Comparative Analysis of Bi-LSTM + CRF and BERT for Named Entity Recognition

Anantha Narayanan Sampath Varadharajan

School of Engineering and Sciences, George Washington University
Washington, DC, USA

Abstract

Named Entity Recognition (NER) is a critical task in Natural Language Processing (NLP) that involves identifying and categorizing entities such as persons, organizations, locations, and dates within unstructured text. This study compares the performance of traditional and modern neural network models for NER, specifically evaluating a Bidirectional Long Short-Term Memory with Conditional Random Fields (Bi-LSTM + CRF) model and a Transformer-based model, Bidirectional Encoder Representations from Transformers (BERT), using the Conference on Natural Language Learning (CoNLL)-2003 dataset. The comparison emphasizes key metrics—precision, recall, and F1-score—to comprehensively assess their effectiveness. Experimental results reveal that both models achieve comparable performance across the evaluated metrics under standard conditions. However, fine-tuned BERT demonstrates superior adaptability and robust performance, particularly when tested under adversarial conditions, where Bi-LSTM + CRF shows notable limitations. This analysis highlights the strengths and weaknesses of each model, offering valuable insights into their suitability for diverse NER tasks.

1 Introduction

Named Entity Recognition (NER) is a pivotal task in NLP that has seen significant advancements. Neural networks, including Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM) improved NER by capturing sequential dependencies, with Conditional Random Fields (CRF) layers enhancing label modeling. Recently, Transformer-based models like Bidirectional Encoder Representations from Transformers (BERT) have set new benchmarks by capturing deep contextual relationships across sequences.

This study focuses on a comparative analysis of a Bidirectional Long Short-Term Memory (Bi-LSTM) combined with CRF and the Transformer-based BERT model on the benchmark dataset for NER, Conference on Natural Language Learning (CoNLL)-2003 dataset. The Bi-LSTM + CRF model employs recurrent networks to capture contextual dependencies in text, while BERT leverages self-attention mechanisms to effectively model long-range dependencies, offering a distinct advantage in handling complex contexts.

The BiLSTM+CRF is manually trained, a pre-trained BERT is fine-tuned, and their performance is assessed using standard metrics: precision, recall, and F1-score.

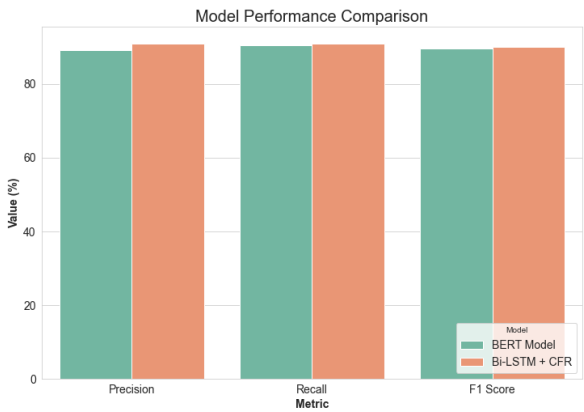

Fig 1: Model Metrics

From Fig 1, Results indicate that the model metrics comparison shows that both Bi-LSTM + CRF and BERT perform similarly, with no significant differences. Bi-LSTM + CRF slightly outperforms BERT in precision (91% vs. 89.07%), Recall (91% vs. 90.5%), and F1 score (90% vs. 89.7%). Overall, the performance of both models is comparable

across precision, recall, and F1-score suggesting that both approaches are viable for NER tasks, each with unique strengths.

When tested on adversarial sentences, intentionally crafted to mislead or confuse machine learning models, BERT significantly outperforms Bi-LSTM + CRF, showcasing its robustness and adaptability to challenging inputs. These findings highlight the robustness, generalization, and performance of the models.

2 Methodology

2.1 Dataset Overview

The CoNLL-2003 dataset, introduced during the 2003 Conference on Natural Language Learning (CoNLL), is widely recognized as a benchmark for Named Entity Recognition (NER) tasks. It was designed to evaluate models for identifying and classifying named entities in text. The dataset includes English and German texts, but for this study, only the English dataset is used.

The English portion of the dataset was sourced from Reuters news articles published between August 1996 and August 1997. For the training and development sets, data from the end of August 1996 were used, while the test set consisted of articles from December 1996. The dataset includes annotations for four entity types: PER (Person), ORG (Organization), LOC (Location), and MISC (Miscellaneous), which are labeled at the token level.

The English dataset consists of 14,987 sentences and 203,621 tokens, with 7,140 instances of PER, 3,438 of LOC, 6,321 of MISC, and 6,600 of ORG. This dataset has become a crucial resource for training and evaluating NER models, providing a standardized benchmark for performance and enabling the development of more accurate and efficient entity recognition systems.

2.2 Data Preprocessing

2.2.1 Bi-LSTM Preprocessing

The preprocessing for the Bi-LSTM + CRF model begins by loading the CoNLL-2003 dataset and converting all tokens to lowercase, ensuring uniformity in the data. A vocabulary lookup layer is created by mapping tokens to numerical IDs based on the training set. This step allows the model to handle out-of-vocabulary words during inference. Additionally, the preprocessing includes adjusting tag IDs to account for padding tokens, ensuring that the model correctly handles variable-length sequences. Once the transformations are applied, the data is batched and cached to enhance the training process by ensuring consistent input shapes and faster data loading.

2.2.2 BERT Preprocessing

For the BERT model, preprocessing also starts with loading the CoNLL-2003 dataset. A pre-trained BERT tokenizer is used to tokenize the text, and special care is taken to align the tokenized words with their respective labels. This alignment is crucial, as BERT tokenizes text into subword units, which may result in a mismatch between the original words and their labels. To address this, the labels are adjusted so that each token, including special tokens like [CLS] and [SEP], receives the appropriate label. The tokenized data is then grouped into batches, which are further processed to prepare it for training and evaluation.

2.2.3 Key Differences

Tokenization Approach

- In the Bi-LSTM model, tokens are directly mapped to numerical IDs, with lowercase conversion ensuring consistency across the dataset.
- In the BERT model, tokenization is handled by a pre-trained tokenizer, which may split words into subword units, requiring additional label alignment to ensure the correct labels are assigned to these subword tokens.

Special Tokens

- Bi-LSTM preprocessing does not need to handle special tokens like [CLS] and [SEP] as explicitly as BERT, since these tokens are integral to BERT's architecture for classification and sequence segmentation.

Handling Padding

- Both models handle padding, but the Bi-LSTM model requires adjustments to tag IDs to accommodate the padding token, whereas BERT uses its pre-trained tokenizer to manage tokenization and padding efficiently.

2.3 Model Implementation

Two models have been evaluated for the study: A transformer-based BERT model and a classical Bi-LSTM + CRF model.

2.3.1 BERT

BERT-Base-Cased is a case-sensitive transformer-based language model widely used for natural language processing (NLP) tasks, here it is used for Named Entity Recognition (NER), by fine-tuning the model on the dataset.

Characteristics of Pre-Trained BERT

1. Model Architecture

- **Size**: 12 Transformer layers, 768 hidden units, and 12 attention heads.
- **Case Sensitivity**: Differentiates uppercase and lowercase letters, crucial for tasks like NER where capitalization provides meaningful cues.

2. Pre-training

- Trained on Book Corpus and English Wikipedia. It was trained to do the following tasks:
- Masked Language Modeling (MLM): Predicting masked words in sentences.
- Next Sentence Prediction (NSP): Understanding sentence relationships.

The bert-base-cased model will take the input of the tokenized text sequence and output the predicted label corresponding to the named entity of the token.

2.3.2 Bi-LSTM with CRF

The Bi-LSTM-CRF model combines Bidirectional Long Short-Term Memory (Bi-LSTM) networks with a Conditional Random Field (CRF) layer to perform sequence labeling tasks. It is designed to leverage both the sequential representation power of LSTMs and the structured prediction capability of CRFs, making it well-suited for this NER task.

What is Long Short-term Memory (LSTM), and how is Bi-LSTM developed based on LSTM?

1. Long Short-Term Memory (LSTM)

LSTM is a type of recurrent neural network (RNN) designed to capture long-term dependencies in sequential data while addressing the vanishing gradient problem. It uses three gates (input, forget, and output) to control the flow of information:

- **Input Gate**: Decides how much new information to allow into the memory cell.
- **Forget Gate**: Determines which information to discard from the memory cell.
- **Output Gate**: Filters the information to pass on to the next layer.

2. Bidirectional LSTM (Bi-LSTM)

A Bi-LSTM processes input sequences in both forward and backward directions. This allows the model to incorporate both past and future context for each token in the sequence, improving its understanding of sequential dependencies.

Characteristics of the Bi-LSTM-CRF Model

Model Architecture

Embedding Layer:
Converts integer word indices (input tokens) into dense vector representations, capturing semantic information about each word.

Bi-LSTM Layer:
Processes the embeddings bidirectionally, capturing contextual information from both preceding and succeeding tokens.
Outputs a sequence of hidden states for all input tokens.

CRF Layer:
Predicts the most likely sequence of labels for the given inputs by enforcing valid label transitions.
Learns transition probabilities between labels, allowing it to model dependencies.

The CRF layer gives out 4 outputs of which the decode sequence is used, which returns the predicted sequence of labels for the input tokens.

Feature	Bi-LSTM-CRF	BERT-Base-Cased
Input Representation	Embedding layer with learned word vectors	Pre-trained contextual embeddings
Contextual Learning	Captures past and future context explicitly using LSTM	Captures context via self-attention layers
Prediction Layer	CRF for structured predictions	Linear layer for token-level predictions

Table 1: Model's Architecture Comparison

3 Experiments

3.1 Model Training

The Bi-LSTM-CRF and BERT models were trained on the CoNLL-2003 dataset under configurations optimized for their respective architectures.

3.1.1 Bi-LSTM-CRF

The Bi-LSTM-CRF model was trained with a large batch size of 2048 to efficiently process the dataset. The Adam optimizer, with a learning rate of 0.02, was employed for adaptive gradient updates, facilitating faster convergence. The training was conducted over 10 epochs, providing the model with sufficient iterations to learn from the data. While dropout regularization was not explicitly applied, it could enhance generalization in future implementations. Early stopping criteria were not used, and the model trained for the full number of epochs as predefined. The model was evaluated on the test data after it was completely trained on the training data.

3.1.2 BERT

The fine-tuning of the pre-trained 'bert-base-cased' BERT model utilized a batch size of 16 to balance memory requirements and training efficiency. The AdamW optimizer, with a learning rate of 2e-5 and weight decay of 0.01 for regularization, was employed to fine-tune the model. Similar to the Bi-LSTM-CRF setup, the BERT model was trained for 10 epochs without early stopping. Evaluation was conducted at the end of each epoch to track performance on the test data.

Both models were trained to maximize performance on Named Entity Recognition tasks, with the training configurations tailored to their architecture-specific requirements. These consistent settings ensured a fair comparison of their effectiveness in extracting named entities trained based on the dataset.

3.2 Evaluation Metrics

The performance of the Bi-LSTM-CRF and BERT models was evaluated using standard metrics widely employed in sequence labeling tasks: precision, recall, and F1-score. These metrics provide a comprehensive assessment of the model's ability to correctly identify and classify named entities in text while balancing false positives and false negatives.

Precision: Evaluate the proportion of correctly predicted positive instances among all predicted positive instances. This metric is crucial for understanding how often the model's predictions are accurate when it identifies an entity.

Recall: Measures the proportion of correctly predicted positive instances among all actual positive instances in the dataset. This metric highlights the model's ability to identify all relevant entities.

F1-Score: Combines precision and recall into a single metric by calculating their harmonic mean, offering a balanced evaluation of the model's performance. This is particularly useful in cases where precision and recall need to be equally prioritized.

The models were evaluated on the test set of the CoNLL-2003 dataset, ensuring that the results reflect their generalization capability. These metrics were computed for each entity type and averaged to obtain an overall score for the model. The evaluation results were used to compare the effectiveness of the Transformer-based BERT model with the traditional Bi-LSTM-CRF model, providing insights into their respective strengths and weaknesses in Named Entity Recognition tasks.

4 Results

The performance of the models on the CoNLL-2003 test dataset is summarized below:

Metric (%)	BERT Model	Bi-LSTM + CFR
Precision	89.07%	91%
Recall	90.5%	91%
F-1 Score	89.7%	90%

Table 2: Model's Metric Comparison

The evaluation metrics show that the performance of the Bi-LSTM + CRF model and the BERT model is very similar. The precision, recall, and F1-score of the BERT model are 89.07%, 90.5%, and 89.7%, respectively, when compared to the Bi-LSTM + CRF metrics, which are 91%, 91%, and 90%. These slight differences in metrics imply that both models produce similar and accurate predictions. The overall performance of both models is quite alike based on the metrics suggesting that both models are equally effective for NER tasks according to these metrics tabulated in above Table 2.

When put to test on sample data, both the models accurately recognized entities in sentences. Where both models accurately recognized entities as displayed in Table 3.

Example Entity Recognitions
Input: EU rejects German call to boycott British lamb **Actual Entities:** EU – ORG, German – MISC, British – MISC
BiLSTM+CRF: EU – ORG, German – MISC, British – MISC **BERT:** EU – ORG, German – MISC, British – MISC

Table 3: Entity Recognition Model Predictions

5 Analysis

5.1 Computational Analysis

The computational analysis reveals that the Bi-LSTM+CRF model is significantly more efficient than BERT. From Table 4, Bi-LSTM+CRF takes just 4.2 seconds per epoch and has 691,074 parameters, while BERT requires 144 seconds per epoch and has 107,726,601 parameters. This highlights Bi-LSTM+CRF's lower computational cost, making it more suitable for resource-constrained environments, though BERT may offer superior performance in terms of robustness.

Computations	BERT Model	Bi-LSTM + CFR
Time for Epoch (in seconds)	144s	4.2s
Parameters	107726601	691074

Table 4: Model's Computation Comparison

5.2 Adversarial Analysis

To further compare and evaluate the robustness of the models subjected to adversarial conditions (intentionally messing with the input text to mislead the model), it is important to note that while both models exhibited robust performance on the standard dataset, the BiLSTM+CRF model proved more susceptible to adversarial attacks from Table 5 below.

Example Adversarial Texts
Input: Apple Inc. is based in Cupeptino, California. **Actual Entities:** Apple Inc – ORG, California - LOC
BiLSTM+CRF: Apple – ORG, California - O **BERT:** Apple Inc – ORG, California – LOC
Input: Joe Biden was born in Honolulu, Hawaii. **Actual Entities:** Joe – PER, Biden – PER, Honolulu – LOC, Hawaii – LOC
BiLSTM+CRF: Joe – O, Biden – O, Honolulu – O, Hawaii – O **BERT:** Joe – PER, Biden – PER, Honolulu – LOC, Hawaii – LOC
Input: The GrandCanyon is a major tourist attraction in Arizona. **Actual Entities:** GrandCanyon – LOC, Arizona – LOC
BiLSTM+CRF: GrandCanyon – MISC, Arizona – LOC **BERT:** GrandCanyon – LOC, Arizona – LOC

Table 5: Model's Adversarial Text Prediction

While the reasons for the poor performance of BiLSTM+CRF models compared to BERT on the adversarial text were not thoroughly analyzed,

some possible explanations include BiLSTM+CRF's dependence on sequential data and its restricted capacity to grasp deeper semantic and contextual nuances. This contrasts with BERT, which benefits from pre-trained text corpora that enable it to learn richer contextual representations.

6 Discussions

The findings in the study highlight significant considerations when selecting models for Named Entity Recognition (NER) tasks. Both Bi-LSTM + CRF and BERT exhibit comparable performance, with minor differences in precision, recall, and F1-score. However, the computational analysis underscores the practical trade-offs: Bi-LSTM + CRF demonstrates significantly lower computational costs, requiring just 4.2 seconds per epoch and having fewer parameters (691,074) compared to BERT's 144 seconds per epoch and over 107 million parameters. This efficiency makes Bi-LSTM + CRF more suitable for deployment in resource-constrained environments.

Conversely, BERT's resilience to adversarial inputs and its ability to capture nuanced contextual dependencies through its transformer-based architecture make it more robust for real-world applications with noisy or adversarial data. These findings suggest that while Bi-LSTM + CRF is advantageous for standard tasks where computational efficiency is critical, BERT's superior adaptability and robustness render it ideal for complex and dynamic NER applications. Future research could further explore this trade-off by assessing model performance on domain-specific datasets and under varied computational constraints.

7 Conclusion and Future Work

7.1 Conclusion

This study provides a comparative analysis of a manually trained Bi-LSTM + CRF model and a fine-tuned BERT model for Named Entity Recognition (NER) using the CoNLL-2003 dataset. The results reveal that while both models achieve comparable performance in terms of standard metrics and normal conditions, BERT demonstrates superior robustness in handling adversarial scenarios. These findings underscore the value of modern transformer-based architectures in addressing complex and noisy data conditions.

The implications of this study highlight that traditional models like Bi-LSTM + CRF remain relevant due to their efficiency and effectiveness in standard settings, while transformer-based models like BERT offer enhanced adaptability and contextual understanding.

In conclusion, the choice of model for NER tasks should be informed by the specific requirements of the application, balancing performance, robustness, and computational considerations. This study provides a foundation for further exploration and optimization of NER models in diverse contexts.

7.2 Future Work

Building on these findings, future research could:

- Evaluate the performance of these models on domain-specific datasets, such as biomedical, financial, or legal texts, to assess their adaptability across different fields.
- Explore adversarial training strategies to enhance the robustness of Bi-LSTM + CRF models, potentially closing the gap with transformer-based models.
- Investigate the trade-offs between computational efficiency and performance in low-resource environments, guiding model selection in diverse use cases.

Ethics Statement

Named Entity Recognition (NER) systems have significant societal and ethical implications due to their application in areas like information extraction, real-time analytics, and knowledge graph construction. While the models in this study show high effectiveness, their deployment must address several ethical considerations:

- **Privacy Concerns:** NER systems often handle sensitive data, such as personal information, raising risks of privacy violations if not properly secured. Adhering to data privacy laws and implementing robust anonymization measures are crucial to mitigate these risks.

- **Misuse Risks:** NER systems could be exploited for unethical purposes, such as mass

surveillance or misinformation. Safeguards must be implemented to prevent misuse and ensure responsible deployment.

- **Transparency and Accountability:** NER models influence critical decisions in domains like healthcare and law enforcement. Explainable AI (XAI) techniques can enhance transparency and foster trust in their predictions.

Despite these challenges, NER technology holds immense potential for societal benefit. Balancing technological progress with ethical responsibility is essential to maximize its positive impact while minimizing potential harm. This work aims to contribute to the responsible advancement of NER systems.

References

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Panchendrarajan, Rrubaa, and Aravindh Amaresan. 2018. Bidirectional LSTM-CRF for Named Entity Recognition. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation (PACLIC 32)*, pages 531–539, Hong Kong. December 1–3, 2018. Association for Computational Linguistics.

Chiu, Jason P. C., and Eric Nichols. 2016. Named Entity Recognition with Bidirectional LSTM-CNNs. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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: Human Language Technologies (NAACL-HLT)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics