Named Entity Recognition for Code-Mixed Text:Project Outline

Project Outline

LexiCoders

March 30, 2025

1 Problem Statement

Named Entity Recognition (NER) in code-mixed text presents unique challenges compared to monolingual text. Code-mixing, where speakers alternate between multiple languages within a conversation or even within a sentence, is extremely common in multilingual communities, particularly in Indian social media contexts. This project focuses on developing an effective NER system specifically for Hindi-English code-mixed text that can accurately identify and classify named entities despite the complexities introduced by language mixing.

1.1 Problem Niche

Standard NER systems struggle with code-mixed text due to:

- Language identification challenges (determining which parts are in which language)
- Non-standard spellings and transliterations
- Mixing of scripts (Roman, Devanagari)
- Culturally-specific entities that might not exist in monolingual datasets
- Lack of annotated code-mixed corpora for training

2 Project Scope

The project will encompass:

1. **Dataset Creation and Annotation**:
    - Collection of Hindi-English code-mixed social media text from platforms like Instagram, Twitter, Facebook and WhatsApp
    - Development of comprehensive annotation guidelines for named entities in code-mixed text
    - Creation of a gold-standard annotated corpus with more than 1000 sentences
2. **Entity Types**:
    - Person (PER)
    - Location (LOC)
    - Organization (ORG)
    - Date/Time (DATE)
    - Culturally-specific entities (e.g., festivals, local events) (CSE)
    - Products, brands (PROD)
3. **Model Development**:

- Feature-based Conditional Random Field (CRF) model

4. **Evaluation**:
    - Standard metrics: Precision, Recall, F1-score
    - Analysis across different entity types
    - Analysis based on code-mixing density and patterns

3 Data Exploration and Creation

3.1 Existing Datasets

Following are the datasets that we will be using in the project:

1. **Replication Data for Automatic language identification in code-switched Hindi-English social media text**
    - Link for the dataset: https://dataverse.harvard.edu/dataset.xhtml;jsessionid=d99a0d881cd919d13cd21af594bf?persistentId=doi:10.7910/DVN/QD94F9
2. **Code-Mixed Dataset for Hindi-English**: Contains code-mixed social media posts
    - Data will be scraped manually from social media posts and chats, Link to the dataset will be provided later.

3.2 Data Creation Process

1. **Collection**:
    - Scrape tweets and posts with Hindi-English hashtags and topics
    - Collect public social media posts and comments from Indian pages
2. **Filtering**:
    - Filter for sentences with substantial mixing (at least 20% of both languages)
3. **Annotation**:
    - Manually annotate the dataset extracted from social media
    - Dataset taken from Harvard is already annotated

3.3 Annotation Guidelines

Detailed guidelines will be provided:

- Entity boundaries in mixed-language contexts
- Handling transliteration variations
- Culturally-specific entities classification
- Script variation handling

4 Technical Approach

4.1 A. Preprocessing

1. **Text Normalization**:
   - Handling of non-standard spellings
   - Transliteration normalization using tools like indic-trans
2. **Language Identification**:
   - Word-level language tagging using models like MultiLID
   - Script identification (Roman vs. Devanagari)

4.2 B. Feature Engineering for CRF

1. **Lexical Features**:
   - Word identity, prefix/suffix n-grams
   - Word shape (capitalization, digits, etc.)
   - Language tag of word
2. **Contextual Features**:
   - Previous/next words and their language tags
   - Word n-grams
3. **Dictionary Features**:
   - Gazetteers for both Hindi and English entities
   - Transliteration dictionaries
4. **Syntactic Features**:
   - POS tags from code-mixed POS taggers
   - Chunking information if available

4.3 C. Model Development

1. **CRF Implementation**:
   - Using sklearn-crfsuite or CRF++
   - Feature selection using grid search and cross-validation
2. **Deep Learning Approach**:
   - BiLSTM-CRF with code-mixed word embeddings
   - Using Glot500 or MBERT for representations
3. **Transformer-based Models**:
   - Fine-tuning models like MuRIL (Multilingual Representations for Indian Languages)
   - Code-mixed BERT variants using HuggingFace Transformers

5 Literature Review

Several research papers provide valuable insights for this project:

1. **Named Entity Recognition on Code-Switched Data: Overview of the CALCS 2018 Shared Task by Aguilar et al. (2018)**:
    - Though focused on Spanish-English and Arabic dialects, their annotation guidelines can be adapted for Hindi-English entity labeling
    - Their error analysis reveals that person names are easier to detect than organizations across language pairs, which may inform our approach
    - Their finding that deep learning methods outperform traditional CRF models supports our plan to explore both approaches
    - Their code-switching evaluation metrics can be directly applied to measure our system's performance on different mixing patterns
2. **Part-of-speech tagging of code-mixed social media content: Pipeline, stacking and joint modelling by Vyas et al. (2014)**:
    - Directly addresses Hindi-English code-mixing in social media text, matching our target domain
    - Their approach to handling romanized Hindi words will be critical for our preprocessing stage
    - Their language identification techniques can be incorporated into our feature extraction process
    - Their joint modeling approach suggests we should integrate language identification with NER rather than treating them as separate steps
3. **Named Entity Recognition in Code-Switched Data Using Neural Architecture by Winata et al. (2018)**:
    - Their character-level approach addresses the exact challenge we face with non-standard spellings and transliterations in Hindi-English text
    - Their model architecture can be studied adapted for our Hindi-English NER system
    - Their work proves that strong results are possible without explicit language identification, which could simplify our pipeline
    - Their error analysis on entity boundaries in mixed-language contexts will inform our annotation guidelines

6 Implementation Plan

6.1 Phase 1: Data Collection and Annotation

- Develop annotation guidelines
- Collect raw code-mixed data
- Begin annotation process
- Achieve target corpus size

6.2 Phase 2: Model Development

- Implement preprocessing pipeline
- Develop and train CRF model
- Optimize model

6.3 Phase 3: Evaluation and Analysis

- Comprehensive evaluation
- Error analysis and model refinement
- Performance analysis across entity types and code-mixing patterns

7 Evaluation Metrics

1. **Standard NER Metrics**:
   - Precision, Recall, and F1-score for each entity type
   - Overall micro and macro F1 scores
2. **Code-Mixing Specific Analysis**:
   - Performance as a function of code-mixing index
   - Analysis based on script variation patterns
   - Performance on mixed-language entities vs. single-language entities

8 Required Resources and Tools

8.1 Libraries and Frameworks:

- spaCy for NLP pipeline components
- sklearn-crfsuite for CRF implementation
- PyTorch for deep learning models
- HuggingFace Transformers for transformer models
- Indic NLP Library for Indian language processing

8.2 Annotation Tools:

- Doccano for collaborative annotation
- BRAT as an alternative annotation tool

9 Potential Challenges and Solutions

1. **Data Sparsity**:
    - Solution: Data augmentation techniques
2. **Annotation Consistency**:
    - Solution: Detailed guidelines and regular checks
3. **Handling Transliteration Variations**:
    - Solution: Normalization rules, character-level models, phonetic matching
4. **Model Performance on Low-Resource Languages**:
    - Solution: Transfer learning from high-resource languages, multilingual representations

10 Conclusion

This project will create a valuable resource and system for NER in Hindi-English code-mixed text, addressing a significant gap in current NLP capabilities for Indian languages. The developed methods can potentially be extended to other Indian language pairs and can inform approaches for other code-mixed NLP tasks. The annotated corpus will also serve as a benchmark dataset for future research in this area.
Language Technologies Research Lab at IIIT-Hyderabad, LCS2 at IIT-Madras and Microsoft Research India have all conducted significant work in code-mixed NLP for Indian languages and their methodologies and resources will be valuable references for this project.