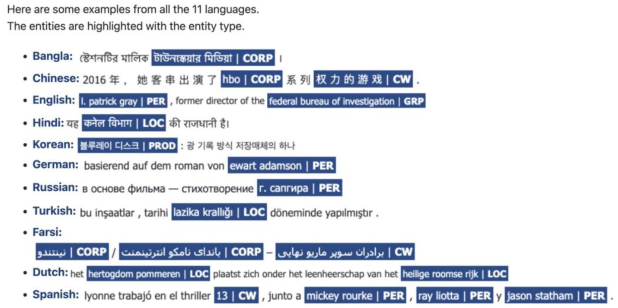
# Multilingual Complex Named Entity Recognition

***Abstract— In this project, we worked on the multilingual complex named entity recognition task in this project. We experimented with multiple baselines and improved models for this task and devised our approach based on an ensemble of architectures with a voting mechanism based on confidence scoring.***

***Keywords— named entity recognition, multilingual NER***

## I. INTRODUCTION

Understanding and identifying complex and ambiguous Named Entities (NEs) is a significant challenge in practical NLP, especially in open-domain contexts, yet it hasn't received adequate attention from researchers. NEs, such as titles of creative works (movies, books, songs, software names), pose difficulties due to their intricate nature. Unlike simple nouns, they can take linguistic forms, like imperative clauses ("Dial M for Murder"), diverging from typical NEs like personal names or locations. This inherent ambiguity complicates recognition, often requiring contextual cues for accuracy. Additionally, these titles may carry semantic ambiguity; for example, "On the Beach" could mean a preposition or a movie title. Moreover, these entities proliferate rapidly, further complicating the task, especially with the emergence of new entities.



Here are some examples from all the 11 languages.
The entities are highlighted with the entity type.

- **Bangla:** স্টেশনটির মালিক টাউনস্কোয়ার মিডিয়া | CORP ।
- **Chinese:** 2016 年， 她 客 串 出 演 了 hbo | CORP 系 列 权 力 的 游 戏 | CW .
- **English:** l. patrick gray | PER , former director of the federal bureau of investigation | GRP
- **Hindi:** यह कनेल विभाग | LOC की राजधानी है।
- **Korean:** 블루레이 디스크 | PROD : 광 기록 방식 저장매체의 하나
- **German:** basierend auf dem roman von ewart adamson | PER
- **Russian:** в основе фильма — стихотворение г. сапгира | PER
- **Turkish:** bu inşaatlar , tarihi lazika krallığı | LOC döneminde yapılmıştır .
- **Farsi:**
  نینتندو | CORP / باندای نامکو انترتینمنت | CORP – برادران سوپر ماریو نهایی | CW
- **Dutch:** het hertogdom pommeren | LOC plaatst zich onder het leenheerschap van het heilige roomse rijk | LOC
- **Spanish:** lyonne trabajó en el thriller 13 | CW , junto a mickey rourke | PER , ray liotta | PER y jason statham | PER .

Complex Multilingual Named Entity Recognition (NER) involves identifying entities within text, such as individuals, organisations, and locations. For instance, in the sentence "B.R. Ambedkar was the head of the constitutional committee of India," "B.R. Ambedkar" is recognised as a person's name, while "India" denotes a country.

This task is complex because it handles multilingual data, where the input text can be in different languages. The second is the nature of tags, which, unlike simple nouns, can be a short sentence describing the entity. This necessitates accurate NER tagging across various linguistic contexts. Addressing these challenges is crucial for developing robust NER systems that accurately identify entities across diverse language datasets.

While neural models like Transformers have achieved impressive scores on benchmark datasets such as CoNLL03/OntoNotes (Devlin et al., 2018), it's important to note, as highlighted by Augenstein et al. (2017), that these scores are often inflated due to the use of well-structured news text and the prevalence of "easy" entities like person names. Additionally, memorisation stemming from entity overlap between training and testing sets contributes to these high scores. However, these models tend to perform notably worse when faced with complex or unseen entities, as pointed out by Meng et al. (2021) and Fetahu et al. (2021). Researchers utilizing Named Entity Recognition (NER) in downstream tasks have also observed a significant number of errors attributable to NER systems struggling to identify complex entities, as documented by Luken et al. (2018) and Hanselowski et al. (2018).

## II. LITERATURE REVIEW

We came across various methodologies to solve this task during our initial study. Baseline models, like Facebook's Roberta-XML and Google's Bert-based multilingual model, followed a simple pipeline of getting the input, generating its embedding through Roberta-tokeniser or Bert-tokeniser and doing a linear classification followed by CRF (Conditional Random Field). The generated logits are then maximised to find the final tagging placement.

| Preained-model | dev f1 |
|---|---|
| BERT$_{base}$ | 0.854 |
| BERT$_{large}$ | 0.871 |
| RoBERTa$_{base}$ | 0.836 |
| RoBERTa$_{large}$ | 0.877 |
| DistilBERT$_{base}$ | 0.835 |
| BERT-WWM$_{large}$ | **0.883** |
| LUKE$_{base}$ | 0.856 |
| LUKE$_{large}$ | 0.878 |

Table: 1

While these simple techniques work extremely well for simple noun-based entities, these models suffer in multilingual scenarios because of a lack of context as well as proper attention mechanisms, which helps the model focus more on words surrounding the entity.

A popular technique followed is using an external knowledge bank [3] to infuse contextual information around the text. This knowledge bank (Wikipedia entries) can provide enough contextual information for the machine-learning model to capture the semantics of words. Another technique based on entity-aware attention tries to build attention that incorporates data from an external databank and a specialised attention mechanism trained separately. So, two separate losses are used, one cross-entropy-loss for the attention mechanism and the second for the overall named entity recognition task.

Approaches that infuse external context work better than baseline code, but the increase is only marginal and doesn't justify the inclusion of this much complexity.

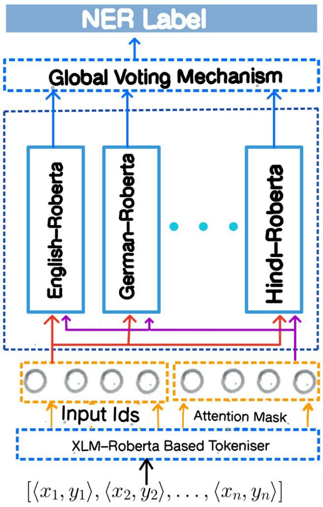
## III. Methodology

Inspired by the above results, we devised the idea to tackle the multilingual problem with an ensemble of Roberta-based models trained on different sets of languages (11 in total) and then used a voting-based confidence scoring.

The confidence scoring is based on this idea:

1. Each Roberta generates a list of logits $L_1, L_2, \ldots, L_n$.
2. We then compute the maximal score per tag per logit, **Max** $\left(L_{11}, L_{12}, \ldots, L_{1n}\right)$ and then, based on this, we return the tag.

This approach is novel in its simplicity and reduced size. This model performs well enough for given tasks.


NER Label
Global Voting Mechanism
English–Roberta
German–Roberta
Hindi–Roberta
Input Ids
Attention Mask
XLM–Roberta Based Tokeniser
$[\langle x_1, y_1\rangle, \langle x_2, y_2\rangle, \ldots, \langle x_n, y_n\rangle]$

EnsembleNER

## IV. Experiment and Results

Baseline performance evaluation of BERT (Devlin et al., 2018), BERT-WMM (Liu et al., 2020), DistilBERT (Sanh et al., 2019), Roberta (Delobelle et al., 2020), and LUKE (Yamada et al., 2020) is conducted by directly fine-tuning the MultiCoNER dataset (Malmasi et al., 2022a) on the pre-trained language models. All models are uncased. The results in Table 1 demonstrate that BERT-RoBerta outperforms the others by 1 to 5 percentage points, leading us to select BERT-WWM as our encoder. This superiority may stem from the whole-word-masking mechanism adopted by BERT-WWM. Most hyperparameters remain consistent with previous settings, notably a learning rate of 2e-5, a batch size of 32, a maximum sequence length of 100, and a maximum of 20 epochs. All performance data are based on the development set.

| Model | F1 | M-F1 |
|---|---|---|
| RoBerta(BASE) | 0.85 | 0.78 |
| BERT(BASE) | 0.83 | 0.74 |
| LUKE | 0.86 | 0.84 |
| Ours | 0.76 | 0.75 |

While we have achieved lesser results, our methodology is simple and works well.

## V. Conclusions

In conclusion, our project focused on tackling the challenge of multilingual complex Named Entity Recognition (NER). We experimented with various baseline models and devised an ensemble approach based on

Roberta-based architectures combined with a voting mechanism using confidence scoring. This novel approach simplifies the model while maintaining competitive performance.

Throughout our study, we observed the complexity inherent in NER tasks, especially in multilingual contexts where entities can vary significantly in linguistic structure and context. While neural models like Transformers have shown promise on benchmark datasets, their performance tends to suffer when dealing with complex or unseen entities. Our research underscores the importance of addressing these challenges to develop robust NER systems capable of accurately identifying entities across diverse language datasets.

Our methodology involved training an ensemble of Roberta-based models on different sets of languages and employing a voting-based confidence scoring mechanism. This approach proved effective in achieving competitive performance while simplifying the model architecture.

In evaluating various pre-trained language models, we found that BERT-WWM outperformed others by a notable margin, leading us to select it as our encoder. The superiority of BERT-WWM may be attributed to its whole-word-masking mechanism.

Moving forward, our work highlights the need for further exploration and refinement of NER techniques, particularly in multilingual settings. Future research could focus on enhancing model robustness, improving performance on complex entities, and addressing challenges related to entity overlap between training and testing sets.

Overall, our project contributes to advancing the field of multilingual complex Named Entity Recognition and underscores the importance of developing effective and scalable solutions to address real-world NLP challenges.

## REFERENCES

1. S. Malmasi, A. Fang, B. Fetahu, S. Kar, and O. Rokhlenko, “SemEval-2022 task 11: Multilingual complex named entity recognition (MultiCoNER),” in Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), G. Emerson, N. Schluter, G. Stanovsky, R. Kumar, A. Palmer, N. Schneider, S. Singh, and S. Ratan, Eds. Association for Computational Linguistics, pp. 1412–1437. [Online]. Available: https://aclanthology.org/2022.semeval-1.196
2. L. Ma, X. Jian, and X. Li, “PAI at SemEval-2022 Task 11: Named Entity Recognition with Contextualized Entity Representations and Robust Loss Functions,” in *Proc. 16th Int. Workshop Semantic Eval. (SemEval-2022)*, pp. 1665-1670, Jul. 2022.
3. I. Yamada, A. Asai, H. Shindo, H. Takeda, and Y. Matsumoto, “LUKE: Deep contextualized entity representations with entity-aware self-attention.” [Online]. Available: http://arxiv.org/abs/2010.01057