International Journal of Computer Vision and Signal Processing, 15(1), 60-66(2025) ORIGINAL ARTICLE

An Ensemble Approach to Named Entity Recognition for Bangla-English Code-Switched Texts Using XLM-R, BiLSTM-CRF, and CRF

Sultana Tasnim Jahan*, Rashed Mustafa

Department of Computer Science and Engineering
University of Chittagong, Bangladesh

Abstract

Named Entity Recognition (NER) in Bangla-English code-switched text is quite challenging due to tokenization irregularities, linguistic variances, and the absence of annotated datasets. Current NER algorithms focus mainly on monolingual or organized multilingual texts, with little attention paid to code-switched data. The efficacy of conventional methods is further constrained by the difficulties in managing grammatical errors, unclear entity boundaries, and tokenization issues. This paper suggests an ensemble-based NER method that combines the CRF, BiLSTM-CRF, and XLM-R models to solve this problem. Entity labels are predicted independently by each model, and a fourth component—an ensemble model that combines the three – is also present. A majority vote process among these four sources determines the final projections. By combining sequence-labeling techniques with transformer-based contextual embeddings, this hybrid approach enhances generalization and lowers recognition mistakes. Our work shows empirical improvements through extensive experimentation and architectural integration and, in contrast to usual surveys, presents a comprehensive, functional pipeline. The results of experiments show that the suggested method greatly enhances entity detection in intricate, code-switched texts by achieving more accuracy and robustness when compared to individual models. The study has significant applications in social media analysis, customer support automation, and multilingual information extraction, all of which depend on the ability to handle mixed-language text. This study confirms performance using both comparison benchmarks and real-case sentence structures from the newly released dataset, in contrast to previous work that assessed code-switched NER separately. We intend to investigate self-learning strategies for domain adaptability, add more varied linguistic patterns, and enlarge the dataset in subsequent research. Furthermore, the performance of NER in code-mixed, low-resource environments may be further improved by using meta-learning techniques and adaptive fine-tuning.

Contribution of the Paper: A newly created Bangla-English code-switched NER dataset is presented in this research along with a novel ensemble-based approach that combines CRF, BiLSTM-CRF, XLM-R, and an ensemble of these models under a single majority voting framework. Accuracy and stability are improved across linguistically inconsistent code-switched texts by combining syntactic, sequential, and contextual modeling.

Keywords: Named Entity Recognition (NER), Bangla-English code-switched text, CRF, BiLSTM-CRF, XLM-R, Ensemble

IJCVSP

ISSN: 2186-1390 (Online)
http://cennser.org/IJCVSP

Article History:
Received: 21/3/2025
Revised: 10/7/2025
Accepted: 1/11/2025
Published Online: 23/11/2025

*Corresponding author
Email addresses: tasnim047sultana@gmail.com (Sultana Tasnim Jahan), rashed.m@cu.ac.bd (Rashed Mustafa)

1. INTRODUCTION

Traditional NER models usually fall short in areas where informal communication (such as social media postings, customer support chats, and conversational AI) regularly combines Bangla and English, mostly because tokenization is irregular and there aren't strict syntactic rules when switching languages.

In this study, a voting-based ensemble of CRF, BiLSTM-CRF, and XLM-R is used for the first time to perform Named Entity Recognition (NER) on data that has been code-switched between Bangla and English. The goal of this ensemble strategy is to increase robustness in noisy informal texts by utilizing both deep contextual embeddings and conventional sequence labeling.

Other motivation behind these:

Increasing Code-Switched Usage: Because Bangla and English are frequently mixed in informal conversations and social media sites, NER is crucial for comprehending the context and purpose of the user.

Real-World Applications: For multilingual users, enhanced NER can enhance search engines, chatbots, and sentiment analysis.

Existing Gaps in Research: Prior research on NER concentrated on English and Bangla as individual languages, making significant progress in monolingual settings.

Furthermore, traditional natural language processing (NLP) models encounter considerable difficulties in comprehending and evaluating such mixed-language data due to the growing frequency of Bangla-English code-switched messages in social media, online communication, and other digital platforms.

Nevertheless, there are still no reliable solutions, especially to deal with the noisy and erratic structure of texts that have been code-switched between Bangla and English. This creates new difficulties in preserving context and settling unclear entity boundaries.

By creating a hybrid framework that combines several models to better handle irregularities, enhance context preservation, and lessen error propagation, this paper closes that gap. The approach increases accuracy while lowering dependence on the shortcomings of any one model by using majority voting among models. By combining the advantages of separate models by majority vote, the ensemble seeks to enhance generalization on mixed-language inputs and lessen error propagation from token misalignment.

2. RELATED WORKS

A crucial field of study in Natural Language Processing (NLP), Named Entity Recognition (NER) has undergone substantial development throughout time. Figure 1 illustrates the steadily rising number of published publications on NER and code-switched text processing, underscoring the field's increasing significance.

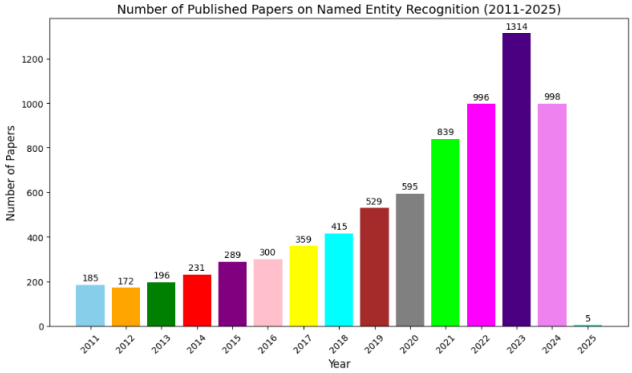


Figure 1: Bar Diagram Representing Number of Published Papers on Named Entity Recognition from 2011 to 2025 Based on the Data acquired through IEEE Xplore Number of published paper, Years (2011 to 2025)

Named Entity Recognition (NER) has evolved from statistical and rule-based models like CRF and HMM to sophisticated deep learning techniques like transformers and BiLSTM-CRF. Even with increased precision and generalization, the majority of current research still concentrates on monolingual data, paying little attention to texts that are multilingual or code-switched. In comparison to other language pairs like Hindi-English or Tamil-English, Bangla-English code-switching is still not well studied. This section examines pertinent research, emphasizing methods, findings from experiments, and major issues in code-mixed NER.

2.1. Traditional Machine Learning and CRF-Based Approaches:

Conventional Machine Learning and Methods Based on CRF Rule-based and machine learning methods, like Conditional Random Fields (CRF) and Hidden Markov Models (HMM), were among the first methods for NER. CRF and LSTM models were used by Singh and Vijay [1] to identify things in social media text that were code-mixed in Hindi and English. Their model's remarkable F1-score of 0.95 shows how well deep learning and statistical sequence models work together. The study did not examine Bangla-English data, which has particular syntactic and morphological difficulties, and was restricted to Hindi-English code-switching.

2.2. Methods Based on Neural Networks:

By lowering the dependency on manually created features and rule-based techniques, deep learning has greatly improved NER. Long-range relationships within sequences might now be captured by models thanks to the development of recurrent neural networks (RNNs) and long short-term memory (LSTM) networks.

By combining BERT, BiLSTM, and CRF, Dai et al. [2] suggested an NER system for Chinese Electronic Health

Records (EHRs). Their findings demonstrated a significant improvement in medical entity recognition thanks to BERT's pre-trained contextual representations. However, the model was not suitable for multilingual contexts such as Bangla-English since it was restricted to Chinese text and did not handle code-mixed data.

2.3. Transformer-Based NER Methods:

By enabling self-attention techniques that capture contextual linkages across lengthy sequences, transformers transformed natural language processing. By utilizing deep contextual embeddings, Devlin et al.'s BERT (Bidirectional Encoder Representations from Transformers) [3] greatly enhanced NER performance. However, because traditional BERT models are mainly trained on monolingual datasets, they have trouble with code-mixed text. Support vector machines (SVM) [4], conditional random fields (CRF) [5], structured support vector machines (SSVM) [6], recurrent neural networks (RNN) [7], and convolutional neural networks (CNN) and their variation models [8] are the most effective methods for identifying named items. The proposed study developed a method for performing CNER based on BiLSTM and CRF.

The research also goes into domain-specific NER models, which are designed to address the unique difficulties of certain domains. Models like [9], which was painstakingly created for legal and financial documents, and BioBERT[10], which was optimized for the complex field of medical NER, are prime examples of how NER approaches may be tailored to meet the needs of different contexts. This study explores reinforcement learning[11], going beyond conventional methods and revealing the potential of Gaussian prior[12] and distantly supervised NER techniques[13].

Large-scale models like GPT-3 [14] and PaLM [15] show good few-shot performance with little task-specific data, whereas recent developments like E-NER [16] add uncertainty-aware loss techniques to increase trustworthiness in NER tasks. These advancements reinforce our hybrid architectural approach by demonstrating the increasing applicability of large language models (LLMs) in multilingual and low-resource contexts such as code-switching between Bangla and English. The study also examines methods for optical character recognition (OCR) [17].

3. PROPOSED METHOD

3.1. Premises and Justification

The linguistic intricacy of code-switching creates special hurdles for Named Entity Recognition (NER) in Bangla-English code-switched literature. Sentences that contain terms from several languages in one utterance are referred to as code-switched texts.

3.2. Model of the System

Every input sentence is initially tokenized into distinct tokens in our suggested system. Following that, these tokens are concurrently run through four distinct models: XLM-R, BiLSTM-CRF, CRF, and our recently created ensemble model, which integrates the predictions of the previous three models.

By using a voting process, the ensemble model is intended to capitalize on the advantages of each unique model. However, because of its overwhelming presence in the dataset, we found that the ensemble model frequently forecasts the label "O"—which stands for non-entity—incorrectly. When real named entities are mistakenly classified as non-entities, this can result in a significant number of false negatives. To address this issue, we designed a fallback mechanism:

- A label other than "O" is chosen as the final output if that is what the ensemble model predicts.
- We next verify the predictions of the three base models (XLM-R, BiLSTM-CRF, and CRF) whether the ensemble predicts "O":
  - We declare a non-"O" label to be the majority and choose it as the final label if at least two models concur on it.
  - We use a predetermined priority criterion depending on each model's performance if the three models predict different labels: $BiLSTM\text{-}CRF > CRF > XLM\text{-}R$. The final output is the label with the highest importance out of the three.

With this method, we can lessen the issue of class imbalance and make sure that entity tokens aren't mistakenly suppressed by an ensemble "O" label prediction.

3.3. Preparing the Dataset

Bangla-English code-switched text with entity label annotations makes up the dataset used in this study. Person (PER), Location (LOC), Organization (ORG), Facility (FAC), Date (DATE), Time (TIME), Money (MONEY), Percentage (PERCENT), Quantity (QUANTITY), Product (PRODUCT), Miscellaneous (MISC), and O (non-entity tokens) are the twelve classes into which the entities are divided.

Steps in Data Processing:

1. Text Tokenization: Each sentence is divided into tokens while maintaining structural consistency.
2. Entity Annotation: The appropriate entity class is explicitly assigned to each token.
3. Data Splitting: Three subsets of the dataset are created:
   - For training: Individual models are trained using the training set (80%).

- For validation: Model hyperparameters are optimized using the validation set (10%).
- For test: Evaluation Set (10%): For the last assessment.

4. Managing Code-Switching: To maintain consistency, specific preprocessing methods like transliteration and language identification are used because some tokens may be in Bangla and others in English.

3.4. Majority Voting Mechanism for Ensemble Learning

On the training data, each model (XLM-R, BiLSTM-CRF, and CRF) is trained separately. For every token in a phrase:

1. A label for an entity is predicted by all three models.
2. A majority vote decides the ultimate label.
3. A priority order is used in the event of a tie (i.e., each model makes a different prediction) according to their contextual learning capacities:
   Ensemble model > BiLSTM-CRF > CRF > XLM-R.

Figure 2 provides a visual representation of this procedure, showing the entire ensemble architecture from ensemble prediction to model training.

3.5. Illustrative Diagrams

The following diagrams are presented to provide a clearer illustration of the suggested methodology:

- Data flow from preprocessing to ensemble prediction is shown in the system workflow diagram.
- XLM-R, BiLSTM-CRF, and CRF architectures are represented visually in model architecture diagrams.
- Voting Mechanism Flowchart: Outlining the process of combining entity predictions from many models.

3.6. Probability Estimation for Conditional Random Fields (CRF)

Given an input sequence, the likelihood of a label sequence is as follows:

The probability of a label sequence given an input sequence is defined as

$$\frac{\exp\left(\sum_t \mathbf{W}_{y_t,x_t} + \mathbf{T}_{y_{t-1},y_t}\right)}{\sum_{\mathbf{y}'} \exp\left(\sum_t \mathbf{W}_{y'_t,x_t} + \mathbf{T}_{y'_{t-1},y'_t}\right)}$$

where:

- $\mathbf{W}$ denotes learned model parameters (token features),
- $\mathbf{T}$ represents the transition matrix enforcing label consistency,
- y_t is the label at position t, and
- The denominator sums over all possible label sequences $\mathbf{y}'$.

3.7. Representation of BiLSTM Hidden States

At any given time, the hidden state is provided here: The hidden state at time t is given by:

$$\mathbf{h}_t = \tanh(\mathbf{W}_{ih}\mathbf{x}_t + \mathbf{b}_{ih} + \mathbf{W}_{hh}\mathbf{h}_{t-1} + \mathbf{b}_{hh})$$

where, $\mathbf{W}_{ih}$ and $\mathbf{W}_{hh}$ are input and hidden weights, and $\mathbf{b}_{ih}$, $\mathbf{b}_{hh}$ are biases.

When biases are present, along with input and hidden weights.

3.4.3 Rule of Majority Voting A token's final entity label is calculated as follows: The final entity label E_i for a token is computed as

$$E_i = \text{mode}([E_{\text{XLM-R}},\ E_{\text{BiLSTM-CRF}},\ E_{\text{CRF}}])$$

3.8. Algorithm Formulation

Algorithm 1 Ensemble-based Entity Label Prediction with Fallback Logic

Require: Tokenized sentence $S = \{w_1, w_2, \ldots, w_n\}$
Ensure: Entity labels $L = \{l_1, l_2, \ldots, l_n\}$

```
for each token w_i ∈ S do
    Obtain predictions:
        Ensemble model → E_Ens
        XLM-R → E_XLMR
        BiLSTM-CRF → E_BiLSTM
        CRF → E_CRF
    if E_Ens ≠ "O" then
        l_i ← E_Ens {Use ensemble if not "O"}
    else
        Votes ← {E_XLMR, E_BiLSTM, E_CRF}
        Count occurrences of each label in Votes
        if a majority label exists (appears ≥ 2 times) then
            l_i ← MajorityLabel
        else
            if E_BiLSTM ≠ "O" then
                l_i ← E_BiLSTM
            else if E_CRF ≠ "O" then
                l_i ← E_CRF
            else
                l_i ← E_XLMR
            end if
        end if
    end if
end for
return L = {l_1, l_2, ..., l_n}
```

3.9. Evaluation Metrics:

Equation of evaluation matrix given here:

Performance is assessed using precision, recall, F1-score, and the confusion matrix.

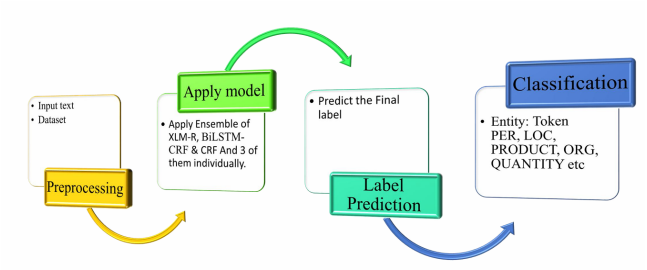


Figure 2: Concept of my methodology.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The confusion matrix provides insights into entity classification accuracy.

Dataset: A manually labeled dataset that has been code-switched between Bangla and English is used. 80%-10%-10% is the split between training, validation, and testing. Hardware: TensorFlow/PyTorch framework, T4 GPU.

The pipeline and a section of the annotated dataset are further depicted in Figure 3. The end-to-end flow from preprocessing to ensemble prediction is demonstrated, along with actual instances of mixed Bangla-English tokens.

A thorough schematic of our process is shown in Figure 4, which shows how each model is trained separately, makes predictions at the token level, and contributes to the final ensemble output. Figure 7 provides comparative visualizations: classification metrics, dataset distribution, and accuracy across models.

Here, included my technique figure that details each stage.

The flow of data in the NER system:

Input: Text with a code switch between Bangla and English.

Preprocessing includes noise reduction, normalization, and tokenization.

Model Training: Separate training is done for XLM-R, BiLSTM-CRF, and CRF.

Ensemble Decision: The final entity classification is chosen by majority vote.

Output: Sentences with entity labels annotated. A portion of my created dataset is shown in figure 7. Here, both Bangla and English mixed words are present.

To confirm the efficacy of the suggested Named Entity Recognition (NER) system for texts with code switches between Bangla and English, the experimental evaluation is essential. The performance of the model is thoroughly examined in this chapter, which also covers experimental

2224	416	last	O
2225	416	week	DATE
2226	417	we	O
2227	417	will	O
2228	417	go	O
2229	417	to	O
2230	417	Canada	LOC
2231	417	next	O
2232	417	month	DATE
2233	418	the	O
2234	418	project	O
2235	418	was	O
2236	418	completed	O
2237	418	on	O
2238	418	১২ই ডিসেম্বর	DATE
2239	419	the	O
2240	419	gift	PRODUCT
2241	419	cost	O
2242	419	২৫০০৳	MONEY

Figure 3: A portion of my created dataset.

conditions, dataset properties, parameter selection, accuracy measurements, and comparison analysis.

Accuracy, precision, recall, and F1-score are among the quantitative performance indicators that are thoroughly examined to evaluate each model's advantages and disadvantages. Misclassifications, model behavior under various entity categories, and enhancements made possible by the ensemble approach are also covered in this chapter. The contributions of this work are highlighted by comparing the results with previous research, which further validates them.

Furthermore, several visualizations are used to show

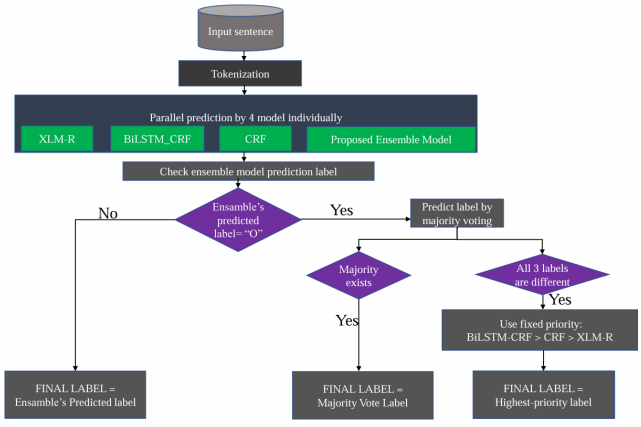


Figure 4: Proposed methodology

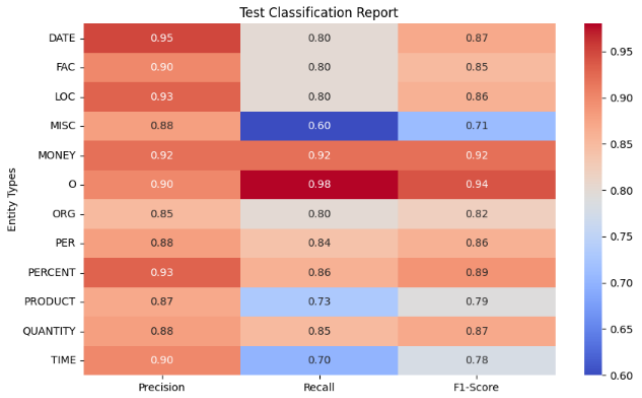
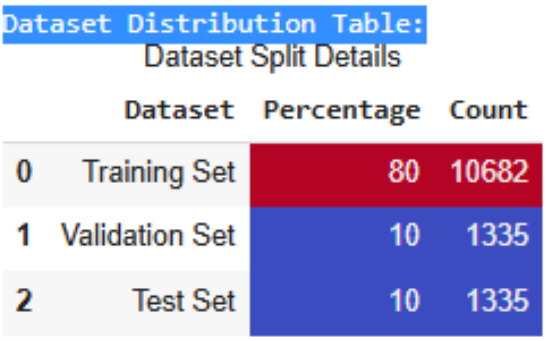


Figure 5: Classification Report



Dataset Distribution Table:

Dataset Split Details

	Dataset	Percentage	Count
0	Training Set	80	10682
1	Validation Set	10	1335
2	Test Set	10	1335

Figure 6: Dataset distribution

the performance patterns, such as classification reports, confusion matrices, and accuracy comparison graphs.

Figure 5: Classification Report for Different Models

Figure 6: Dataset distribution table here:

Based on the findings, we note the following:

- Because of its high precision score, XLM-R produces fewer false positive predictions. This implies that entity and non-entity words are successfully distinguished by the transformer-based approach.
- While BiLSTM-CRF is better at collecting more real entities than XLM-R, it occasionally misclassifies non-entities as entities due to its higher recall.
- CRF alone performs worse in terms of precision and recall, indicating that it has trouble correctly classifying items in the code-switched dataset.
- The ensemble model achieves the optimal balance between precision and recall, outperforming all individual models and resulting in the highest F1 score.

The accuracy comparison table is presented in Figure 7.

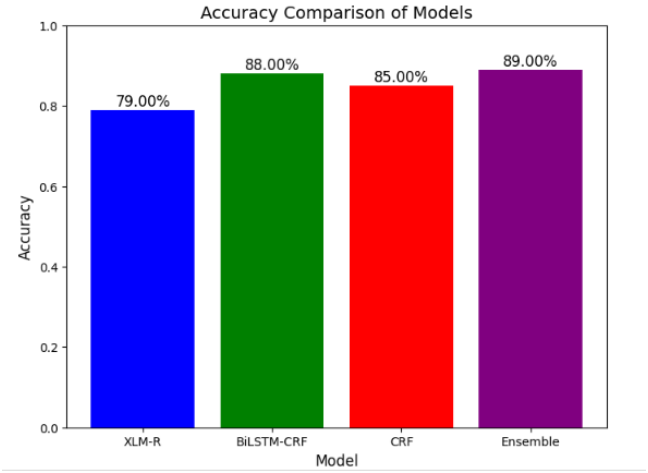


Figure 7: Accuracy comparison table

The accuracy attained by each model and the suggested ensemble technique are shown in Figure 7. Despite being effective in multilingual settings, the XLM-R model had the lowest accuracy (79%) in code-switched Bangla-English text, most likely as a result of domain-specific variances and poor contextual adaptation. With 85% accuracy, traditional CRF did mediocrely well; it benefited from handcrafted characteristics but lacked in-depth contextual knowledge.

With an accuracy of 88%, the BiLSTM-CRF model beat both XLM-R and CRF by skillfully utilizing contextual patterns and sequential dependencies. Nevertheless, the ensemble model, which employs a majority voting method to integrate the advantages of XLM-R, BiLSTM-CRF, and CRF, obtained the greatest accuracy of 89%, suggesting enhanced generalization and resilience. This demonstrates that combining several model predictions lessens the biases of each model and better manages the complexity of code-switching.

4. CONCLUSIONS

In this paper, we used the XLM-R, BiLSTM-CRF, and Conditional Random Fields (CRF) models in conjunction with a majority voting ensemble strategy to tackle the problem of Named Entity Recognition (NER) in Bangla-English Code-Switched Texts. The main objective was to improve the accuracy of entity recognition in code-mixed data, where contextual ambiguities and language alternation cause typical monolingual NER models to perform poorly.

Among our key contributions are

- A new voting-based ensemble model specifically designed for code-switched NER between Bangla and English is proposed.

- A majority vote method is used to resolve prediction inaccuracies at the token level.
- obtaining better performance than separate models.
- offering a benchmark dataset that has been manually annotated to aid in upcoming code-switched NER research.

Directions for the future work are:

- Domain adaptation is the process of applying the model to data that is peculiar to a given domain, like financial, medical, or legal materials.
- Expanding to additional language pairs, such as Hindi-English and Chinese-English, is known as multilingual scalability.
- Transformer Improvements: Including adapters for language-specific transformers, such as mBERT or BanglaBERT.
- Low-Resource Strategies: Investigating prompt-based or semi-supervised learning techniques to enhance performance in low-resource environments.
- Dataset Expansion: To enhance model generalization, a bigger, more varied Bangla-English code-switched dataset should be curated.

This study shows that NER performance in code-switched scenarios is greatly enhanced by merging transformer-based models with sequential tagging models via majority voting. It establishes the groundwork for multilingual, scalable NER solutions in linguistically varied environments.

References

[1] G. Singh, T. Vijay, Named entity recognition for hindi-english code-mixed social media text, ACL Anthology.

[2] Z. Dai, X. Wang, P. Ni, Y. Li, G. Li, X. Bai, Named entity recognition using bert bilstm crf for chinese electronic health records, IEEE.

[3] J. Devlin, et al., Bert: Pre-training of deep bidirectional transformers for language understanding, NAACL-HLT.

[4] M. Asahara, Y. Matsumoto, Japanese named entity extraction with redundant morphological analysis, in: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, Vol. 1, Association for Computational Linguistics, 2003, pp. 8–15.

[5] A. McCallum, W. Li, Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons, in: Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4, Association for Computational Linguistics, 2003, pp. 188–191.

[6] Y.-J. Lee, O. L. Mangasarian, Ssvm: A smooth support vector machine for classification, Computational Optimization and Applications 20 (1) (2001) 5–22.

[7] Z. Huang, W. Xu, K. Yu, Bidirectional lstm-crf models for sequence tagging, arXiv preprint arXiv:1508.01991.

[8] E. Strubell, et al., Fast and accurate entity recognition with iterated dilated convolutions, arXiv preprint arXiv:1702.02098.

[9] W. Lin, Q. Gao, L. Sun, Z. Zhong, K. Hu, Q. Ren, Q. Huo, Vibertgrid: A jointly trained multi-modal 2d document representation for key information extraction from documents, in: Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part I 16, Springer, 2021, pp. 548–563.

[10] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (4) (2020) 1234–1240.

[11] D. A. Permatasari, D. A. Maharani, Combination of natural language understanding and reinforcement learning for booking bot, Journal of Electrical, Electronic, Information, and Communication Technology 3 (1) (2021) 12–17.

[12] Y. Yang, X. Hu, F. Ma, A. Liu, L. Wen, S. Y. Philip, Gaussian prior reinforcement learning for nested named entity recognition, in: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2023, pp. 1–5.

[13] L. Ding, T.-Y. Huang, H. Liu, Y. Wang, Z. Zhang, Distantly supervised named entity recognition with category-oriented confidence calibration, in: International Conference on Asian Digital Libraries, Springer, 2022, pp. 46–55.

[14] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[15] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, et al., Palm: Scaling language modeling with pathways, arXiv preprint arXiv:2204.02311.

[16] Z. Zhang, M. Hu, S. Zhao, M. Huang, H. Wang, L. Liu, Z. Zhang, Z. Liu, B. Wu, E-ner: Evidential deep learning for trustworthy named entity recognition, arXiv preprint arXiv:2305.17854.

[17] P. Bhatia, B. Celikkaya, M. Khalilia, S. Senthivel, Comprehend medical: A named entity recognition and relationship extraction web service, in: 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA), IEEE, 2019, pp. 1844–1851.