Improving Generalisation Capacities of GLiNER with ModernBERT

Léo Stepien
ENSAE
leo.stepien@ensae.fr

Abstract

Named Entity Recognition (NER) is a core task in Natural Language Processing, often tackled using Transformer-based models such as BERT. GLiNER is a recent framework that reframes NER as a span-to-entity matching problem and achieves strong generalization to unseen entity types by leveraging pretrained encoders. In this project, we hypothesize that replacing GLiNER's BERT backbone with ModernBERT—a recently released Transformer architecture optimized for long contexts and large-scale pretraining—can further improve generalization in zero-shot settings. Our approach uses ModernBERT's pretrained weights while preserving GLiNER's core architecture. Although we were unable to complete full-scale training or evaluation, our work highlights the practical barriers involved in adapting modern Transformer variants to span-based entity recognition systems and offers a foundation for future improvements.

1 Introduction

Named Entity Recognition (NER) is a cornerstone task in Natural Language Processing (NLP), essential for structuring unstructured text by identifying and classifying spans of text into predefined semantic categories such as persons, organizations, locations, or dates. For instance, in the sentence *"Apple acquired Beats in 2014"*, the words *"Apple"* and *"Beats"* can be recognized as organizations, and *"2014"* as a temporal expression. NER has wide-ranging applications, including information retrieval, question answering, content recommendation, biomedical text mining, and knowledge graph construction. Over the years, various NER-specific models have been developed, such as Conditional Random Fields (CRFs), BiLSTM-CRF architectures, and more recently, span-based approaches [4, 8]. However, the advent of the Transformer architecture has dramatically shifted the landscape by enabling models to capture long-range dependencies and contextual nuance through self-attention mechanisms. This innovation culminated in the introduction of BERT (Bidirectional Encoder Representations from Transformers) [2], which demonstrated that large-scale pretraining on raw text, without explicit supervision, can yield general-purpose language representations that transfer effectively to a wide variety of downstream tasks.

BERT's core innovation lies in its use of bidirectional masked language modeling: instead of predicting the next word in a sequence (as in traditional left-to-right models), BERT masks random tokens in the input and learns to predict them using both left and right contexts. This dual context awareness allows BERT to generate richer and more semantically grounded representations. Unsurprisingly, this architecture has proven valuable for NER, where understanding the full context of an entity is crucial. GLiNER (Generalist Language-independent Named Entity Recognizer) [5] builds on this idea by employing a pretrained BERT encoder to produce contextualized token representations, which are then used to match against candidate entity descriptions, enabling a more flexible and scalable approach to entity recognition across domains and languages.

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

One persistent challenge in NER is generalization to unseen entities or domains—a scenario where traditional models tend to falter. However, because BERT is trained on massive corpora, it often encounters similar structures or semantically related entities even if the exact form of an entity was not present in the training data for the NER task. This means that an unknown entity might still occupy a semantically appropriate region in the representation space, making it possible to associate it with known categories through contextual similarity. In this way, BERT facilitates a form of implicit generalization by anchoring new inputs to familiar patterns. The effectiveness of this mechanism hinges on the quality of the learned representations.

Recently, a new generation of Transformer-based models has pushed the boundaries of generalization and efficiency in NLP. Among these, **ModernBERT** [9] stands out as a dedicated architecture designed to overcome the limitations of earlier BERT variants. Unlike RoBERTa [6], DeBERTa [3], or Longformer [1], which each introduce incremental improvements to BERT—such as better pretraining objectives, enhanced positional encodings, or extended context windows—ModernBERT integrates a suite of architectural optimizations aimed at improving memory efficiency, inference speed, and scalability to extremely long sequences (up to 8192 tokens). Trained on a massive corpus exceeding two trillion tokens, including diverse web and code data, ModernBERT achieves state-of-the-art results across various NLP benchmarks, often outperforming its predecessors in both classification and retrieval tasks. Its ability to encode longer contexts with greater computational efficiency makes it a strong candidate for improving the robustness of downstream models, including those used in Named Entity Recognition.

These advances naturally lead to the hypothesis that integrating ModernBERT into GLiNER may enhance its ability to generalize to Out-of-Distribution (OOD) entities by providing more expressive and semantically rich token representations.

2 Background

2.1 BERT and Transformer Encoders

The Transformer architecture is built around a self-attention mechanism that enables the model to capture dependencies between tokens, regardless of their position in the sequence. Each input token is embedded and then linearly projected into three vectors: a query (Q), a key (K), and a value (V), using learned parameter matrices W_Q, W_K, and W_V. For a given token representation x, these projections are computed as:

$$Q = xW_Q, \quad K = xW_K, \quad V = xW_V \tag{1}$$

The attention scores between a token and all other tokens in the sequence are calculated using scaled dot-product attention:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{2}$$

where d_k is the dimensionality of the key vectors. This mechanism allows the model to assign dynamic importance weights to surrounding tokens, enabling it to contextualize a token's meaning based on its full sentence or document context. Multiple attention heads are used in parallel to capture diverse linguistic phenomena.

BERT leverages this architecture with two major innovations in its pretraining: bidirectional masked language modeling (MLM) and next sentence prediction (NSP). In MLM, random tokens in the input are masked and the model learns to predict them using both left and right context. This bidirectional context modeling allows for richer and more semantically coherent token embeddings. NSP further encourages the model to understand inter-sentence relationships by training it to predict whether two given sentences follow each other in the original text. Together, these objectives help BERT develop a robust understanding of both local and global context.

The combination of self-attention, bidirectional encoding, and unsupervised pretraining leads to contextualized token representations that generalize effectively across NLP tasks. This foundation is particularly valuable for applications like Named Entity Recognition, where understanding the semantic role of a token depends heavily on its surrounding context.

2.2 Named Entity Recognition and GLiNER

NER is a foundational task in NLP that involves identifying spans of text corresponding to semantic categories such as persons, organizations, locations, and more. Traditional approaches to NER typically rely on supervised sequence labeling models, such as CRFs or BiLSTMs, trained on domain-specific data and outputting IOB tags for tokens. While effective in-distribution, these models often lack the flexibility and robustness required for generalization to out-of-distribution (OOD) entities or domains.

GLiNER [5] proposes a unified, generalist framework for NER that significantly improves cross-domain and zero-shot performance by reframing the task as a span-to-entity matching problem. At its core, GLiNER leverages a pretrained bidirectional encoder such as BERT or RoBERTa to obtain contextualized token representations. These token embeddings form the foundation for a **span representation module**, which computes embeddings for all valid candidate spans in a sentence. For a span that begins at position i and ends at j, its embedding $S_{ij} \in \mathbb{R}^D$ is given by:

$$S_{ij} = \text{FFN}([h_i \,\|\, h_j]) \tag{3}$$

where h_i and h_j are the contextual embeddings of the start and end tokens, $\|$ denotes concatenation, and FFN is a two-layer feedforward network. To maintain tractability and computational efficiency, only spans up to a maximum length $K = 12$ are considered, which empirically preserves high recall while enabling linear-time span enumeration.

Simultaneously, an **entity representation module** maps entity types into the same latent space as span embeddings. Each entity type embedding q_t is obtained by refining its initial input vector p_t (e.g., the average of the label's token embeddings) through a feedforward transformation:

$$q_t = \text{FFN}(p_t) \tag{4}$$

The model computes a **matching score** between each span embedding S_{ij} and each entity embedding q_t as:

$$\phi(i, j, t) = \sigma(S_{ij}^\top q_t) \tag{5}$$

where σ is the sigmoid function, and $\phi(i, j, t)$ is interpreted as the probability that the span (i, j) expresses the entity type t. This formulation allows the model to treat NER as a **binary classification task over span-type pairs**, rather than a traditional multiclass tagging problem. During training, the objective is to maximize $\phi(i, j, t)$ for gold-labeled (positive) span-type pairs and minimize it for all others (negative pairs) using binary cross-entropy loss.

One of the key strengths of GLiNER lies in its ability to generalize to unseen types and domains. It achieves this by explicitly learning a shared latent space for both entities and spans, which facilitates effective **zero-shot inference**. The model has been evaluated across diverse settings, including two OOD benchmarks as well as multilingual zero-shot evaluations. It consistently outperforms prior models, especially in challenging zero-shot settings.

2.3 ModernBERT Enhancements

ModernBERT [9] introduces a series of innovations over the original BERT architecture, targeting both scalability and representational robustness. These enhancements are motivated by the goal of producing a general-purpose encoder that is efficient, highly contextual, and capable of operating in memory-constrained environments without compromising performance.

Training Improvements. One of the most impactful modifications in ModernBERT is the scale of its pretraining data. The model is trained on a massive and diverse corpus totaling over 2 trillion tokens, drawn from a wide range of domains and genres. This breadth of linguistic exposure enables ModernBERT to develop more generalizable and semantically rich representations than earlier models like BERT or RoBERTa, especially when deployed in OOD scenarios.

Architectural and Computational Enhancements. ModernBERT also introduces key architectural refinements and regularization techniques that permit efficient processing of long input sequences—up to 8192 tokens—without sacrificing speed or memory efficiency. These optimizations are designed to make training and inference feasible on commonly available GPU hardware, in contrast to many long-context models that require specialized accelerators. This makes ModernBERT not only more powerful but also more accessible for downstream tasks.

Implications for GLiNER. The benefits of ModernBERT directly translate into advantages for GLiNER. First, the improved contextual encoding capabilities and massive training corpus result in higher-quality token and span representations, which are crucial for accurately matching entities to spans in a zero-shot or cross-domain setting. Second, the enhanced generalization potential offered by ModernBERT's training data aligns well with GLiNER's objective of being robust to unseen entity types and domains. Finally, the model's ability to process longer contexts efficiently and run on standard GPUs improves the practicality of deploying GLiNER in real-world systems, enabling broader adoption in resource-constrained environments.

Overall, replacing GLiNER's backbone with ModernBERT is a promising direction to further push the boundaries of generalizable and efficient named entity recognition.

3 Methodology and Experimental Setup

Our central hypothesis is that replacing the BERT encoder in GLiNER with ModernBERT will lead to improved performance on OOD named entity recognition tasks. This is motivated by the observation that ModernBERT produces more robust and semantically rich representations, thanks to its large-scale pretraining and architectural improvements. These characteristics make it especially well-suited for zero-shot scenarios where entity types at inference time may not have been seen during training. By leveraging ModernBERT's enhanced generalization capabilities, we aim to boost GLiNER's ability to correctly identify novel entities across diverse domains, without modifying the core architecture of the model.

3.1 Dataset Overview

Our experimental setup relies on two datasets: **PileNER** and **CrossNER** from [7], chosen to evaluate the generalization capability of GLiNER when equipped with a ModernBERT backbone. PileNER is a large-scale GPT-generated NER dataset created using type-based data construction prompts. It offers wide coverage and serves as the training corpus in our experiments. In contrast, CrossNER is designed specifically for evaluating OOD NER performance. It contains highly domain-specific entities and is thus well-suited for assessing zero-shot generalization.

To understand the rationale behind selecting CrossNER for OOD evaluation, we examine its key differences from PileNER. CrossNER is divided into five distinct topical domains: *AI*, *Literature*, *Politics*, *Music*, and *Science*. This topical segmentation introduces domain-specific vocabulary and entity types not typically found in generic NER datasets. Additionally, as shown in Figure 1, CrossNER examples tend to be significantly shorter in sentence length compared to those in PileNER, introducing a structural form of domain shift that further challenges the model's robustness.

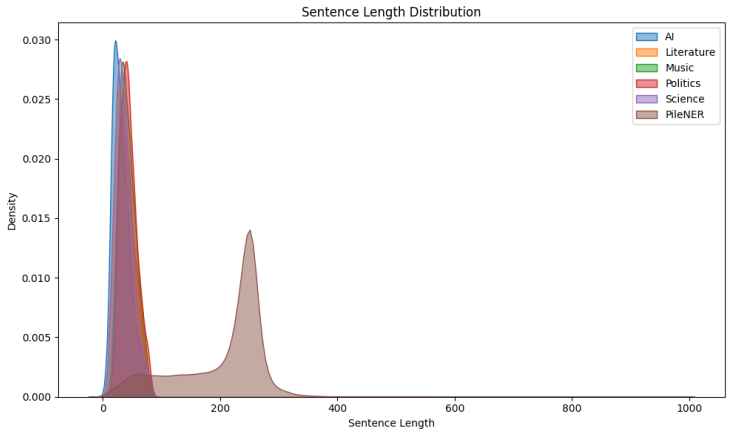

Figure 1: Distribution of Sentence Length for each dataset

A more critical factor motivating our zero-shot evaluation is the low overlap in named entities between the two datasets. Figure 2 quantifies this by showing the percentage of entities in each CrossNER

subset that also appear in PileNER. The overlap never exceeds 8%, highlighting the novelty of CrossNER's entity types relative to the training distribution. Overall, PileNER contains 495,056 annotated entities, of which only 1,467 are found anywhere in CrossNER. This minimal intersection justifies our use of CrossNER as a genuine zero-shot benchmark for evaluating the generalization capability of the GLiNER+ModernBERT model.

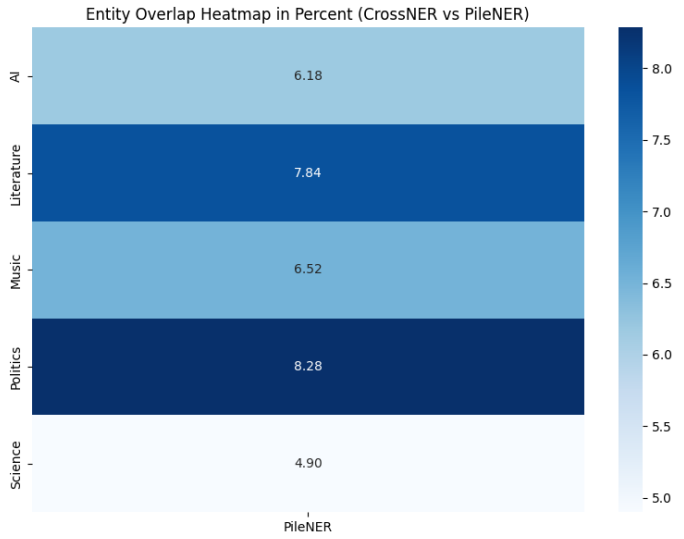

Figure 2: Proportion of entities overlap between CrossNER and PileNER

3.2 Model and Training Details

Our experimental setup consists of replacing the standard BERT encoder in GLiNER with ModernBERT (base version: 768 hidden dimensions, 22 layers, 12 attention heads). The ModernBERT encoder is initialized using pre-trained weights from `answerdotai/ModernBERT-base` and configured with optimized hyperparameters, including an extended maximum sequence length of 8,192 tokens and specialized embedding parameters tailored for long-context processing.

For training, we employ the PileNER dataset. The model is optimized using AdamW with separate learning rates: 5×10^{-5} for the encoder and 1×10^{-4} for the remaining components. A weight decay of 0.01 is applied, along with focal loss (parameters: $\gamma = 2.0$, $\alpha = 1.0$) to better handle class imbalance. The model is trained for 3 epochs with a batch size of 4, leveraging gradient accumulation and mixed-precision training to efficiently manage GPU memory usage.

During training, the ModernBERT encoder is kept frozen to preserve its general-purpose linguistic representations, while the span detection and entity classification layers are briefly fine-tuned to adapt to the specific NER task.

Following training, the model is evaluated on the CrossNER dataset to assess its zero-shot performance on out-of-distribution (OOD) entity types. Results will be compared against those reported in [5], which tested GLiNER with various transformer backbones. If time permits, we also plan to replicate training using other encoders to provide a more comprehensive comparison.

4 Results

Despite our efforts, we encountered a series of technical obstacles that ultimately prevented us from successfully training the GLiNER model with a ModernBERT backbone. Our implementation was based on the original code provided by [5], but several issues arose during model initialization and training.

Initially, we faced incompatibilities between GLiNER's model initialization logic and Hugging Face's model loading mechanisms, particularly related to caching and internal handling within the

transformers library. These discrepancies made it difficult to load pre-trained BERT-based models directly into the GLiNER framework. We mitigated this by manually loading and aligning the model parameters to ensure compatibility with GLiNER's architecture.

However, further complications emerged during training. Although the training pipeline appeared to function correctly at first, we eventually encountered persistent shape mismatches between the model's outputs and target labels. These inconsistencies occurred despite following the original data preprocessing steps outlined in the public implementation. We suspect that these errors stem from subtle mismatches introduced by the integration of ModernBERT's extended sequence handling and layer configuration, which diverge from the original GLiNER assumptions.

We were unable to proceed with a full experimental evaluation.

5 Conclusion

This work set out to explore the potential of improving zero-shot NER performance by integrating ModernBERT into the GLiNER framework. ModernBERT's superior contextual encoding capabilities, extended sequence handling, and extensive pretraining make it an attractive candidate for enhancing span-based NER systems like GLiNER.

However, our efforts revealed a number of technical roadblocks that limited our ability to train and evaluate the combined model. Key difficulties arose in model initialization due to incompatibilities between the Hugging Face Transformers library and GLiNER's parameter loading routines. Even after circumventing these issues through manual weight alignment, we encountered persistent shape mismatches during training that we were unable to resolve. These problems appear to stem from assumptions embedded in GLiNER's architecture that do not hold when using an alternative encoder with modified dimensions and sequence handling strategies.

Despite these setbacks, our investigation underscores the need for more modular and encoder-agnostic NER frameworks that can seamlessly leverage recent advances in Transformer modeling. Future work could involve deeper architectural adaptation, better debugging tools for span-based models, and systematic benchmarking of GLiNER with a wider range of encoders. Our partial integration effort offers a valuable case study in the practical challenges of combining state-of-the-art language models with task-specific architectures.

References

[1] Iz Beltagy, Matthew E Peters, and Arman Cohan. "Longformer: The Long-Document Transformer". In: *arXiv preprint arXiv:2004.05150* (2020).

[2] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[3] Pengcheng He et al. "DeBERTa: Decoding-enhanced BERT with Disentangled Attention". In: *arXiv preprint arXiv:2006.03654* (2020).

[4] Guillaume Lample et al. "Neural Architectures for Named Entity Recognition". In: *Proceedings of NAACL*. 2016.

[5] Yujie Li et al. "GLiNER: Generalist Entity Recognizer with Multitude of Entity Types". In: *arXiv preprint arXiv:2311.08526* (2023). URL: https://arxiv.org/abs/2311.08526.

[6] Yinhan Liu et al. "RoBERTa: A Robustly Optimized BERT Pretraining Approach". In: *arXiv preprint arXiv:1907.11692* (2019).

[7] Zihan Liu et al. "CrossNER: Evaluating Cross-Domain Named Entity Recognition". In: (2020). arXiv: 2012.04373 [cs.CL].

[8] Xuezhe Ma and Eduard Hovy. "End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF". In: *Proceedings of ACL*. 2016.

[9] Ashish Sabharwal et al. "Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference". In: *arXiv preprint arXiv:2412.13663* (2024). URL: https://arxiv.org/abs/2412.13663.