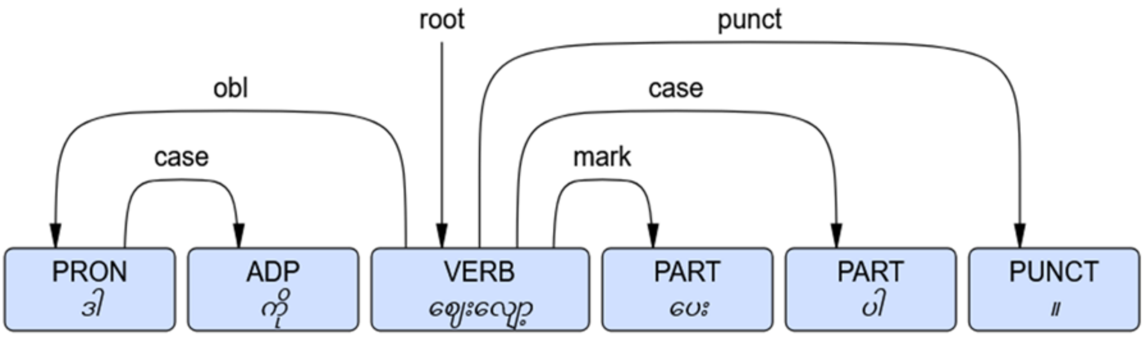

Fig. 2. Sample of Myanmar Annotated Parsing Tree

5. Implementation

A series of training experiments were undertaken to assess the effectiveness of transformer-based dependency parsing for the Myanmar language by utilizing both multilingual and Myanmar-specific pre-trained language model. To isolate the impact of pre-trained encoders in Myanmar dependency parsing, we conducted all experiments using a unified parsing setup by varying only the embedding backbone with different configurations. Our experiments primarily focused on Myanmar dependency parsing based RoBERTa model. The first dependency model was developed using XLM-RoBERTa-Large, a multilingual extension of RoBERTa trained on around 100 languages. This model is well-suited for capturing the syntactic complexity and unique grammatical structures of the Myanmar language because it provides robustness about cross-lingual representations. RoBERTa-base-driven dependency models are limited to processing sentences no longer than 512 tokens. To overcome this, the second model was built using XLM-RoBERTa-base-longformer-4096. It is a variant of Roberta and specifically designed for sequences up to 4096 tokens.

In addition to this RoBERTa-based models, the dependency model based on MyanBERTa was also undertaken to investigate how the distinct architectural characteristics of BERT-base and RoBERTa-base impact Myanmar dependency parsing. The parameters used in these experiments are shown in Table 2:

Models	**RoBERTa-Large**	**RoBERTa-Longformer**	**MyanBerta**
Epochs	15	60	80
Batch Size	8	8	16
Learning Rate	0.00005	0.00005	0.00005
Embedding	512	4096	512
Patience	3	3	3

Table 2. Comparison of Parameters

Training is conducted end-to-end using a combined cross-entropy loss for both tasks, optimized with AdamW. Key training parameters include a specific learning rate (often with a scheduler like linear warmup and decay), a defined batch size, and a maximum sequence length. The training runs for a set number of epochs with patience for early stopping. Finally, tree decoding algorithms are applied to the model's output to form valid dependency trees, with performance assessed by Labeled and Unlabeled Attachment Scores (LAS/UAS).

6. Implementation

This analysis provides a comprehensive evaluation of different dependency parsing models for the Myanmar language by focusing on Unlabeled Attachment Score (UAS) and Labeled Attachment Score (LAS). These metrics are crucial for assessing the accuracy of identifying syntactic heads and their corresponding dependency labels respectively. To evaluate the effectiveness of different pretrained transformer models for Myanmar dependency parsing, we compared the performance of RoBERTa-Large (512-token input), RoBERTa-Longformer (4096-token input), and MyanBERTa (512-token input) on both development and test dataset.

Models	Dataset	POS (%)	UAS (%)	LAS (%)
RoBERTa-Large	Dev	**98.253**	**95.734**	**93.629**
RoBERTa Longformer	Dev	96.787	92.561	89.712
MyanBERTa	Dev	94.560	85.558	81.360
RoBERTa -Large	Test	**98.182**	**95.820**	**93.520**
RoBERTa -Longformer	Test	96.474	92.069	89.105
MyanBERTa	Test	94.429	85.614	81.628

Table 3: Comparison of LAS and UAS scores

As shown in Table 3, RoBERTa-Large achieves the highest UAS of 95.73% and LAS of 93.63% on the development set, and 95.82% UAS and 93.52% LAS on the test set. In comparison, RoBERTa-Longformer yielded 92.56% UAS and 89.71% LAS on the development set, and 92.07% UAS and 89.11% LAS on the test set. MyanBERTa, despite being a Myanmar-specific model, achieved only 85.56% UAS and 81.36% LAS on the development set, and 85.61% UAS and 81.63% LAS on the test set. These results show that relative performance drop of approximately **3.3% UAS and 4.2% LAS** for RoBERTa-Longformer compared to RoBERTa-Large, and a **10.6% UAS and 13.1% LAS** drop for MyanBERTa compared to RoBERTa-Large on the test set.

Although RoBERTa-Longformer extends the input capacity to 4096 tokens, it did not achieve better performance than RoBERTa-Large in dependency parsing. This outcome implies that increasing the maximum input length through architectural changes alone does not necessarily enhance the model’s ability to capture syntactic dependencies. Even with a 512-token limit, RoBERTa-Large's optimized architecture and extensive pretraining on diverse datasets enable it to effectively encode the syntactic structures vital for high-quality dependency parsing.

While While MyanBERTa is pretrained specifically on Myanmar language data and utilizes a 512-token embedding, its performance in dependency parsing is inherently limited by the restricted scale and syntactic diversity of the available Myanmar corpora. Although it adopts the RoBERTa architecture, the relatively smaller number of layers, hidden dimensions, and attention heads constrains its ability to capture long-range dependencies and hierarchical syntactic structures. In contrast, RoBERTa-large with substantially more layers, hidden units, and attention heads benefits not only from higher model capacity but also from pretraining on massive multilingual corpora that expose it to a wide variety of syntactic patterns across languages. This cross-linguistic exposure enables RoBERTa-large to generalize more effectively for Myanmar language by transferring knowledge of structural patterns that may not be present in Myanmar-specific data. Due to the optimized pretraining methodology of RoBERTa-large such as dynamic masking and longer training schedules on diverse data, it enables to model complex morphological and syntactic dependencies. As a result, RoBERTa-large consistently demonstrates superior performance in constructing accurate dependency trees as demonstrated in higher LAS and UAS scores compared to MyanBERTa. These observations indicate that leveraging large-scale, multilingual pretraining can provide more substantial gains in Myanmar dependency parsing than focusing solely on language-specific corpora. Moreover, the model scale and pretraining quality play a critical role in capturing the intricate structures inherent in morphologically rich and syntactically flexible languages for Myanmar language.

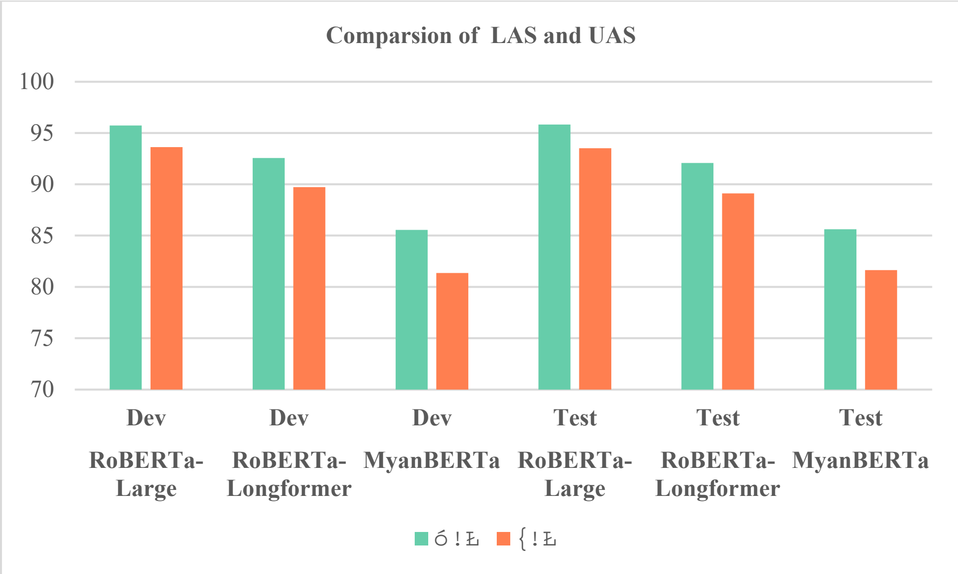

Fig. 3. Comparison of LAS and UAS Scores for Myanmar Dependency Parsing Based on Three Pretrained-. Transformers

7. Conclusion

This study presented a systematic evaluation of transformer-based dependency parsing for the Myanmar language based on three pretrained-transformers: XLM-RoBERTa-Large, XLM-RoBERTa-base-Longformer-4096, and MyanBERTa within a unified biaffine parsing and multi-task learning framework. Our results clearly showed that XLM-RoBERTa-Large achieved the highest Labeled Attachment Score (LAS) and Unlabeled Attachment Score (UAS) on both development and test datasets. Experimental results demonstrated that XLM-RoBERTa-Large consistently outperforms both the long-sequence variant and the language specific pretrained model. These findings indicate that model scale, robust architectural design, and high-quality multilingual pretraining contribute more significantly to parsing accuracy than extended input capacity or language-specific pretraining alone. Furthermore, the parameter-efficient integration of adapter layers with biaffine parsing enabled balanced optimization across syntactic and morphological tasks which are essential for morphologically rich and low-resource languages. Future research will focus on hybrid pretraining strategies that combine multilingual representations with targeted Myanmar-specific syntactic data as well as advanced sequence modeling techniques to further enhance the handling of complex sentence structures.

8. Conflict of Interest

The authors have no conflicts of interest to declare.

References

[1] Aye, H. T.; Pa, P. W. (2020): Dependency Head Annotation for Myanmar Dependency Treebank. Advances in Science, Technology and Engineering Systems Journal, 5 (6), 788-800.

[2] Beltagy, I.; Peters, M.E.;Cohan, A. (2020). Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

[3] Cui, D.; Bi, Y. (2024): Korean Dependency Parsing Using Deep Biaffine Dependency Parser. International Conference on Asian Language Processing (IALP), Hohhot, China, pp. 451-455, doi: 10.1109/IALP63756.2024.10661183.Berry M. W., Dumais S. T., O'Brien G. W. (1995): Using linear algebra for intelligent information retrieval, SIAM Review, 37, pp. 573-595.

[4] Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M., Zettlemoyer, L. and Stoyanov; V. (2019): Unsupervised cross-lingual representation learning on a scale. *arXiv preprint arXiv:1911.02116*.

[5] Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. (2019): BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*, 1 (2), pp. 4171-4186.

[6] Dozat, T.; Manning, C.D. (2016): Deep Biaffine Attention for Neural Dependency Parsing. *arXiv preprint arXiv:1611.01734*.

[7] Hlaing, A. M.; Pa, P. W. (2022): MyanBERTa: A Pre-trained Language Model for Myanmar. In 2022 International Conference on Communication and Computer Research (ICCR2022), Online and Seoul, Republic of Korea

[8] Hromei, C. D.; Croce, D.; Basili, R. (2023): End-to-end Dependency Parsing via Auto-regressive Large Language Models. In *CEUR WORKSHOP PROCEEDINGS*, Vol. 3596.

[9] Kiperwasser, E.; Goldberg, Y. (2016): Simple and Accurate Dependency Parsing Using Bidirectional LSTM Feature Representations. *Transactions of the Association for Computational Linguistics*, *4*, pp.313-327.

[10] Li, Z.; Cai, J.; He, S.; Zhao, H. (2018): Seq2seq dependency parsing. In Proc of the 27th International Conference on Computational Linguistics.

[11] Liu, Y.; Ott, M.; Goyal, N.; Du, J., Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. (2019): RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv e-prints, pp. arXiv-1907.
[12] Nguyen, L.T.; Nygyen, D.Q. (2021): A joint multi-task learning model for Vietnamese part-of-speech tagging, named entity recognition and dependency parsing. 2021 Conference of the North American Chapter of the Association for Computational Linguistics, pp. 1-7.
[13] Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; et al. (2020): Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Association for Computational Linguistics.

Authors Profile

Nwe Nwe Win is a PhD candidate at the University of Computer Studies, Yangon, Myanmar. She is an Assistant Director at the Department of Myanmar Nationalities' Languages under the Ministry of Education, Myanmar. She received B.Sc.(Maths) from Mandalay University in 1991, Post Graduate Diploma in Multimedia Art (PGDMA) from University of Education, Mandalay in 2000. Diploma in Computer Science.(D.C.Sc.) in 2003 from Computer University (Monywa) and Master of Information Science(M.I.Sc) in 2006 from Univeristy of Computer Studies, Mandalay

Win Pa Pa, is a Professor and researcher on Artificial Intelligence, and Language and Speech Processing. She received Ph.D(IT) from University of Computer Studies, Yangon(UCSY), Myanmar in 2009. She was a researcher at Natural Language Processing Lab (UCSY) from 2009 to 2024, She is a Dean of the Faculty of Computing at the Naypyitaw State Polytechnic University (NSPU), Myanmar since November 2025

Nang Kham Htwe,received the B.C.Sc. (Hons;), M.C.Sc. from University of Computer Studies, Mandalay in 2005 and 2009 and Ph.D (IT) from Univesity of Computer Studies, Yangon in 2024. She is currently working as Associate Professor at Polytechnic University, Bamaw in Myanmar. She can be contacted at email: nangkhamhtwe@ucsy.edu.mm.