BTP: Temporal Entity Extraction using LLM API & Cross-Lingual NER

Submitted By:

Samarth Rathore
22075073
Computer Science and Engineering (B.Tech)

Submitted To:

Prof. Anil Kumar Singh

Project Overview: Tackling Low-Resource NLP Challenges

This project addresses critical gaps in Natural Language Processing for low-resource scenarios, specifically focusing on two challenging areas:

Task 1: Temporal Entity Extraction

Automating the identification of temporal expressions and events from English text.

Task 2: Cross-Lingual NER Dataset Generation and Model Training & Evaluation

Creating Named Entity Recognition datasets for under-resourced Indic languages by leveraging Hindi as a pivot and fine-tuned XLM-R and evlauted it on the data.

Task 3: Hindi Temporal Entity Recognition

Developing robust model for temporal entity recognition within Hindi texts.

Our innovative approach combines advanced Large Language Model (LLM) prompting techniques with multilingual fine-tuning to achieve state-of-the-art results.

Task 1: English Temporal Entity Extraction

Our primary goal was to accurately extract TIMEX3 and EVENT entities from the established TempEval-3 dataset, comprising diverse English news sources.

Dataset & Methodology

- **Dataset:** Comprehensive English TimeBank TempEval3 dataset.
- **Approach:** We explored four distinct prompting strategies, refining our method to an optimised decomposed and filtered dependency approach:

- Baseline prompting
- Dummy tag prompting
- Dependency-based prompting
- **Final optimized approach (decomposed + filtered dependencies)**

Key Performance Indicators

Our final optimised approach yielded impressive results:

0.94

Precision

High accuracy in identifying relevant temporal entities.

0.81

Recall

Effectively capturing a significant portion of all temporal entities.

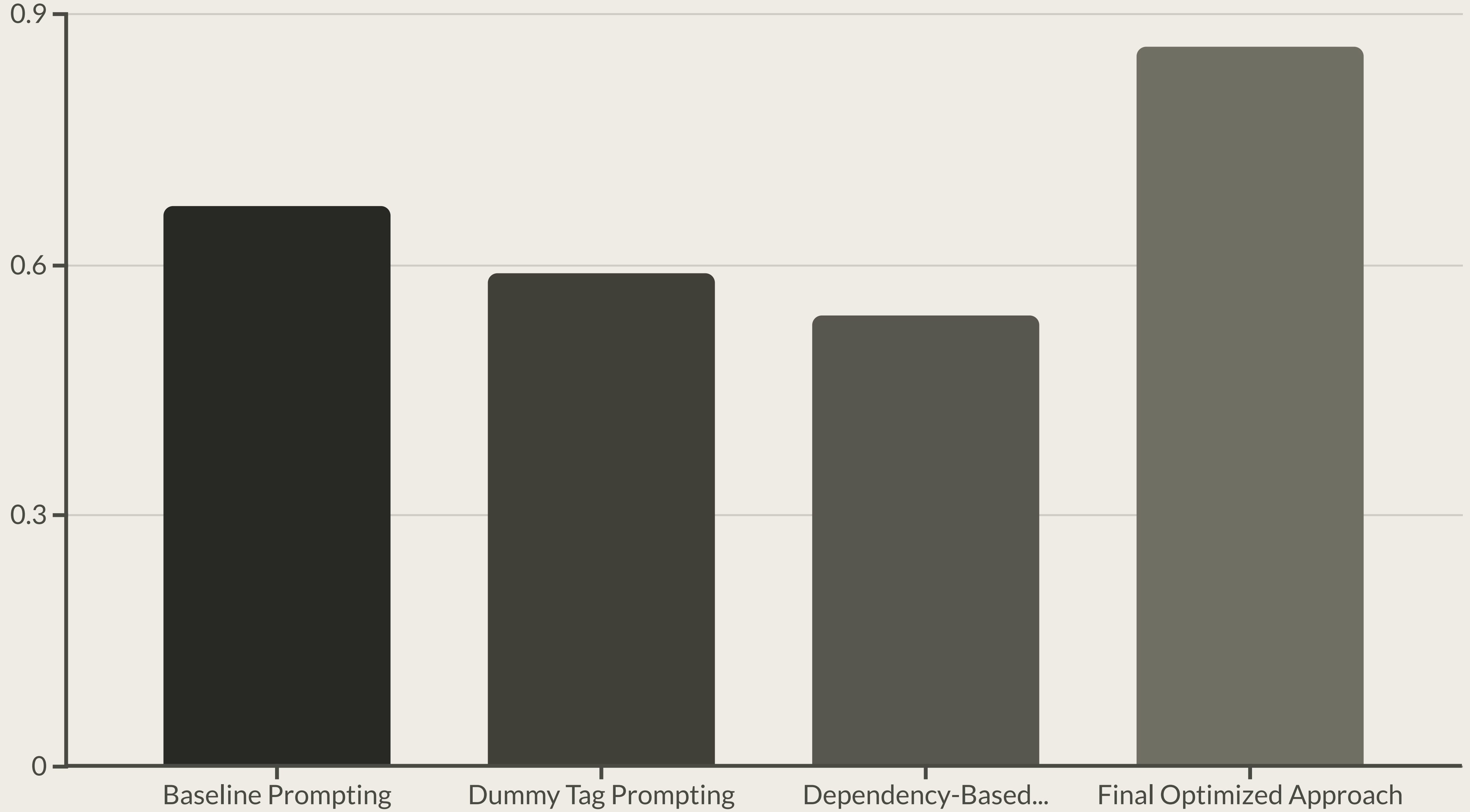
0.86

F1 Score

Balancing precision and recall, matching PTime (0.87) and surpassing HeidelTime/SUTime.

Temporal Extraction: Comparative Analysis of LLM Prompting Strategies

A detailed comparison of F1 scores across different prompting approaches highlights the significant impact of our optimized strategy.



Conclusion: Our refined LLM prompting, particularly with the decomposed and filtered dependencies, demonstrates that near-state-of-the-art performance can be achieved without the intensive effort of rule engineering, marking a significant advancement in temporal entity extraction.

Task 2: Cross-Lingual NER Dataset Generation for Indic Languages

Our goal was to create high-quality NER datasets for seven under-resourced Indic languages, using Hindi as a pivotal language for projection.

Target Languages

We focused on Assamese (as), Bengali (bn), Gujarati (gu), Malayalam (ml), Marathi (mr), Tamil (ta), and Telugu (te).

Labels

Standard Named Entity Recognition labels were used: Person (PER), Location (LOC), and Organisation (ORG), formatted in BIO (Beginning, Inside, Outside) tagging scheme.

Generation Pipeline

The process involved word alignment (using SimAlign), followed by sophisticated Hindi to Target language tag projection, and concluded with a rigorous 3-method evaluation for validation.

Cross-Lingual Dataset Quality & Reliability

Rigorous evaluation confirmed the high quality and reliability of the generated cross-lingual NER datasets for Indic languages.

0.757

Entity Matching

High agreement on identified entities across languages.

0.916

Tag Consistency

Strong consistency in NER tag assignments.

0.82

Token-level F1

Excellent performance at the token level, validating the dataset's accuracy.

> **Conclusion:** The substantial Entity Matching, Tag Consistency, and Token-level F1 scores unequivocally demonstrate the high quality of the generated datasets and the reliability of our cross-lingual projection methodology. This paves the way for effective NER model training in these low-resource languages.

Multilingual NER Model Training: XLM-R Base

Our multilingual Named Entity Recognition model was trained on the newly generated datasets for the seven target Indic languages, leveraging the powerful XLM-R Base architecture.

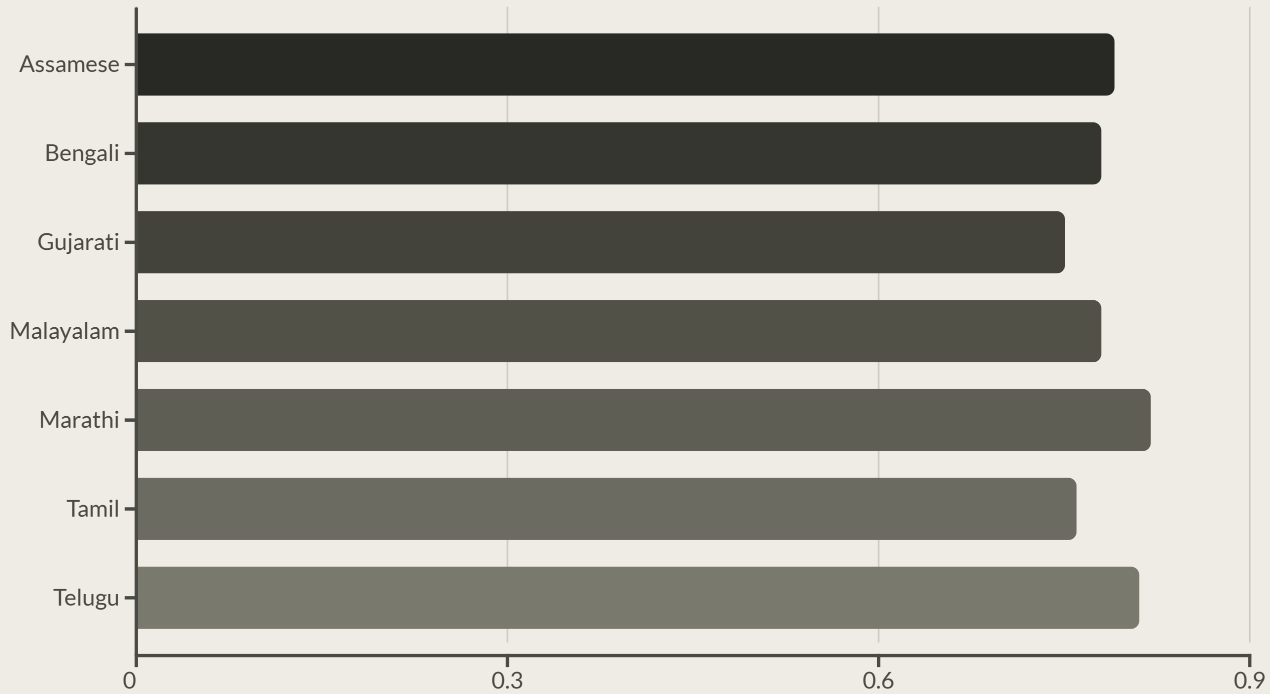
Model & Training Parameters

- **Model:** XLM-R Base
- **Task:** Token Classification (BIO scheme)
- **Epochs:** 10
- **Learning Rate:** Optimised from 2e-5 to 3e-5

Overall Performance

- **Overall F1:** 0.7891
- **Accuracy:** 97.97%

Per-Language F1 Scores



Task 3: Hindi TimeBank Temporal Extraction

We extended our temporal extraction capabilities to Hindi, developing a model specifically for identifying EVENT, STATE, and TIMEX entities within Hindi TimeBank data.

Dataset & Overall Performance

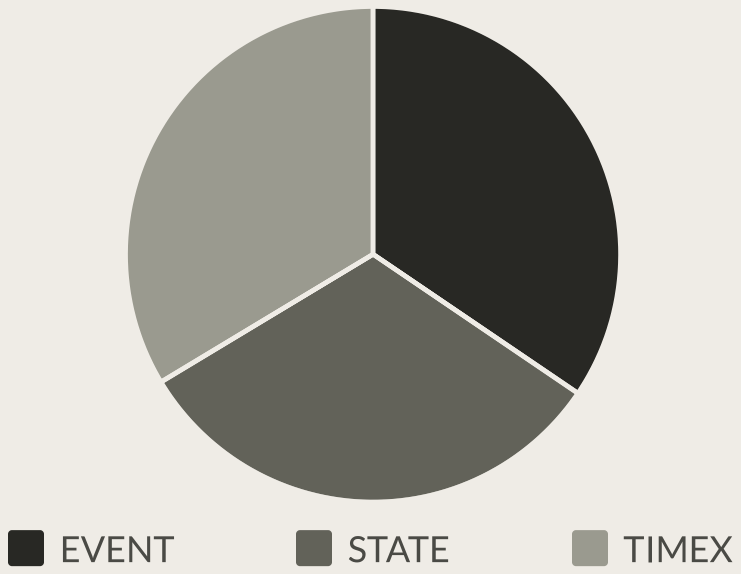
- **Labels:** EVENT, STATE, and TIMEX entities.

Precision: 0.80

Recall: 0.76

F1 Score: 0.78

Entity-Wise F1 Scores



Final Summary: Unified Pipeline for Enhanced NLP

This project successfully developed and validated a unified pipeline, delivering robust solutions for temporal information extraction and cross-lingual NER, particularly for Indic languages.

1

English Temporal Extraction

Achieved a competitive **0.86 F1 score**, matching state-of-the-art systems with an optimized LLM prompting approach.

2

Cross-Lingual NER Dataset Generation

Produced high-quality datasets for 7 Indic languages with a strong **0.82 Token-level F1** for projection reliability.

3

XLM-R Multilingual NER

Trained a robust model with an **overall 0.7891 F1 score** and **97.97% accuracy** across the Indic languages.

4

Hindi Temporal Extraction

Demonstrated effective temporal entity recognition in Hindi with a solid **0.78 F1 score**.

Our contributions significantly enhance low-resource NLP capabilities, providing valuable tools and insights for researchers and practitioners.

Thank You