CNER: Concept and Named Entity Recognition

Ivan Tchomba*
Finance Risk and Data Department
ENSAE
itchomba@ensae.fr

Abstract

This project investigates the relative contribution of the classifier head in the task of joint Concept and Named Entity Recognition (CNER). Using the dataset proposed by Babelscape and the DeBERTa-v3 encoder, we freeze the transformer backbone and focus solely on the effectiveness of different classifier architectures. We compare three approaches: a BiLSTM with a Conditional Random Field (CRF), a BiLSTM followed by a 1D-CNN and linear layer, and a simple feed-forward network (FFN). Our results show that despite the encoder being frozen, the BiLSTM+CRF model achieves surprisingly competitive performance, reaching up to 56% F1-score. In contrast, other architectures achieve notably lower performance, revealing the crucial role of the encoder in capturing contextual representations. Furthermore, we analyze the impact of class imbalance and label frequency on model performance. This study provides insights into the classifier head's limitations and strengths when decoupled from fine-tuned language models.

1 Introduction

Named Entity Recognition (NER) is a foundational task in Natural Language Processing (NLP), consisting of identifying and classifying spans of text that correspond to predefined entity types, such as persons, organizations, or locations. Concept recognition, on the other hand, focuses on identifying mentions of abstract or domain-specific concepts, often in specialized texts such as biomedical literature. While traditionally treated separately, the **CNER** task—*Concept and Named Entity Recognition*—seeks to jointly recognize both types of expressions, reflecting a richer and more realistic understanding of semantic spans in text.

In this project, we investigate the task of CNER using the dataset released by Babelscape [1], introduced in the paper *CNER: Concept and Named Entity Recognition* [2], accepted at NAACL 2024. The dataset provides annotations over multilingual corpora (we focus on the English portion) where tokens are labeled with both named entity and concept tags. This unified framework introduces new challenges for span detection, label imbalance, and semantic ambiguity.

Due to limited computational resources, particularly in terms of GPU availability, this work does not explore large-scale pretraining or architecture search. Instead, the objective is to better understand the relative contribution of the classification head to the overall performance, assuming a pretrained transformer encoder. Specifically, we compare two alternative classification heads applied on top of a frozen or partially fine-tuned transformer encoder:

- A **BiLSTM-CRF** head, inspired by classic sequence labeling pipelines.
- A **feedforward head** with two linear layers, a dropout rate of 0.1, GeLU activation, and layer normalization, as recommended in the original CNER paper.

*4th year student at ENASE

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

In both cases, we analyze how much predictive power can be extracted from the transformer representations alone, with minimal additional modeling. Our results suggest that, consistent with prior work, the bulk of predictive performance in CNER tasks comes from the pretrained encoder, while classification head choices play a secondary—yet still non-negligible—role.

While we do not attempt to outperform the state of the art reported in [2], we aim to shed light on architectural simplicity and trade-offs under resource constraints.

Table 1: Summary of tested classifier heads and results (DeBERTa encoder frozen).

Model	Classifier Head	Best F1 (%)	Epoch
BiLSTM + CRF	2-layer BiLSTM, LayerNorm, Dropout, CRF	**56**	10
BiLSTM + CNN + Linear	BiLSTM, Conv1D, LayerNorm, Linear	26	140
FFN	2 Linear Layers, GELU, LayerNorm	25	8

2 Related Work

2.1 Named Entity Recognition (NER)

Named Entity Recognition (NER) has been a cornerstone task in Natural Language Processing (NLP), aiming to identify and classify entities such as persons, organizations, and locations within text. Traditional approaches utilized rule-based systems and statistical models like Hidden Markov Models (HMMs) and Conditional Random Fields (CRFs). The advent of deep learning introduced neural architectures, notably BiLSTM-CRF models, which significantly improved performance by capturing contextual information.

The introduction of transformer-based models, particularly BERT [3], revolutionized NER by enabling deep contextual understanding. Fine-tuning pre-trained transformers on NER datasets has become a standard practice, yielding state-of-the-art results across various benchmarks.

2.2 Concept Recognition

Concept recognition extends beyond traditional NER by identifying abstract or domain-specific terms, often prevalent in specialized fields like biomedical or legal texts. Unlike named entities, concepts may not have clear boundaries or capitalization cues, making their detection more challenging. Approaches have included dictionary-based methods, machine learning classifiers, and more recently, neural models that leverage contextual embeddings to capture nuanced meanings.

2.3 Unified Concept and Named Entity Recognition (CNER)

The integration of concept and named entity recognition into a unified task, known as CNER, addresses the need for comprehensive semantic understanding. Martinelli et al. [2] introduced a framework that jointly models both entities and concepts, demonstrating that a unified approach outperforms separate models. Their work includes the release of the Babelscape CNER dataset, annotated with fine-grained labels across multiple languages.

Their proposed model fine-tunes a DeBERTa-v3-base transformer with a token classification head, achieving significant improvements in macro F1 scores compared to specialized systems. This highlights the effectiveness of leveraging pre-trained contextual embeddings for joint entity and concept recognition.

2.4 Advancements in NER Architectures

Recent studies have explored enhancements to NER architectures. For instance, SCANNER [4] introduces a knowledge-enhanced approach that integrates external information to improve recognition of unseen entities. Similarly, SUNER [5] employs contrastive learning to align span and entity type representations, facilitating better generalization across domains.

These advancements underscore the trend towards models that not only rely on contextual embeddings but also incorporate external knowledge and sophisticated training objectives to handle the complexities of entity and concept recognition.

3 Dataset

The experiments in this project are conducted on the **CNER dataset** released by Babelscape [1], which is introduced in the paper *CNER: Concept and Named Entity Recognition* [2]. The dataset contains annotations for both named entities and general concepts, unifying two previously distinct labeling tasks into a single tagging scheme. Annotations are provided in IOB2 format, covering a wide range of entity and concept types in multiple languages. This work focuses exclusively on the English portion of the dataset.

3.1 Overview and Statistics

The dataset is composed of tokenized sentences, each annotated with concept or named entity labels. Each token is associated with a BIO-tag and a corresponding semantic category (e.g., `B-ORG`, `I-Concept_Generic`, etc.). In total, the English subset contains approximately 8,782,688 tokens, within 319,590 sentences (data samples).

A summary of basic statistics is provided below:

- **Number of unique labels**: 59
- **Proportion of concepts vs. named entities**: Concepts account for approximately 20% of annotated spans, while named entities represent the remaining 18%.

3.2 Label Distribution and Class Imbalance

The label distribution is notably imbalanced (see Figure 1). A few classes, such as `Concept_Generic` and `ORG`, dominate the training set, while many fine-grained types have very few occurrences. This skew can negatively affect model performance, particularly on low-resource tags that occur rarely or are semantically ambiguous.

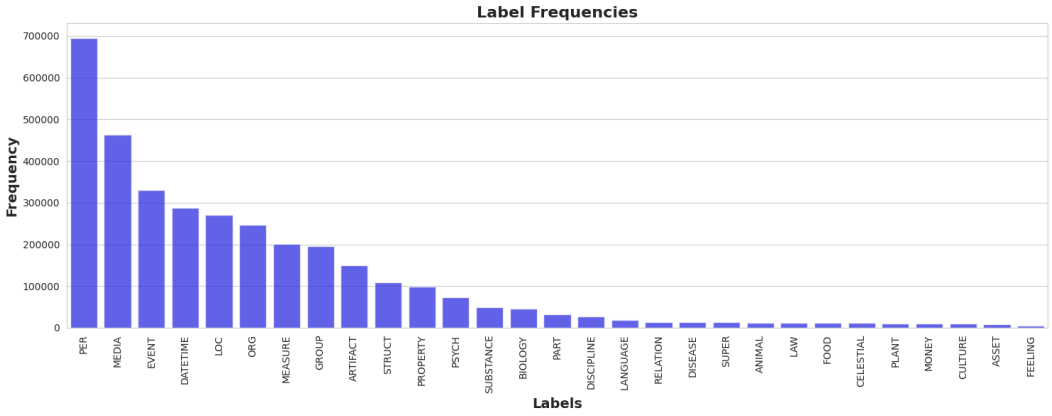

Figure 1: Distribution of token labels in the CNER dataset.

The dataset contains a wide range of concept and named entity types. Excluding the "O" class (non-entity tokens), the remaining annotated tokens are heavily imbalanced. Table 2 presents the relative proportion of each class in the training data.

As seen above, the distribution is highly skewed: the top three classes (PER, MEDIA, and EVENT) alone account for over 88% of the annotated tokens. In contrast, a large number of rare classes (e.g., `ASSET`, `FEELING`, `PLANT`, `LAW`) each represent less than 1% of the dataset. This strong imbalance presents a significant challenge for learning and generalization, particularly when the encoder is frozen and the classifier must rely solely on static embeddings.

Table 2: Label distribution excluding the "O" class.

Label	Proportion (%)	Label	Proportion (%)
PER	41.78	PSYCH	4.28
MEDIA	27.06	SUBSTANCE	2.86
EVENT	19.30	BIOLOGY	2.66
DATETIME	16.74	PART	1.86
LOC	15.76	DISCIPLINE	1.60
ORG	14.42	LANGUAGE	1.06
MEASURE	11.74	RELATION	0.84
GROUP	11.46	DISEASE	0.82
ARTIFACT	8.76	SUPER	0.80
STRUCT	6.36	LAW	0.70
PROPERTY	5.78	ANIMAL	0.70
		CELESTIAL	0.66
		FOOD	0.66
		PLANT	0.62
		MONEY	0.58
		CULTURE	0.58
		ASSET	0.44
		FEELING	0.32

3.3 Entity vs. Concept Repartition

We further examine the distribution between named entities and concepts (Figure 2). Named entities tend to be more frequent in news-like text (e.g., `PER`, `LOC`, `ORG`), whereas concepts (e.g., `BIOLOGIE`, `PSYCH`) are more prevalent in specialized or domain-specific contexts. This diversity poses additional challenges for generalization, particularly when fine-tuning on a general-domain encoder.

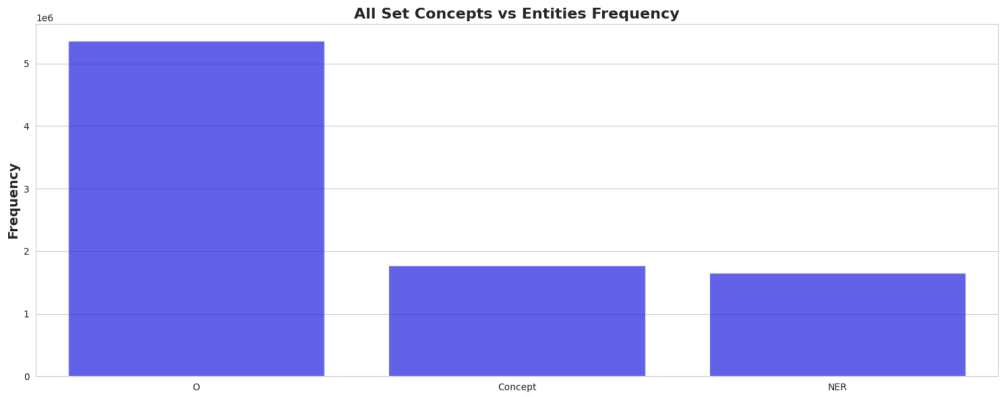

Figure 2: Repartition of named entities and concepts across the dataset.

3.4 Sentence Length and Annotation Density

Sentence lengths vary from very short utterances to longer, more complex structures. Figure 3 shows the distribution of sentence lengths. On average, longer sentences contain more annotated spans, although the relationship is not strictly linear. Higher density of labeled spans tends to increase model difficulty due to overlapping or ambiguous spans, particularly when using BIO tagging.

3.5 Implications for Modeling

The imbalance across labels and the coexistence of both broad and fine-grained categories imply that models must not only be robust to skewed distributions but also semantically precise. Furthermore, the token-level annotation scheme introduces challenges with nested or discontinuous spans, which

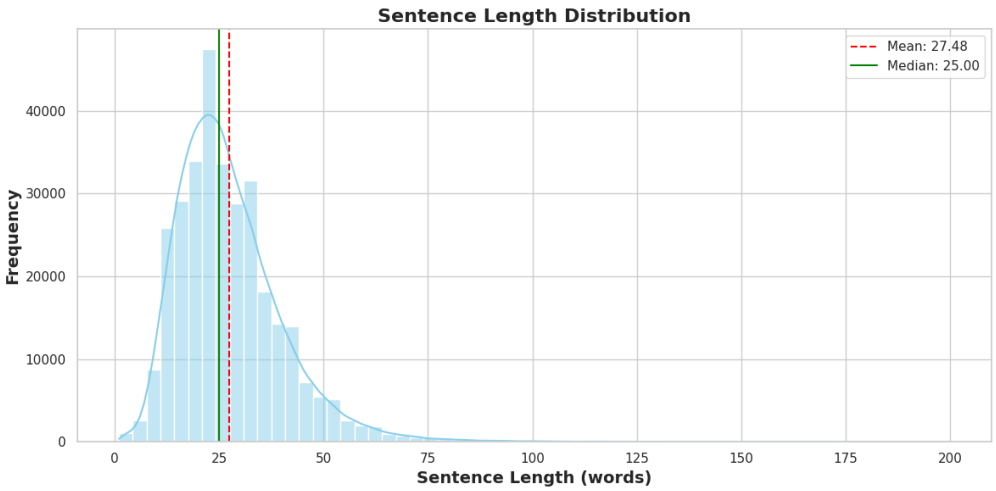

Figure 3: Distribution of sentence lengths in the dataset.

are not explicitly handled by standard BIO tagging. These aspects inform our modeling decisions and evaluation strategy, as discussed in later sections.

4 Methodology

4.1 Problem Framing and Hypothesis

The task addressed in this project is joint *Concept and Named Entity Recognition* (CNER), where the objective is to assign a semantic label to each token in a sentence. The dataset used is the Babelscape CNER corpus [1], and the goal is to explore the contribution of the classification head in a transformer-based architecture, under the constraint that the encoder is kept frozen.

In contrast to typical end-to-end fine-tuning approaches where the entire model is updated, our methodology deliberately isolates the classification layer's effect. This choice is motivated by practical constraints (limited GPU resources) and the scientific interest in understanding how much performance can be recovered or optimized when the encoder is fixed.

4.2 Overview of the Approach

Rather than fine-tuning the full transformer model, we precompute the contextualized embeddings using DeBERTa-v3 [6], a state-of-the-art language model known for strong performance in structured prediction tasks. These embeddings are computed once and stored, then reused as inputs to various classifier heads.

The central question is: *How much complexity is needed in the classifier head to achieve strong performance on CNER when the encoder is frozen?*

To explore this, we compare several architectures with increasing complexity:

1. **BiLSTM + CRF:** A strong traditional sequence labeling architecture where a bidirectional LSTM captures contextual dependencies and a Conditional Random Field (CRF) layer models structured output dependencies.
2. **BiLSTM + CNN + Linear:** A hybrid approach where CNN layers extract local n-gram patterns from LSTM outputs, followed by a fully connected classifier.
3. **Feedforward Network (FFN):** A simple two-layer dense network with GeLU activation, dropout, and normalization, as proposed in the original CNER paper [2].

4.3 Embedding Precomputation and Dataset Handling

To mitigate memory and time constraints, we implement a preprocessing step that computes and stores DeBERTa embeddings for all dataset splits. These embeddings are extracted using a frozen DeBERTa-v3 encoder and stored along with attention masks and token-level labels. The dataset is then wrapped in a custom PyTorch `Dataset` class to facilitate batch loading during training.

4.4 Training and Evaluation Protocol

Each classifier head is trained independently on top of the precomputed embeddings. We use the cross-entropy loss for token classification, and the CRF log-likelihood for the CRF-based model. Evaluation is conducted using the macro-averaged F1 score over all entity types, consistent with prior work [2]. Model selection is performed based on F1 score on the validation set.

Hyperparameters such as hidden dimension size, dropout rate (set to 0.1), and activation function (GeLU) follow recommendations from the CNER paper and the DeBERTa literature. Due to the nature of the task, all models incorporate masking using attention masks to avoid penalizing padded tokens during training.

4.5 Expected Outcomes and Evaluation Goals

By comparing the architectures under equal encoder conditions, we aim to quantify the added value of each classifier design. Our analysis will cover both the absolute performance (in terms of F1 score) and a qualitative evaluation of the predictions to understand where more complex heads (e.g., CRF) offer advantages over simpler feedforward classifiers.

5 Results and Discussion

5.1 Quantitative Comparison of Model Heads

Table 3 summarizes the macro F1-scores obtained for each classifier head trained on top of frozen DeBERTa-v3 embeddings. The BiLSTM+CRF model yields the best performance with an F1-score of 56%, significantly outperforming the other architectures.

Table 3: Model performance comparison (macro F1-score on the validation set).

Model	F1 Score (%)	Best Epoch	Comments
BiLSTM + CRF	56	10	Best performance
BiLSTM + CNN + Linear	26	140	Performance plateau at 26% 4
FFN (2-layer)	25	8	Simple, similar to CNN+LSTM

While the CRF-based model requires more complex decoding, it achieves higher performance with fewer epochs, showing more stable training. In contrast, the BiLSTM+CNN+Linear model plateaued at a lower F1-score despite a significantly longer training duration.

Overall, results suggest that when the encoder is frozen, the classifier head alone struggles to compensate. While the CRF structure adds valuable sequential modeling, it cannot match the representational power of fine-tuned transformer layers. This confirms that the main predictive power lies in the encoder and that, in constrained settings, models relying solely on static embeddings face sharp limitations.

Interestingly, despite its relative simplicity, the FFN slightly outperformed the BiLSTM+CNN+Linear model, possibly due to instability or overfitting in the latter. This suggests that not all added complexity leads to better generalization under fixed encoder conditions.

5.2 Impact of Freezing the Transformer Encoder

Our results confirm the crucial role that pretrained encoders play in sequence labeling tasks such as CNER. By freezing the DeBERTa-v3 backbone and limiting learning to the classifier head, we isolate

the downstream capacity of different architectures to leverage contextual embeddings. The sharp performance difference—where the BiLSTM+CRF model outperforms both the CNN-enhanced and FFN baselines—suggests that structured decoding (CRF) combined with temporal modeling (BiLSTM) remains a powerful tool even in the presence of static contextual representations.

However, all models exhibit significantly lower performance than reported by fully fine-tuned transformer systems on the same task. This reinforces the hypothesis that much of the semantic disambiguation and contextual generalization in CNER originates from fine-tuning the encoder itself. A frozen encoder acts as a bottleneck: it provides powerful but inflexible token embeddings, and without the ability to adapt these representations to the specific nuances of entity or concept types, even complex classifier heads struggle.

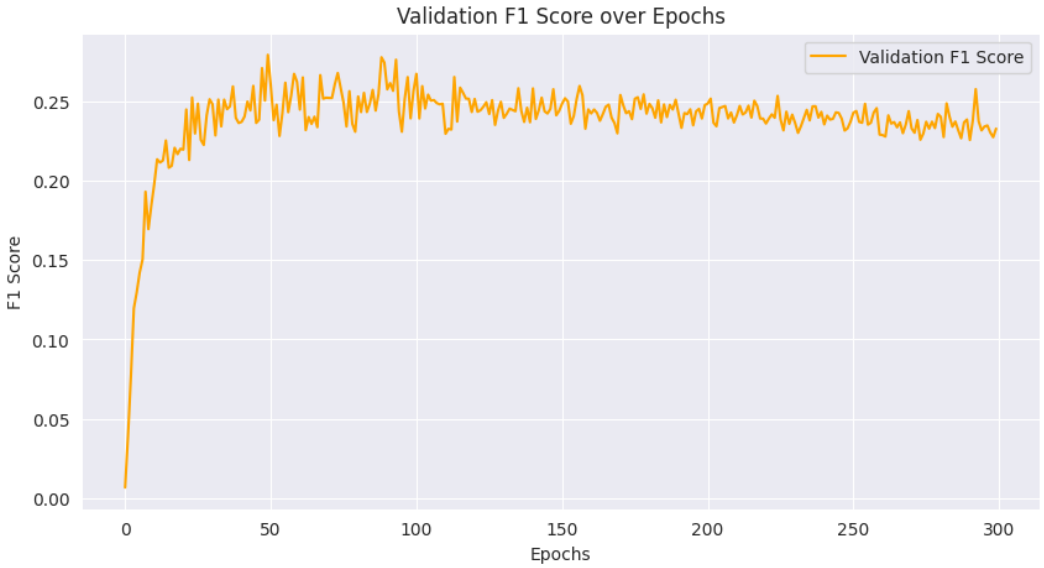

Figure 4: Training data for BiLSTM + CNN + Linear

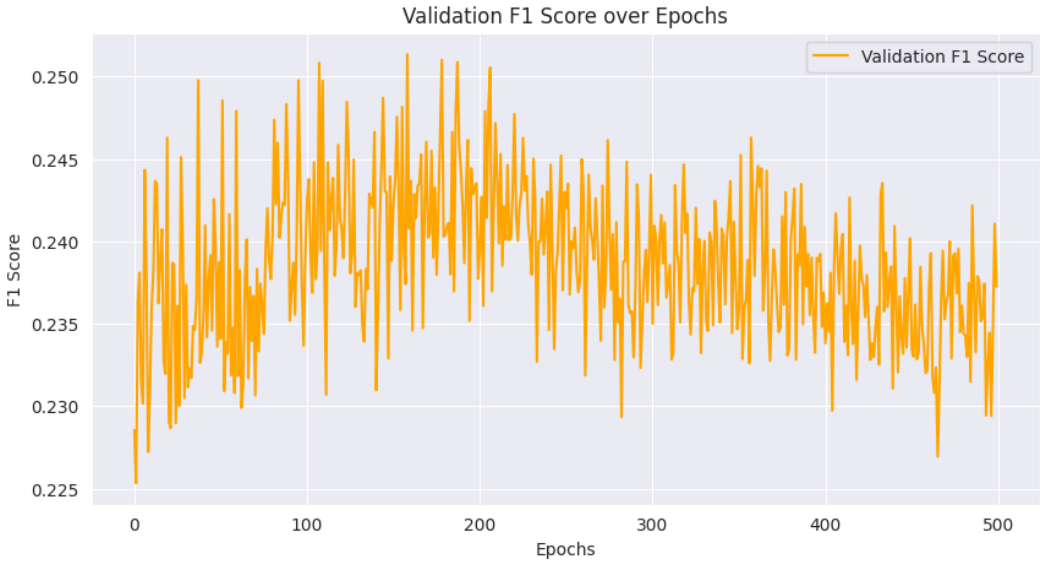

Figure 5: Training data for FFN (2-layers)

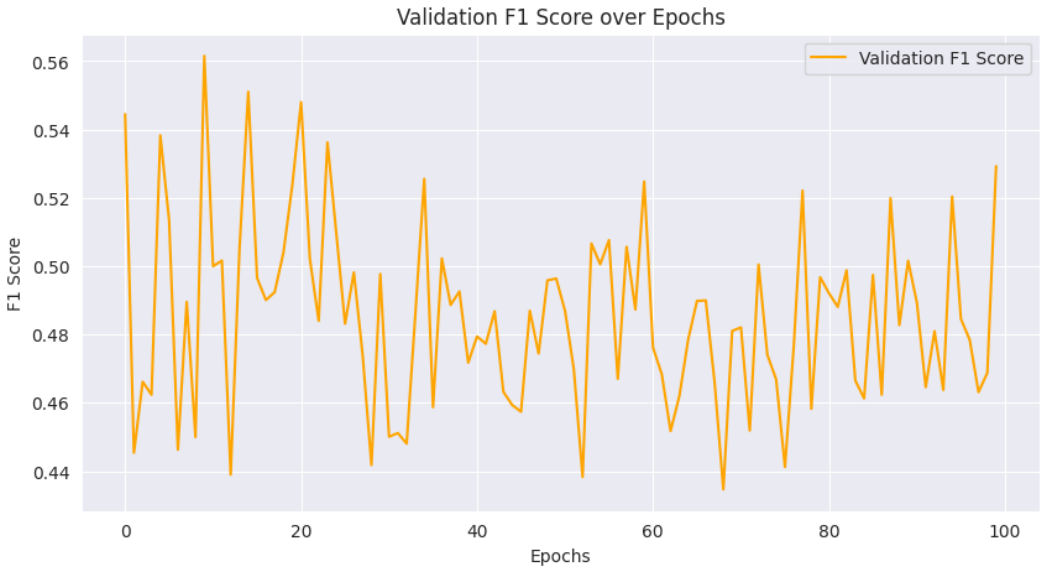

Figure 6: Training data for BiLSTM + CRF

5.3 Effectiveness of Classifier Heads

The performance gap between classifier heads is noteworthy. The BiLSTM+CRF architecture achieves an F1 score of 56%, while the FFN and CNN-augmented models stagnate around 25–26%. This disparity highlights that:

- **Temporal modeling (BiLSTM)** is still effective, even with static embeddings. It allows the classifier to reintroduce sequential dependencies discarded during encoder freezing.
- **Structured prediction (CRF)** enhances robustness by modeling label transitions, which is particularly helpful in BIO-tagged sequences.
- In contrast, **CNN layers and FFNs** may extract local patterns but lack the ability to model long-range dependencies or output label structure.

These findings imply that, in constrained environments where encoder fine-tuning is infeasible, it is more beneficial to invest in structured, sequential decoders than in width or depth of purely feedforward layers.

5.4 Label Imbalance and Generalization Gaps

Our per-class performance analysis reveals significant disparities, with rare labels consistently underperforming. This is unsurprising given the class distribution (cf. Table 2), where a handful of dominant types (e.g., `PER`, `MEDIA`) account for most annotations. Frozen encoders are unable to adapt to these low-frequency types, and classifier heads alone lack sufficient signal to learn discriminative features for underrepresented labels.

Attempts to mitigate this imbalance using weighted losses or sampling strategies may offer marginal improvements, but fundamentally, the problem stems from insufficient supervision for these rare categories. In a low-resource setting, auxiliary strategies such as data augmentation, knowledge injection, or contrastive representation learning could be more effective avenues for future work.

5.5 Loss Curves vs. Evaluation Metrics

A recurring observation during training is the poor alignment between training loss and evaluation performance. While loss steadily decreases, macro F1 often stagnates early—particularly for the FFN and CNN models (Figures 4, 5). This suggests that the classifier is improving in confidence on

already well-classified tokens (likely high-frequency labels) while continuing to misclassify rare or ambiguous types.

This disconnect underscores two key points:

- Cross-entropy loss, while commonly used, may be suboptimal in heavily imbalanced multi-class settings.
- Evaluation should focus on class-sensitive metrics (like macro F1) rather than relying on aggregate loss values.

Future iterations could benefit from loss functions that better reflect per-class performance, such as focal loss or margin-based objectives tailored to span-based prediction.

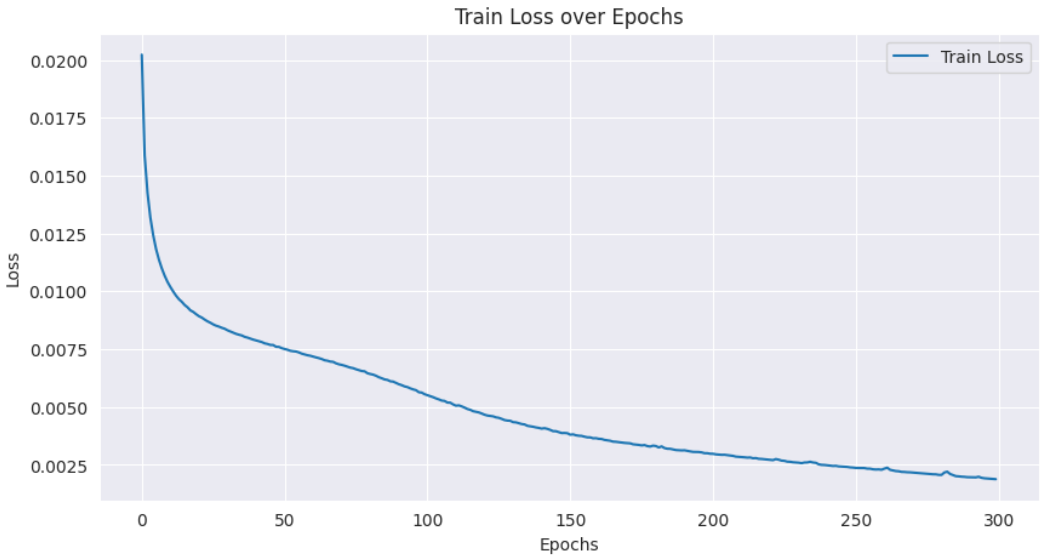

Figure 7: Loss evolution over training eg.: BiLSTM+CNN+Linear

5.6 Architectural Simplicity vs. Performance Tradeoffs

An important practical takeaway is that relatively simple architectures like BiLSTM+CRF still offer a strong baseline in resource-constrained environments. Despite their lower expressiveness compared to end-to-end transformer models, they can extract reasonable performance from frozen embeddings, especially when structured decoding is incorporated.

While our models fall short of state-of-the-art performance, they demonstrate that it is possible to achieve functional CNER systems with limited compute by focusing on smart architectural design rather than full-scale fine-tuning. This could be relevant for on-device or edge NLP applications where transformer fine-tuning is impractical.

6 Conclusion

In this study, we explored the predictive power of different classifier heads in the CNER task, isolating their performance by freezing the DeBERTa-v3 encoder. Our experiments demonstrate that when deprived of encoder fine-tuning, the classifier alone struggles to deliver strong results, especially in the face of the dataset's high class imbalance and long-tailed label distribution.

Among the models tested, the BiLSTM+CRF setup emerged as the most effective, attaining an F1-score of 56% after just a few epochs, while other models such as FFN or BiLSTM+CNN+Linear peaked around 25–26%. This confirms that even with static embeddings, structured prediction through CRFs still adds significant value.

Nonetheless, overall performance remains limited without end-to-end training of the encoder. The observed gap underlines the importance of contextual encoding and transfer learning in NER and concept recognition tasks. Future work could involve partial fine-tuning of encoder layers, class-balancing strategies, or contrastive pretraining for better generalization across rare classes.

Despite the constraints in compute resources, this project highlights the central role of the encoder in such tasks and offers a clear picture of how far classifier architectures can go on their own.

References

[1] Babelscape. Cner dataset. `https://huggingface.co/datasets/Babelscape/cner`, 2024.

[2] Giuliano Martinelli, Francesco Molfese, Simone Tedeschi, Alberte Fernández-Castro, and Roberto Navigli. Cner: Concept and named entity recognition. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 8336–8351, 2024.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186, 2019.

[4] Hyunjong Ok, Taeho Kil, Sukmin Seo, and Jaeho Lee. Scanner: Knowledge-enhanced approach for robust multi-modal named entity recognition of unseen entities. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 7725–7737, 2024.

[5] Hongli Mao, Xian-Ling Mao, Hanlin Tang, Yu-Ming Shang, Xiaoyan Gao, Ao-Jie Ma, and Heyan Huang. Span-based unified named entity recognition framework via contrastive learning. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*, pages 6406–6414, 2024.

[6] Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, 2023.