1 Introduction

During my internship at AIRI, I developed and compared classical machine learning and Large Language Model (LLM)-based approaches for the Named Entity Recognition (NER) task aimed at constructing Knowledge Graphs (KGs). Specifically, I implemented and evaluated two models: the Conditional Random Fields (CRF) framework as a traditional method, and a hybrid architecture combining BERT's contextualized embeddings with CRF for enhanced sequence labeling.

NER is a critical NLP task that involves identifying entities such as people and organizations in unstructured text. Knowledge Graphs (KGs), in turn, are graph-based structures that represent knowledge as interconnected nodes (entities) and edges (relationships). The integration of NER with KGs is essential for transforming raw textual data into structured, machine-readable knowledge. LLMs (Large Language Models) are transformer-based models pre-trained on vast textual data to capture contextual semantics and linguistic knowledge, enabling tasks like NER. State-of-the-art NER methods typically utilize LLMs to take advantage of their deep contextual understanding and linguistic knowledge, enabling more accurate identification and classification of entities in ambiguous or complex texts [5].

In my research, I explore the difference between the use of classical machine learning and LLM-based approaches and assess their performance. The results (see Figure 5) show that the BERT+CRF model outperforms the classical CRF model, achieving higher Micro-F1 (0.95 vs. 0.93) and Macro-F1 (0.86 vs. 0.79), indicating better performance across all types of entities, including unbalanced classes.

2 Related Work

First, I reviewed the literature devoted to the NER task for constructing KGs. One of the most relevant and recent articles [2] focused on biomedical NER and relationship extraction (RE). The authors provided an overview of various approaches to NER and RE, including dictionary-based, rule-based, machine learning-based, and hybrid methods. They highlighted that deep learning techniques, such as LSTM-CRF architectures and pre-trained models like BioBERT, demonstrate the best performance in identifying domain-specific entities (e.g., genes, proteins, diseases).

In this article [6], the authors introduced CoNLL-2003 (based on news articles) and Ontonotes 5.0 (based on multilingual speaking), widely used datasets NER containing annotated texts with entities like persons, organizations, etc. They evaluated traditional models such as CRF (Conditional Random Fields) [4], which served as a strong baseline by leveraging lexical resources (e.g., dictionaries, gazetteers) and Skip-gram embeddings, achieving approximately 90% of F1-score on the test set for CoNLL-2003 and 82% for Ontonotes 5.0.

The [3] article focused on extracting biomedical NER and RE for KGs from real-world 505 patient biomedical unstructured clinical notes. The authors used deep learning models combining different BERT variants in the medical domain (e.g. BioBERT, BioClinicalBERT) with CRF layer. The experiments showed the highest accuracy for the NER task with the Bio_ClinicalBERT + CRF architecture (90.7%).

For my research I decided to use CoNLL04 dataset which is not widely recognized for NER tasks but for RE [1], [7]. However, it has the same entities as CoNLL-2003 but also contains relations for visualizing KGs for news articles.

3 Dataset Description

For my research, I used CoNLL04 from `huggingface_hub`, but the official dataset is available at `https://cogcomp.seas.upenn.edu/page/resource_view/43`. The CoNLL04 focuses on NLP tasks such as information extraction (NER, RE), semantic role labeling, and text classification (sentiment, topics). The provided texts from news articles contain densely distributed named entities which often appear within short textual spans (see Figure 1). Besides, there are texts with title case capitalization which complicate NER due to its mixed use of uppercase letters, for example "Security Forces Widen Ban on Possession of Arms...".

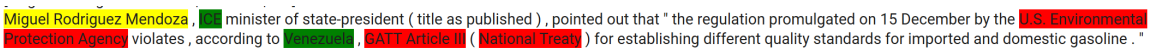

Figure 1: Entity highlighting example for CoNLL04 dataset (different colors represent different types of named entities)

The dataset has a JSON fixed structure: entities, corresponded relations and provided text. The CoNLL04 contains 1,437 sentences, each of which has at least two entities (People, Organization, Location or Other) and one relation (located in, work for, based in, live in, and kill) (see Figure 2). The dataset has already split up in train, val and test subsets (70/15/15).

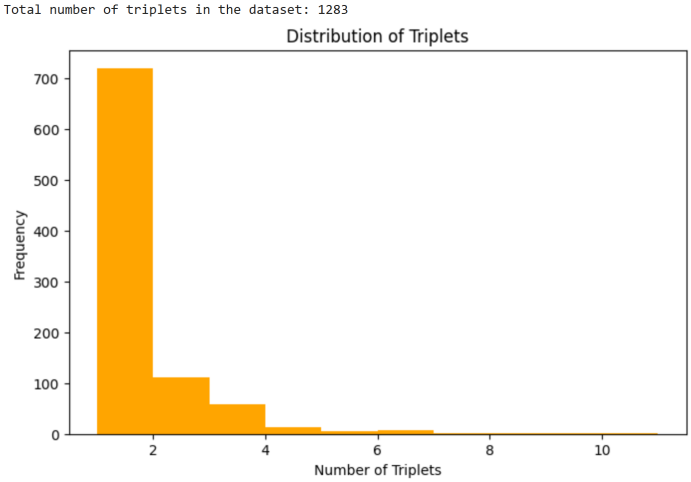

Figure 2: The distribution of triplets in train subset

The provided texts are quite short (Average Tokens Per Row is 28.77) and demonstrate an imbalanced distribution of entity classes. Figure 3 shows the predominance of Loc (Location) entities and low frequency of Org (Organization) entities in train data. This imbalance means that a model might become biased toward the majority class and perform poorly on underrepresented categories. For my research, the dataset was preprocessed by extracting bio-tags (Beginning, Inside and Outside of an entity) and part-of-speech (POS) tags (used only for CRF model).

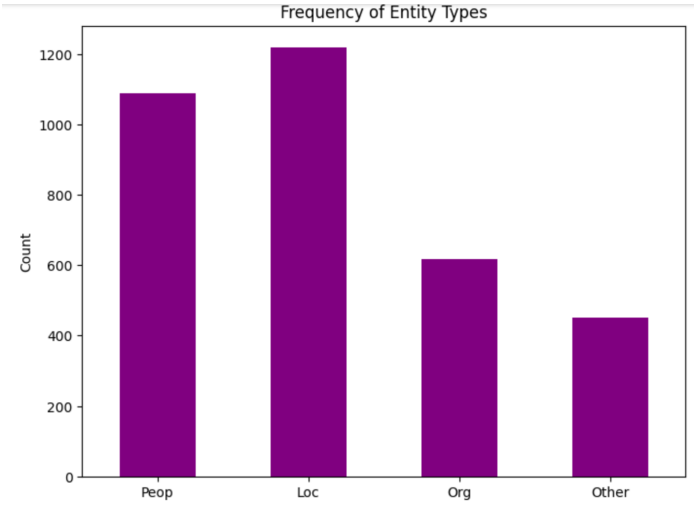

Figure 3: The frequency of named entity types in train subset

4 Traditional Approach: CRF model.

The CRF model is particularly well-suited for sequence labeling tasks such as NER. CRFs are discriminative models that directly estimate the conditional probability $P(Y|X)$, where Y represents the sequence of labels and X is the input sequence of tokens.

This approach assigns labels to each token based on its features and the features of neighboring tokens [6]. The conditional probability of a label sequence $Y = (y_1, y_2, ..., y_n)$ given an input sequence $X = (x_1, x_2, ..., x_n)$ is defined using a log-linear model:

$$P(Y|X;\theta) = \frac{1}{Z(X)} \exp\left(\sum_{i=1}^{n}\sum_{k}\theta_k^{(s)} f_k^{(s)}(y_i, x_i, i) + \sum_{i=1}^{n-1}\sum_{k}\theta_k^{(t)} f_k^{(t)}(y_i, y_{i+1}, x_i, x_{i+1}, i)\right),$$

Here:
- θ: Parameters (feature weights) learned during training
- $Z(X)$: Normalization term (partition function), ensuring probabilities sum to 1
- $f^{(s)}$: State features associated with individual positions i and labels y_i
- $f^{(t)}$: Transition features associated with transitions between labels y_i and y_{i+1}

This formulation enables the CRF to consider both local features of individual tokens and global dependencies between adjacent labels, making it effective for tasks like NER where sequential dependencies are crucial.

In my research state features were taken from the article[6]: POS, lowercase, capitalization pattern and subword information (character-level n-grams, slices of tokens recognized as prefixes and suffixes) of each token. Moreover, I decided to extract neighborhood features as contextual information to improve NER precision: lowercase and POS of previous and next tokens. I do not use dictionaries or embeddings to provide extra context in my experiment of traditional machine learning approach.

5 LLM-based Approach: BERT + CRF architecture.

My LLM-based model also builds upon a CRF framework whose inputs remain the state (contextualized embeddings, pre-trained by BERT) and transition features (BIO-tagging) provided to a model. This approach reduces reliance on manual feature engineering and enables the model to capture nuanced semantic relationships through the contextual understanding provided by the LLM. The Figure 4 shows the architecture used in this research, based on [3], excluding the co-reference resolution step.

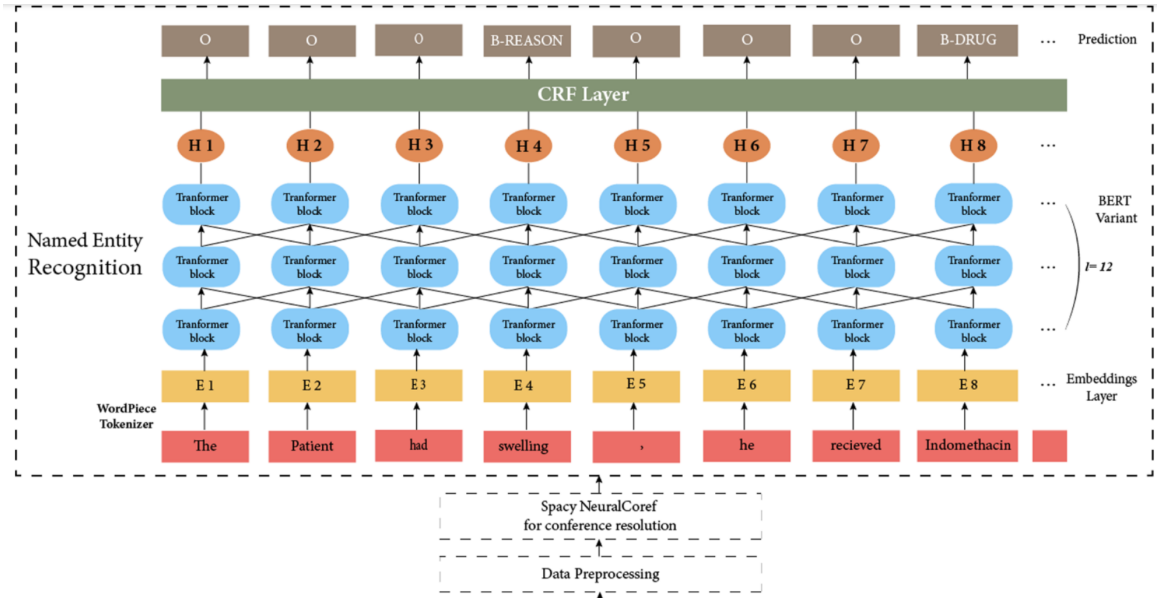

Figure 4: The BERT + CRF architecture for NER task from paper [3]

The architecture consists of 12 stacked encoder layers, where each layer includes a Self-Attention mechanism and a Feed-Forward Network. Additionally, the Multi-Head Attention layers utilize 12 attention heads to capture diverse contextual relationships within the input data. Specifically, the BERT component extracts contextual embeddings which are then processed through a linear layer (fc) and transformed into the label space (mapping BERT embeddings to BIO tagging classes). Finally, the CRF layer models dependencies between tags, ensuring that label transitions follow valid sequences (e.g., avoiding invalid transitions like 'B-Person → I-Org'). By combining LLM-based embeddings with CRF, the model achieves both contextual understanding and sequential coherence in entity recognition.

6 Experimental Results.

The traditional CRF model was implemented using $sklearn - crfsuite$, incorporating POS tags as state features to capture syntactic patterns using $spaCy$ NLP tools (en_core_web_sm-3.8.0 model). The neural model was implemented with $torch - crf$. For sequence labeling, both the CRF and BERT + CRF models utilized BIO tagging to standardize entity boundaries. BIO-tagging was applied without requiring additional libraries.

Predicted labels (BIO tags) of implemented models were evaluated using standard evaluation metrics such as Micro-F1, Macro-F1, and Weighted-Average F1 score (see Figure 5). In addition, Precision, Recall, and F1-score were calculated or each predicted BIO class (see Figure 6 and 7).

METRIC	CRF MODEL	BERT+CRF MODEL	IMPROVEMENT
Micro-F1	93.28%	95.26%	+2.1%
Macro-F1	79.90%	86.99%	+8.9%
Weighted avg	93.16%	95.25%	+2.2%

Figure 5: Comparison of CRF Model and BERT+CRF Model performance metrics

	precision	recall	f1-score	support
O	0.9692	0.9781	0.9737	6313
B-Loc	0.8682	0.8173	0.8420	427
B-Org	0.8045	0.7273	0.7639	198
I-Org	0.6920	0.7925	0.7389	241
B-Other	0.7981	0.6241	0.7004	133
I-Other	0.8155	0.6462	0.7210	130
B-Peop	0.8116	0.8318	0.8215	321
I-Peop	0.8471	0.9458	0.8937	369
I-Loc	0.8375	0.6569	0.7363	204
accuracy			0.9328	8336
macro avg	0.8271	0.7800	0.7990	8336
weighted avg	0.9323	0.9328	0.9316	8336

Figure 6: Classification report for CRF Model

Thus, the BERT+CRF model consistently provides better performance in predicting each BIO tag, especially for the I-entity class, achieving near-perfect F1-scores. Moreover, the LLM-based approach excels in handling complex entity types like Organization and Other, likely due to its contextualized embeddings from BERT. However, both models struggle slightly with less frequent classes, such as I-Loc.

By constructing KGs from existing dataset relations, we can observe that the CRF model frequently splits compound entities into fragments due to punctuation or abbreviation ambiguity, reducing precision. However, it often recognizes correct entity type classification (e.g., labeling both fragments as LOCATION in 8) The BERT+CRF model exhibited the same splitting issue but showed also one critical advantage: identified unseen entities absent from the dataset (for example extracted 'Cambodian government' in **??**), demonstrating strong generalization.

Overall, the BERT+CRF model achieves higher Micro-F1 (0.93 vs. 0.95), Macro-F1 (0.79 vs. 0.86), and Weighted-Average F1 (0.93 vs. 0.95) compared to the CRF model, indicating better general accuracy and class balance. To provide a clearer visualization of these results, a confusion matrix was generated for both models (see Figure 8 and Figure 9 a. It offers a detailed breakdown

	precision	recall	f1-score	support
O	0.9282	0.9391	0.9336	427
B-Peop	0.8152	0.8687	0.8411	198
I-Peop	0.6544	0.6692	0.6617	133
B-Org	0.9536	0.9595	0.9565	321
I-Org	0.9163	0.8703	0.8927	478
B-Loc	0.8302	0.8420	0.8361	424
I-Loc	0.7595	0.6061	0.6742	198
B-Other	0.9690	0.9704	0.9697	742
I-Other	0.9725	0.9768	0.9746	7516
accuracy			0.9509	10437
macro avg	0.8665	0.8558	0.8600	10437
weighted avg	0.9504	0.9509	0.9505	10437

Figure 7: Classification report for BERT + CRF Model

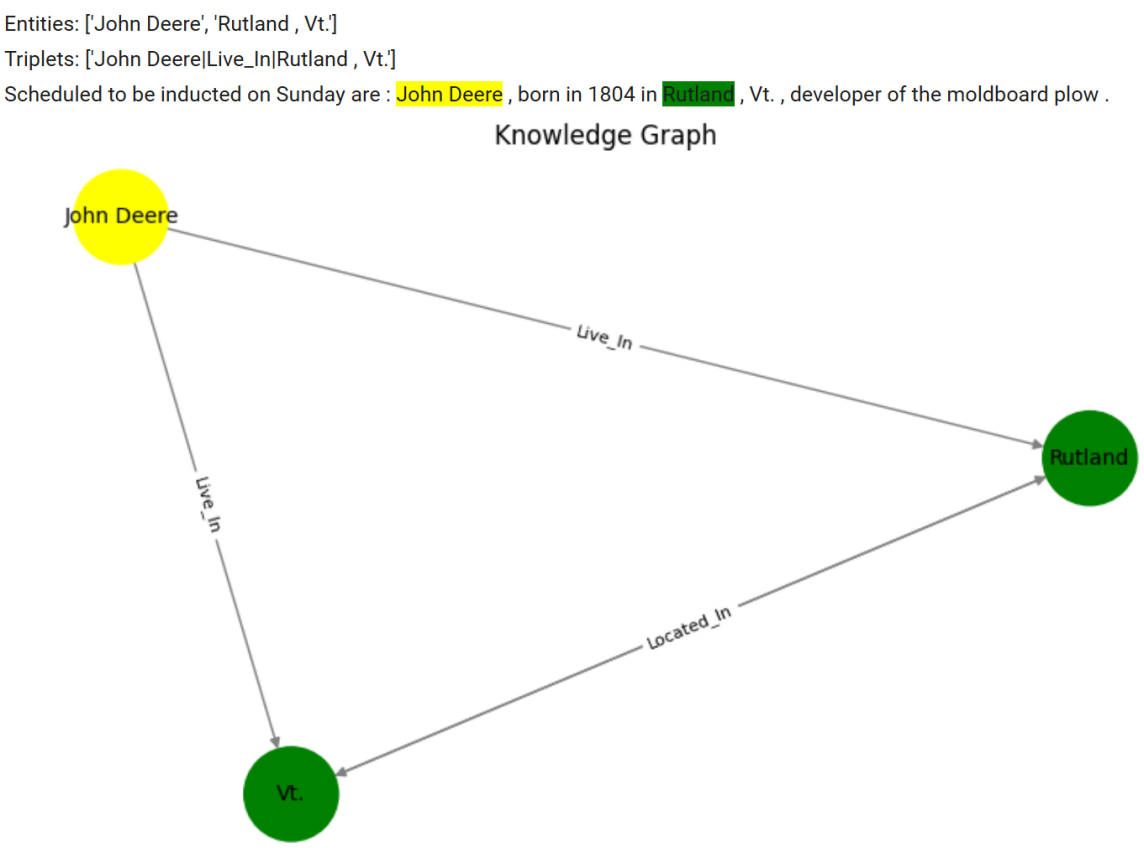

Figure 8: A knowledge graph constructing using CRF model for NER

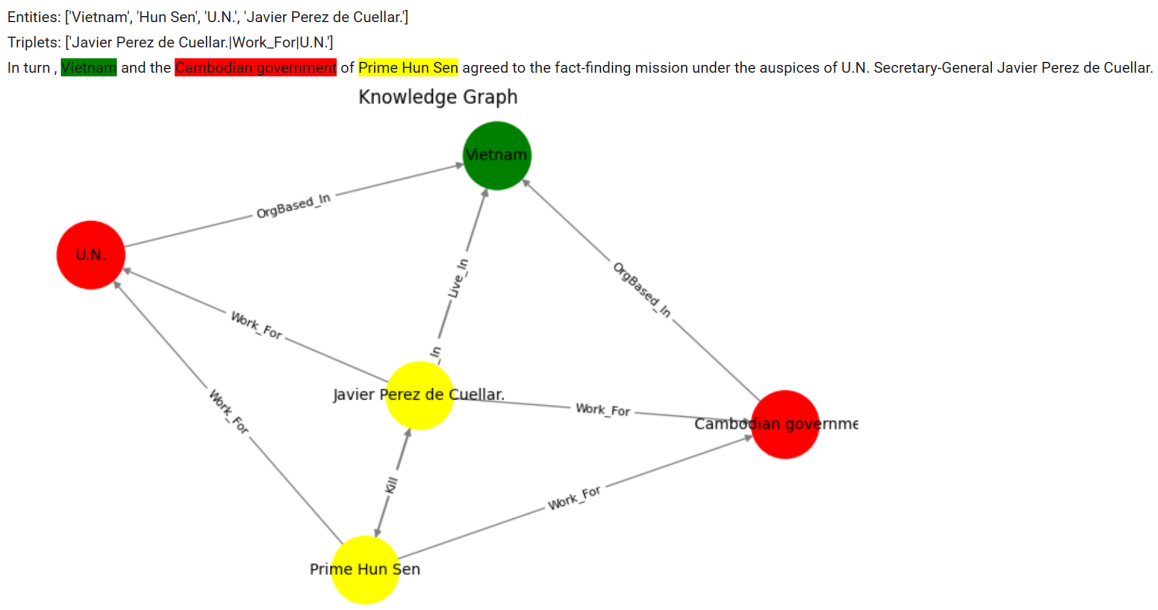

Figure 9: A knowledge graph constructing using BERT+CRF model for NER

of true positives, false positives, false negatives, and true negatives for each entity type (e.g., B-Peop, I-Loc, etc.), allowing us to observe how well the models perform on individual classes. All models and preprocessing scripts used in this study are publicly available at `https://github.com/Ulyana-Nagornaya/NER_for_KG`, ensuring full reproducibility of the experiments.

7 Conclusion

This research compares classical and LLM-based approaches for NER, showing that BERT+CRF outperforms CRF in accuracy and class balance. The results confirm the value of contextual embeddings in handling ambiguous and rare entities. Since accurate entity recognition is a key step in building structured representations of information, these findings are particularly relevant for improving Knowledge Graph (KG) construction from unstructured text across various domains.

A critical area for future improvement involves training the model to avoid fragmenting unified entities (e.g., 'Memphis , Tenn') into separate segments due to punctuation, abbreviation ambiguity, or syntactic noise. Future work may explore domain-specific fine-tuning, integration of coreference resolution, and investigating other BERT variants, such as multilingual BERT, RoBERTa, or lightweight versions like DistilBERT. These improvements could provide better performance and efficiency across diverse linguistic contexts.

References

[1] Pere-Lluís Huguet Cabot and Roberto Navigli. "REBEL: Relation extraction by end-to-end language generation". In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. 2021, pp. 2370–2381.

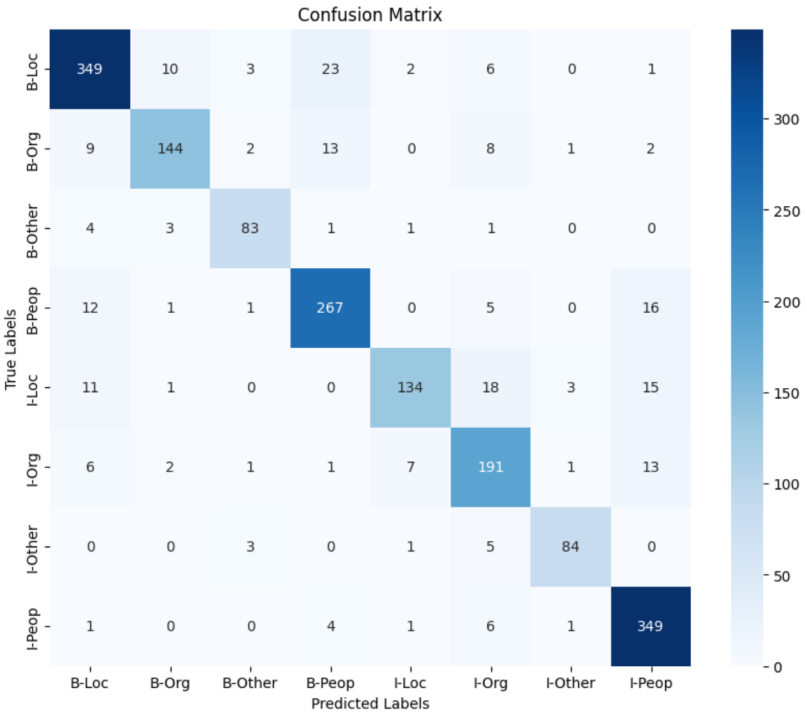

Figure 10: Confusion matrix for CRF Model

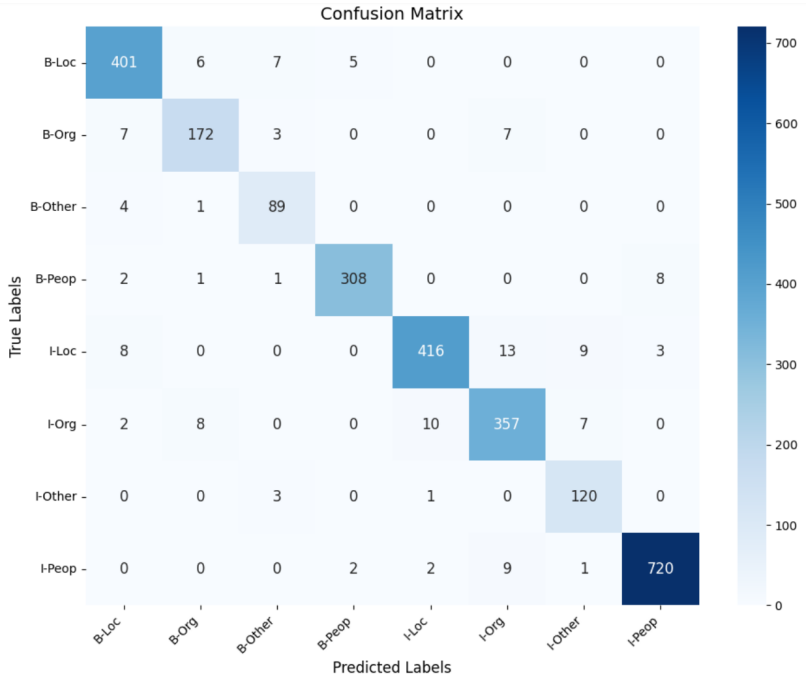

Figure 11: Confusion matrix for BERT + CRF Model

[2] Nandita Goyal and Navdeep Singh. "Named Entity Recognition and Relationship Extraction for Biomedical Text: A comprehensive survey, recent advancements, and future research directions". In: *Neurocomputing* (2024), p. 129171.

[3] Ayoub Harnoune et al. "BERT based clinical knowledge extraction for biomedical knowledge graph construction and analysis". In: *Computer Methods and Programs in Biomedicine Update* 1 (2021), p. 100042.

[4] John Lafferty, Andrew McCallum, Fernando Pereira, et al. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data". In: *Icml*. Vol. 1. 2. Williamstown, MA. 2001, p. 3.

[5] Shirui Pan et al. "Unifying large language models and knowledge graphs: A roadmap". In: *IEEE Transactions on Knowledge and Data Engineering* 36.7 (2024), pp. 3580–3599.

[6] Alexandre Passos, Vineet Kumar, and Andrew McCallum. "Lexicon infused phrase embeddings for named entity resolution". In: *arXiv preprint arXiv:1404.5367* (2014).

[7] Jue Wang and Wei Lu. "Two are better than one: Joint entity and relation extraction with table-sequence encoders". In: *arXiv preprint arXiv:2010.03851* (2020).