Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 47 No. 03 (2026)

Enhancing Named Entity Recognition through Advanced Sequence Labeling using Conditional Random Fields with Contextual Feature Integration

Simran Patil[1*], Dr. Sunny Sall[2]

[1*]Student, Department of Computer Engineering, St. John College of Engineering & Management (SJCEM), Palghar, Mumbai-401404, India.

[2]Assistant Professor, Department of Computer Engineering, St. John College of Engineering & Management (SJCEM), Palghar, Mumbai-401404, India.

Abstract:- Objectives: This study aims to improve the accuracy and consistency of Named Entity Recognition (NER) systems by addressing the limitations of transformer-based models that rely on independent token-level predictions. It focuses on exploring how structured prediction techniques can enhance sequence labeling performance. **Methods:** A comparative evaluation is carried out using three hybrid architectures: BERT combined with CRF, BiLSTM combined with CRF, and BERT integrated with a Transformer layer. All models are tested within a consistent experimental setup using the CoNLL-2003 dataset. The methodology emphasizes combining contextual representations with sequence-level optimization through Conditional Random Fields. **Findings:** The results indicate that the BERT + CRF model delivers the highest performance, achieving an F1-score of 0.94. In comparison, BERT + Transformer attains 0.81, while BiLSTM + CRF reaches 0.72. These outcomes demonstrate that incorporating CRF significantly improves label sequence coherence and overall extraction effectiveness. **Novelty:** The work presents a unified and systematic comparison of hybrid NER models, highlighting the importance of structured decoding in modern architectures. It offers valuable insights into balancing model performance and computational efficiency, supporting the development of reliable and scalable NER solutions for practical applications. The other NER models often fail to classify multi-token organization entities and entities with ambiguous contextual meaning correctly. The BERT + CRF model got better boundary detection by structured sequence optimization.

Keywords: Named Entity Recognition, Natural Language Processing, Conditional Random Fields, Deep Learning, Contextual Embedding, Sequence Labeling, Information Extraction, Intelligent Systems.

1. Introduction

Named Entity Recognition (NER) is an essential task in natural language processing (NLP) that focuses on identifying and classifying entities such as persons, organizations, and domain-specific terms within text. Early approaches relied on statistical models like Hidden Markov Models and Conditional Random Fields (CRFs), where CRFs effectively captured label dependencies but depended heavily on handcrafted features.

With the advancement of deep learning, transformer-based models such as BERT have significantly improved NER by learning contextual representations. However, many transformer-based methods perform token-level classification, which can lead to inconsistent label sequences. To address this, hybrid models combining contextual embeddings with CRFs have been proposed, improving sequential correlation.

Earlier the BiLSTM–CRF architectures also demonstrated strong sequence labeling performance by combining bidirectional contextual learning with structured prediction.

Research in domain-specific Named Entity Recognition (NER) has significantly increased in fields such as biomedical, clinical, and multilingual applications. Sun et al. [1] proposed a CRF-based framework integrated with temporal convolutional networks to improve efficiency and accuracy in biomedical text processing. Comprehensive surveys by Nasar et al. [4], and Li et al. [5] reviewed the evolution of NER methods and highlighted ongoing challenges such as model interpretability and robustness. Applications in geoscience and e-commerce were explored by Qiu et al. [6] and Chen et al. [7], respectively, showcasing the effectiveness of domain-tuned models while noting their limited generalization.

The addition of semantic representations also improved the NER performance. Devlin et al. [10] introduced BERT, a transformer-based model that takes in bidirectional context from big corpora. To address the constraints of token-level tagging, various alternative formulations have been investigated. He et al. [16] presented prompt-based approaches for few-shot Named Entity Recognition (NER) independent of pre-established templates. Lou et al. [19] incorporated graph-based knowledge representations for biomedical named entity recognition (NER).

Table 1 Comparative analysis of recent NER approaches based on contextual learning, structured prediction, computational efficiency, and practical applicability.

Metric	**CRF-based NER [1]**	**Deep Learning NER [2][3]**	**BERT-based NER [7][10]**	**Advanced NER Methods [11][12][15]**	**BERT + CRF (Present Research)**	**BERT + Transformer (Present Research)**
Contextual Learning	Low	Moderate	High	Very High	High	Very High
Sequence Optimization	Excellent	Good	Moderate	Good	Excellent	Good
Semantic Understanding	Low	Moderate	High	Very High	High	Very High
Long-Range Dependency	Low	Moderate	High	Excellent	High	Excellent
Structured Prediction	Excellent	High	Moderate	Good	Excellent	Good
Domain Adaptability	Moderate	High	High	Excellent	High	High
Computational Efficiency	High	Moderate	Moderate	Low	Moderate	Low
Nested Entity Recognition	No	Partial	Partial	Yes	Partial	Partial

Practical Suitability	Moderate	High	High	Moderate	Excellent	High

According to Table 1, existing studies demonstrate that combining contextual encoders with CRF-based structured decoding yields strong NER performance. However, most prior work focuses on individual architectures or domain-specific settings, with limited systematic comparison under unified experimental conditions. Furthermore, there is insufficient analysis of performance–efficiency trade-offs across hybrid architectures.

Despite these advances, many recent approaches require complex architectures or high computational resources. In contrast, CRF-based structured decoding remains efficient and interpretable. This study evaluates BERT + CRF, BiLSTM + CRF, and BERT + Transformer models under unified settings, emphasizing accuracy–efficiency trade-offs for practical NER deployment.

Contributions of the Study

The main contributions of this paper are summarized as follows:

(i) proposing a unified experimental framework for evaluating hybrid NER architectures,

(ii) systematically comparing BERT + CRF, BiLSTM + CRF, and BERT + Transformer models,

(iii) analyzing the impact of structured prediction on label consistency and performance, and

(iv) providing practical guidance on model selection for real-world information extraction systems.

2. Methodology

This section provides a comprehensive overview of the methodology employed to integrate deep context-aware representations with structured prediction through CRFs, evaluating three architectures: BERT + CRF, BiLSTM + CRF, and BERT + Transformer. Fig. 1 illustrates the methodology for NER using the three proposed models.

2.1 Proposed Methodology

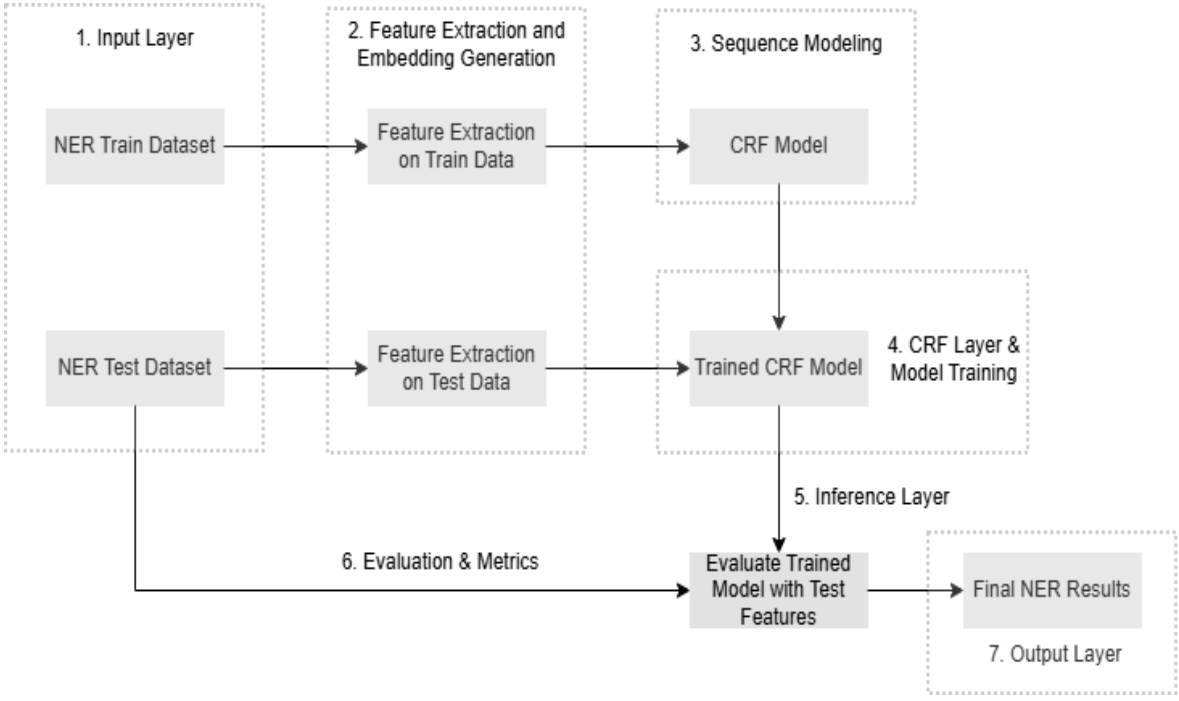

Fig. 1 Proposed methodology for NER using Conditional Random Fields (CRF)

1. Input Layer

The process begins with two datasets: one for training and one for testing. BERT models use WordPiece tokenization, which breaks down words into smaller parts. Meanwhile, BiLSTM uses standard word-level tokenization. For BERT, we use special tokens like [CLS] and [SEP] to mark the start and the end of a sequence. Entity labels follow the BIO format (Beginning, Inside, Outside), which helps track exactly where each entity starts and ends.

2. Feature Extraction and Embedding Generation Layer

Now, we extract features from both the training data and the testing data. Here, we examine features including part-of-speech tags, capitalization patterns, and where each word is placed in the sentence. These features get transformed into dense vectors. BERT embeddings handle deep, context-rich meaning from both directions, while BiLSTM tends to lean on pretrained models like GloVe. By handling the training and testing data identically in this context, we can ensure that the features remain uniform throughout.

3. Sequence Modeling

The system sends the extracted features from the training set to the CRF Model Training module. While using BiLSTM + CRF, the embeddings run through BiLSTM layers first in order to detect dependencies in both directions. The BERT + CRF architecture directly utilizes contextual embeddings which are generated by BERT. Within the BERT and Transformer combination, there's an extra transformer encoder included for modeling extended-range dependencies. This stage focuses on teaching the model to understand context within the words and sequences.

4. CRF Layer & Model Training

Following that, the CRF layer takes effect. It takes the emission scores from BERT or BiLSTM and learns how to transition smoothly between different entity tags—such as ensuring you avoid random label sequences (for example, B-LOC randomly switching to B-PER). Training uses negative log-likelihood loss to fine-tune those sequence-level predictions. The result is a fully trained CRF model.

The CRF layer estimates the conditional probability of the label sequence as depicted in Eq. 1 below:

$$P(y \mid x) = \frac{\exp(score(x,y))}{\sum_{y'} \exp(score(x,y'))} \quad \text{---------- Eq. 1}$$

where:

- x represents the token sequence taken as an input,
- y represents the label sequence which is predicted,
- score(x,y) represents the compatibility score between the input token sequence and label sequence,
- y' represents all possible label sequences.

The CRF layer chooses the label sequence which has the highest conditional probability while considering dependencies between neighbour entity tags. This ensures valid BIO tag transitions and helps improve sequential consistency in NER tasks.

5. Inference Layer

Once trained, the model moves to the inference phase. The test features and trained CRF model feed into the Evaluation module. The CRF uses the Viterbi algorithm to pick the most likely sequence of labels for the test data.

6. Evaluation and Metrics

The system then checks its predictions against the actual labeled data. We can monitor our performance using Precision, Recall, and F1-score. This evaluation covers modules, with final results landing in the output phase.

7. Output Layer

In the end, every token gets a structured entity label. For example, if you feed in "Barack Obama visited Paris," you get tags like [B-PER, I-PER, O, B-LOC]—structured, interpretable, and prepared for subsequent tasks.

2.2 Dataset

CoNLL-2003

The CoNLL-2003 shared task dataset comes from English Reuters news articles and stands as a widely used benchmark dataset for testing Named Entity Recognition (NER) systems. This dataset was introduced by Tjong Kim Sang and De Meulder in 2003 at the Conference on Computational Natural Language Learning.

https://huggingface.co/datasets/tner/conll2003

https://huggingface.co/datasets/eriktks/conll2003

Inside the dataset, there are four entity types: Person (PER), Organization (ORG), Location (LOC), and Miscellaneous (MISC). The MISC label is not simply a catch-all—it includes nationalities, events, products, and others, resulting in a comparatively more challenging dataset, as compared to older datasets that used fewer categories. Annotation adheres to the BIO format: "B-X" marks the beginning of an entity, "I-X" continues it, and "O" indicates you are outside any entity. This setup determines precisely where entities start and end, which really matters when entities span multiple words.

CoNLL-2003 splits everything up into training (about 15,000 sentences), development (3,500 or so), and test sets (around 3,700). Overall, you can expect nearly 300,000 tokens. This structure helps you train models, adjust settings, and test your work without overlap.

The reason CoNLL-2003 stands out in this study is its consistent annotations, the extensive collection of benchmark findings including approaches from BiLSTM–CRF up to BERT–CRF, and its status as a common standard among researchers. Using this dataset means you can compare results against prior studies in a clear and reproducible manner.

2.3 Model Architecture

The Named Entity Recognition (NER) system integrates feature engineering, sequence modeling, and structured prediction through a Conditional Random Field (CRF). Let's break down how everything integrates, using Fig 2.:

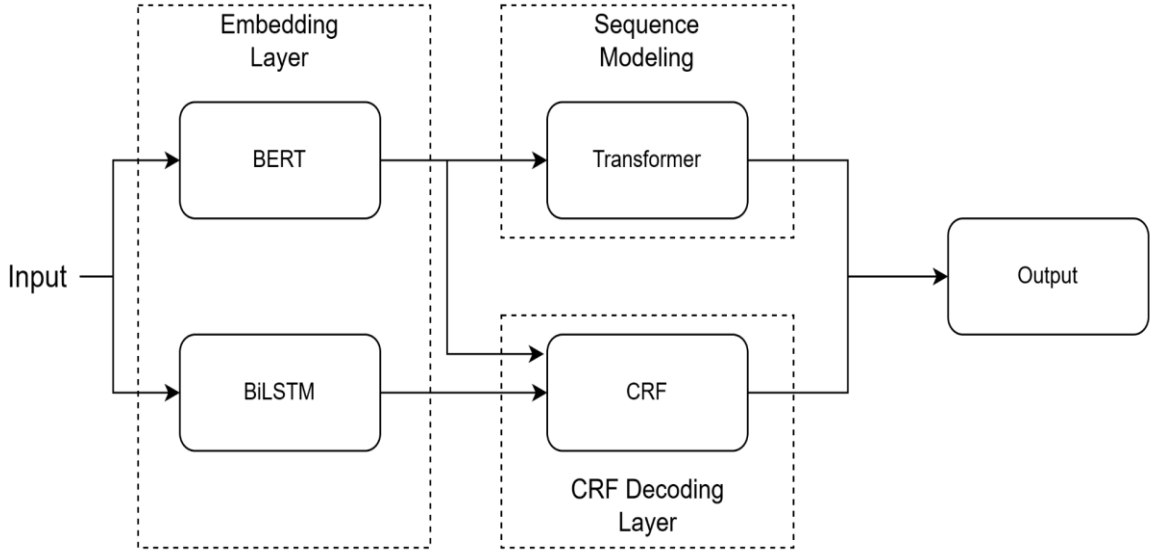

Fig. 2 Model Architecture for NER using BERT + CRF, BiLSTM + CRF and BERT + Transformer models

1. Embedding Layer

BERT models create embeddings that capture the information from both directions in the sentence—left and right. BiLSTM deals with fixed word vectors, usually from something like GloVe, so each word has a single semantic meaning. These vectors initiate the sequence learning.

2. Sequence Modeling

Contextual relationships between tokens are modeled using different architectures. BiLSTM captures the context from both sides—before and after every word. In the BERT + Transformer model, the additional transformer layers capture long-range contextual dependencies and global semantic relationships.

3. CRF Decoding Layer

For decoding, BERT + CRF and BiLSTM + CRF models use a Conditional Random Field layer. That means instead of just guessing each label on its own, the model checks the labels next to it and picks the best overall sequence. This approach maintains consistent labeling and ensures consistent results. The Viterbi decoding algorithm was used during inference to identify the optimal label sequence.

4. Training Strategy

At the training stage, all models use AdamW as the optimizer, with a learning rate schedule for smoother learning. CRF setups use negative log-likelihood loss, while the plain Transformer model relies on cross-entropy loss. Dropout helps avoid overfitting, so the models don't memorize the training data too closely. By using the same dataset and prep pipeline for each model, we can ensure a fair comparative analysis.

The proposed framework integrates contextual feature extraction with structured sequence decoding. Input sentences are tokenized and transformed into dense vector representations using either pretrained contextualized embeddings (BERT) or trainable word embeddings (BiLSTM). These representations are then passed through sequence modeling layers, followed by a Conditional Random Field (CRF) or transformer-based decoding layer to generate valid label sequences.

Three architectures are evaluated under identical preprocessing and evaluation conditions:

1. BERT + CRF
2. BiLSTM + CRF
3. BERT + Transformer

By comparing these models, we highlight how contextual embeddings (BERT/word vectors) combined with CRFs enhance boundary detection and sequence labeling. The methodology ensures a fair comparison by training all models on the same dataset with consistent preprocessing and evaluation pipelines.

BERT + CRF Architecture

This model breaks down the input sentences using WordPiece tokenization. Then, it runs everything through a pretrained BERT-base (uncased) model, which extracts context-rich embeddings — meaning every word captures the meaning of the words before and after it. These embeddings go through a linear layer and are placed in a CRF layer. The CRF doesn't simply assign labels to tokens individually; it looks at how labels connect to each other, making sure you get valid BIO tagging patterns. With the Viterbi algorithm, the CRF picks the best possible sequence, so predictions stay consistent.

BiLSTM + CRF Architecture

This setup starts with word embeddings, which the model tunes while training. Next, it sends them into a BiLSTM network, letting the system learn relationships going both forward and backward in the sequence. The outputs from that network move into a CRF layer, which handles label dependencies — so the structure of your labeling

is understandable. The difference is that this model works effectively with patterns in the sequence, although it captures limited context compared to those elaborate pretrained language models.

BERT + Transformer Architecture

Here, we obtain the BERT-generated embeddings, but instead of using a CRF layer, they are passed through more Transformer encoder layers. This boosts the model's ability for recognizing global patterns and long-range dependencies. At the end, it simply executes a softmax classifier on each token for the final prediction. However, without something like a CRF to link output labels, there might be BIO tag sequences that aren't always as consistent.

2.4 Algorithmic Configuration

- **Data Split:** 80% for training, 10% for validation, and 10% for testing.
- **Batch Size:** 16 (BERT + CRF), 8 (BERT + Transformer), 32 (BiLSTM + CRF)
- **Learning Rate:** 5e-5 (BERT models), 1e-3 (BiLSTM)
- **Optimizer:** Every model used the Adam optimizer.
- **Epochs:** Training was conducted for 3 epochs for CRF-based models with early stopping based on validation loss, but just 1 epoch for BERT + Transformer models to reduce computational overhead and training time.
- **Loss Function:** Negative log-likelihood (CRF-based models)
- **Hardware:** Experiments were conducted on an AMD A6 (7th-generation) processor, paired with integrated Radeon R4 graphics and 16 GB RAM. Since there was no dedicated GPU, all training and evaluation happened on the CPU.
- **Convergence Criterion:** Validation F1-score stabilization

2.5 Evaluation Metrics

The CoNLL-2003 dataset was divided into a training set, testing set and validation set. The model performance was evaluated using entity-level precision, recall, and F1-score. Precision tells us what fraction of predicted entities are actually correct. Recall checks how many of the real, annotated entities that the model identifies. F1-score finds the optimal point between precision and recall. Support merely records how many true instances of each entity type are in the dataset.

All the results are reported at the entity level—so the prediction must match both the boundary and the actual label in the gold-standard annotations to count as correct. This ensures the comparisons are fair between our CRF model and the standard baselines like BiLSTM-CRF and BERT-CRF. This is an effective approach to determine the model that manages entities most effectively.

All three architectures were evaluated under similar preprocessing, training, and evaluation criteria to ensure unbiased comparison and consistent, reproducible results.

3. Results and Discussion

The performance of these three models—BERT + CRF, BiLSTM + CRF, and BERT + Transformer on the CoNLL-2003 dataset is discussed in this section. All the models were trained with identical parameters, observing a fair comparative evaluation. The key metrics are accuracy, precision, recall, and F1-score, and their results really bring out the strengths and weaknesses of each approach, especially when it comes to handling context, keeping sequences consistent, and managing computational load.

First up, BERT + CRF comes out on top with an F1-score of about 0.94, as illustrated in Table 2. The combination achieves highly effective performance—BERT extracts all the fine-grained context from the dataset, while the

CRF layer makes sure the model's predicted labels consistently correspond. This approach helps eliminate problems like invalid BIO sequences, thereby producing more consistent and reliable predictions.

Table 2 Performance metric for BERT + CRF model

Validation Accuracy : 0.9896 Classification Report :					
	Precision	Recall	F1-score	Support	Accuracy
LOC	0.96	0.96	0.96	1837	98.96
MISC	0.89	0.90	0.90	922	98.96
ORG	0.90	0.91	0.91	1341	98.96
PER	0.96	0.97	0.97	1842	98.96
micro avg	0.94	0.94	0.94	5942	98.96
macro avg	0.93	0.94	0.93	5942	98.96
weighted avg	0.94	0.94	0.94	5942	98.96

At this point, the BiLSTM + CRF model presents a different case. Table 3 shows it achieves an F1-score of around 0.72. Certainly, it doesn't achieve the same performance as the transformer-based models, but it performs seamlessly, requires comparatively lower computational resources, and maintains consistent predictions. This is due to BiLSTM scanning the text in both directions and the CRF making the final labels consistent. Even so, the static GloVe word representations limit its potential—it cannot adapt alongside shifting word senses in real time. However, it needs way less computation, which makes it a practical option while being short on GPU power, working on a small project, or needing real-time results.

Table 3 Performance metric for BiLSTM + CRF model

Validation Accuracy : 0.9386 Classification Report :					
	Precision	Recall	F1-score	Support	Accuracy
LOC	0.88	0.72	0.79	1837	93.86
MISC	0.86	0.65	0.74	922	93.86

ORG	0.69	0.61	0.65	1341	93.86
PER	0.80	0.59	0.68	1842	93.86
micro avg	0.81	0.64	0.72	5942	93.86
macro avg	0.81	0.64	0.71	5942	93.86
weighted avg	0.81	0.64	0.71	5942	93.86

Table 4 illustrates that the BERT + Transformer model achieved an intermediate F1-score of around 0.81. This configuration utilizes BERT and adds additional Transformer layers, reaching deeper into the text for better long-range context. It helps, especially with complex sentences, but omitting the CRF causes it to occasionally make mistakes when tagging inconsistently. Additionally, the surplus modeling significantly slows training and slows inference. Unless you are handling a massive pipeline or a task requiring thorough comprehension first and foremost, the computational overhead may not always be justified.

Table 4 Performance metric for BERT + Transformer model

Validation Accuracy : 0.9471 Classification Report :					
	Precision	Recall	F1-score	Support	Accuracy
LOC	0.85	0.83	0.84	198	94.71
MISC	0.59	0.26	0.36	62	94.71
ORG	0.62	0.68	0.65	114	94.71
PER	0.94	0.87	0.90	365	94.71
micro avg	0.84	0.78	0.81	739	94.71
macro avg	0.75	0.66	0.69	739	94.71
weighted avg	0.84	0.78	0.80	739	94.71

Comparing precision and recall across these models shows some key trade-offs. The BERT + CRF model distinguishes itself—it achieves both precision and recall at 0.94. This implies it not only detects entities

accurately but also seldom overlooks details. The BiLSTM + CRF, on the other hand, shows comparatively lower recall performance. It occasionally fails to detect multi-token entities, so it does not capture the entire phrase. BERT + Transformer boosts recall marginally, but its precision drops, which leads to higher entity detection at the cost of additional false predictions.

Token-level accuracy is one aspect, but precision, recall, and F1-score focus on the entity level, using the BIO tagging scheme. Out of all of them, BERT + CRF gets the highest F1-score. This is the result for combining strong semantic representations with a CRF for structured prediction. Basically, the CRF refines the process, making sure the label sequences are coherent, instead of random or inconsistent tags.

BiLSTM + CRF simply lacks the equivalent advantage. It does not include those pretrained contextual representations, does not cope well with long dependencies, and is limited by the size of the dataset. It cannot acquire knowledge to the same extent. BERT + Transformer benefits from understanding context deeply, but without the CRF's structure, it doesn't always get the sequences right. So, BERT + CRF takes the lead.

Instead of predicting labels independently for each token, CRF based structured prediction models dependencies between neighboring tags and significantly improves the performance of sequence labeling. Transformer models perform token-level classification, CRFs jointly optimize the entire label sequence, thus guaranteeing valid BIO transitions and improving sequence consistency. This makes the BERT + CRF model good at both deep contextual understanding and schema-enforced sampling.

There are still challenges. The models sometimes fail to identify the exact boundaries of entities, especially when they're more than one word. The evaluated models demonstrate limited capability in handling nested or overlapping entities, which indicates that span-based methods may be beneficial in the future work. Rare and domain-specific entities remain challenging to identify , mostly because there aren't enough training examples. Adding character-level features or fine-tuning on domain-specific data might help cover those gaps.

As illustrated in Table 5, the results are in line with previous work that has shown that combining contextual embeddings with structured prediction methods for Named Entity Recognition (NER) is effective. Devlin et al. [10] proposed BERT, which significantly improved the understanding of the context of words in NLP tasks by using bidirectional transformer representations. Likewise, multiple studies have shown that the inclusion of CRF layers on top of context-dependent mapping improves sequence consistency and entity boundary detection in NER systems. The proposed BERT + CRF model achieved a better F1-score of 0.94, which is consistent with the results of prior BERT-CRF based approaches [10][15][19].

Compared to traditional CRF-based methods [1], the proposed BERT + CRF architecture allows for better semantic understanding and better handling of long-range contextual dependencies thanks to transformer-based dynamic word embeddings. The results are also in line with the findings of Li et al. [5] and Nasar et al. [4] that contextual representation learning greatly improves NER accuracy over traditional handcrafted feature-based methods.

The performance of BiLSTM + CRF was relatively lower, which is consistent with previous studies that the BiLSTM model relies heavily on static embeddings and is less capable of capturing contextual semantics compared to transformer-based architectures [2][3]. However, the obtained results demonstrate that the BiLSTM + CRF is still computationally efficient and suitable for lightweight or resource-constrained applications.

The BERT + Transformer model demonstrated strong contextual learning capability , but produced less sequential consistency than BERT + CRF. This observation is consistent with previous work showing that token-level transformer classifiers can produce inconsistent BIO tag transitions in the absence of grammar-constrained decoding mechanisms [11][12]. More transformer layers help learn long range dependencies but the lack of CRF based structured prediction hurts label consistency.

The primary novelty of this study lies in the unified comparative evaluation of BERT+CRF, BiLSTM+CRF and BERT+Transformer models together under the same pre-processing, training and evaluation conditions using the CoNLL-2003 dataset. In contrast to many previous studies that explored a single architecture or domain-specific

implementation, this work provides a comprehensive comparison of contextual learning, sequence consistency, computational efficiency and practical applicability under a unified experimental setting. Furthermore, the paper discusses the trade-off between model accuracy and computational overhead, providing practical insights into choosing suitable NER architectures for real-world deployment.

Table 5 Comparison of Proposed NER Models with Existing Literature Based on F1-Score Performance

Model	**Dataset / Domain**	**F1-Score**
Deep Learning Sequential Labeling [2][3]	Arabic NER / Breast Cancer Clinical Data	0.84–0.87
CRF-based Sequence Labeling [1][8][9]	Biomedical / Sequence Labeling	0.88–0.91
MRC-based NER [12][18]	Biomedical NER	0.89–0.90
BiLSTM-CRF Based Models [6][13]	Geological / Medical NER	0.90–0.91
Graph-based NER [19]	Biomedical NER	0.92
Demonstration-based Few-shot Learning [20]	Biomedical NER	0.91
BERT + CRF (Proposed Work)	**CoNLL-2003**	**0.94 (Best Performance)**
BiLSTM + CRF	CoNLL-2003	0.72
BERT + Transformer	CoNLL-2003	0.81

In summary, Table 6 highlights that when you combine contextual embeddings and a CRF, there is a significant improvement in NER performance. BERT + CRF gives you a strong combination of semantic understanding, consistency, and efficiency. BiLSTM + CRF remains a reliable choice if you're working with fewer resources. BERT + Transformer can handle trickier contextual patterns, but at a cost. In the end, blending deep contextual knowledge with structured CRF-based decoding is both theoretically sound and empirically validated to enhance named entity recognition.

Table 6 Comparative analysis of Proposed NER models based on contextual learning, structured prediction, and practical applicability

Criteria	**BERT + CRF**	**BiLSTM + CRF**	**BERT + Transformer**
Contextual Representation	**High (deep bidirectional)**	Moderate (word-level)	**High+ (extended)**
Sequence Consistency	**Excellent**	Excellent	Good
Computational Cost	Moderate	**Low**	**High**
Domain Adaptability	**Strong**	Moderate	Strong
Long-Range Context	Good	Moderate	**Excellent**
Overall Suitability	**Best for high-accuracy systems**	**Best for lightweight setups**	**Best for long-context tasks**

4. Conclusion

This study set out to enhance and compare three deep learning architectures for Named Entity Recognition (NER), namely BERT+CRF, BiLSTM+CRF and BERT+Transformer on the CoNLL-2003 dataset under the same preprocessing, training and evaluation conditions. The experimental results showed that the combination of the contextual representations and the structured sequence decoding can significantly improve the NER performance.

The best overall performance was achieved by the proposed BERT + CRF architecture with an F1-score of 0.94 and validation accuracy of 98.96% outperforming the models of BiLSTM + CRF (F1-score: 0.72), BERT + Transformer (F1-score: 0.81) among all the evaluated models. The better performance of BERT + CRF shows that the integration of bidirectional contextually relevant representations and CRF-based structured prediction can effectively improve the detection of entity boundaries, the consistency of BIO tags, and the accuracy of sequence labeling.

The relatively lower performance of BiLSTM + CRF model was based on the fact that it used static word embeddings and had limited contextual understanding. However, it has lower computational complexity and is applicable to lightweight and resource-constrained environments. The BERT + Transformer model demonstrated better contextual learning and long-range dependency modeling but lacked CRF-based structured decoding, leading to lower contextual coherence and higher computational overhead.

The primary novelty of this work is the comparative evaluation of BERT + CRF, BiLSTM + CRF and BERT + Transformer architectures under the same dataset, preprocessing pipeline, training configuration and evaluation metrics. Unlike many previous studies that focused mostly on one single architecture or domain-specific implementation, the present study provides a comprehensive comparison of contextual learning capability, sequence consistency, computational efficiency and practical applicability in a common experimental framework.

The results of this work show that the combination of contextual embedding models and CRF-based syntax-aware decoding provides an effective and scalable solution for real-world NER applications, such as information extraction, healthcare text mining, enterprise analytics, question answering systems, and knowledge graph construction.

Future research can focus on the model efficiency improvement by the model compression and lightweight transformer architecture for real-time deployment. Other potential avenues for future work include the handling of nested and overlapping entities, multilingual and low-resource NER, fine-tuning for a specific domain, and integration with downstream NLP applications such as relation extraction and knowledge graph generation. In addition, the inclusion of span-based learning, attention optimization, and character-level contextual representations might also boost the entity recognition performance in complex and domain-specific settings.

References

[1] Guang Xun Sun, Cheng Jie Zhou, Han Yu Zhao, Bo Jin, Zhan Gao, "Fast and effective biomedical named entity recognition using temporal convolutional network with conditional random," *in PubMed Mathematical Biosciences and Engineering,* 2020 May 12. https://doi.org/10.3934/mbe.2020200

[2] K. K. Shahina, P. V. Jyothsna, Greeshma Prabha, Premjith B., Soman Kp "A Sequential Labelling Approach for the Named Entity Recognition in Arabic Language Using Deep Learning Algorithm*", in 2019 International Conference on Data Science and Communication (IconDSC)*, March 2019.https://doi.org/10.1109/IconDSC.2019.8817039

[3] Zhang X, Zhang Y, Zhang Q, Ren Y, Qiu T, Ma J, Sun Q, "Extracting comprehensive clinical information for breast cancer using deep learning methods", *Int J Med Inform,* 2019 Dec.https://doi.org/10.1016/j.ijmedinf.2019.103985

[4] Nasar Z., Jaffry S.W., Malik M.K., "Named entity recognition and relation extraction: state-of-the-art.", *ACM Comput. Surv,* 2021.https://doi.org/10.1145/3445965

[5] Li J., Sun A.X., Han J.L., Li C.L., "A survey on deep learning for named entity recognition.", *IEEE Transactions on Knowledge and Data Engineering,* Volume: 34, Issue: 1, 01 January 2022, pp. 50-70.https://doi.org/10.1109/TKDE.2020.2981314

[6] Qiu Q.J., Xie Z., Wu L., Tao L.F., Li W.J., "BiLSTM-CRF for geological named entity recognition from the geoscience literature.", *EARTH SCIENCE INFORMATICS,* 2019, pp. 565–579.https://doi.org/10.1029/2019ea000610

[7] Chen M.J., Luo X., Shen H.L., Huang Z.Y., Peng Q.J., "A novel named entity recognition scheme for steel E-commerce platforms using a lite BERT.", *CMES-COMPUTER MODELING IN ENGINEERING & SCIENCES,* 2021.https://doi.org/10.32604/cmes.2021.017491

[8] Harshil Shah, Tim Xiao, David Barber, "Locally-Contextual Nonlinear CRFs for Sequence Labeling", *arXiv,* 2021, pp. 235–240.https://doi.org/10.18653/v1/P18-2038

[9] Tianwen Wei, Jianwei Qi, Shenghuan He, Songtao Sun, "Masked Conditional Random Fields for Sequence Labeling", *ACL Anthology,* 2021, pp. 2024–2035.https://doi.org/10.18653/v1/2021.naacl-main.163

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", *ACL Anthology,* 2019, pp. 4171–4186.https://doi.org/10.18653/v1/N19-1423

[11] Juntao Yu, Bernd Bohnet, and Massimo Poesio, "Span-based NER using BERT and CRF", *ACL Anthology,* 2020, pp. 6470–6476.https://doi.org/10.18653/v1/2020.acl-main.577

[12] Cong Sun, Zhihao Yang, Lei Wang, Yin Zhang, Hongfei Lin, Jian Wang, "Biomedical named entity recognition using BERT in the machine reading comprehension framework", *Elsevier - Journal of Biomedical Informatics,* 2020.https://doi.org/10.1016/j.jbi.2021.103799

[13] Chaofan Li, Kai Ma, "Entity recognition of Chinese medical text based on multi-head self-attention combined with BILSTM-CRF", *Mathematical Biosciences and Engineering,* 2022, pp: 2206-2218.https://doi.org/10.3934/mbe.2022103

[14] Hangle Hu, Chunlei Cheng, Qing Ye, Lin Peng, Youzhi Shen, "Enhancing traditional Chinese medicine diagnostics: Integrating ontological knowledge for multi-label symptom entity classification"*, Mathematical Biosciences and Engineering,* 2024, pp: 369-391.https://doi.org/10.3934/mbe.2024017

[15] Peng C, Wang X, Li Q, Yu Q, Jiang R, Ma W, Wu W, Meng R, Li H, Huai H, et al., "A New Chinese Named Entity Recognition Method for Pig Disease Domain Based on Lexicon-Enhanced BERT and Contrastive Learning", *Journal of Applied Sciences,* 2024.https://doi.org/10.3390/app14166944

[16] Kai He; Rui Mao; Yucheng Huang; Tieliang Gong; Chen Li; Erik Cambria, "Template-Free Prompting for Few-Shot Named Entity Recognition via Semantic-Enhanced Contrastive Learning", *IEEE Transactions on Neural Networks and Learning Systems ,* Volume: 35, Issue: 12, December 2024.https://doi.org/10.1109/TNNLS.2023.3314807

[17] Shan Zhao; Minghao Hu; Zhiping Cai; Fang Liu, "Dynamic Modeling Cross-Modal Interactions in Two-Phase Prediction for Entity-Relation Extraction", *IEEE Transactions on Neural Networks and Learning Systems,* Volume: 34, Issue: 3, March 2023.https://doi.org/10.1109/TNNLS.2021.3104971

[18] Huang Z, He L, Yang Y, Li A, Zhang Z, Wu S, Wang Y, He Y, Liu X., "Application of machine reading comprehension techniques for named entity recognition in materials science.", *J Cheminform,* 2024.https://doi.org/10.1186/s13321-024-00874-5

[19] Lou Y, Zhu X, Tan K., "Dictionary-based matching graph network for biomedical named entity recognition.", *Sci Rep,* 2023.https://doi.org/10.1038/s41598-023-48564-w

[20] Leilei Su, Jian Chen, Yifan Peng, Cong Sun, "Demonstration-based learning for few-shot biomedical named entity recognition under machine reading comprehension", *Journal of Biomedical Informatics,* Volume 159, Issue C, 2024.https://doi.org/10.1016/j.jbi.2024.104739