Named Entity Recognition for Code-Mixed Text

Project Presentation

LexiCoders

March 30, 2025

Problem Statement

- Named Entity Recognition (NER) in code-mixed text presents unique challenges compared to monolingual text
- Code-mixing: speakers alternating between multiple languages within a conversation or sentence
- Common in multilingual communities, particularly in Indian social media contexts
- Project focus: Developing an effective NER system for Hindi-English code-mixed text
- Goal: Accurately identify and classify named entities despite language mixing complexities

Problem Niche: Why Standard NER Systems Struggle

- 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

Project Scope (1/2)

Dataset Creation and Annotation:

- Collection of Hindi-English code-mixed social media text
  - Platforms: Instagram, Twitter, Facebook, WhatsApp
- Development of comprehensive annotation guidelines
- Creation of gold-standard annotated corpus (1000+ sentences)

Entity Types:

- Person (PER)
- Location (LOC)
- Organization (ORG)
- Date/Time (DATE)
- Culturally-specific entities (CSE)
- Products, brands (PROD)

Project Scope (2/2)

Model Development:

- Feature-based Conditional Random Field (CRF) model

Evaluation:

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

Data Exploration and Creation: Existing Datasets

1. **Replication Data for Automatic language identification in code-switched Hindi-English social media text**
   - Source: Harvard Dataverse
   - Link: `https://dataverse.harvard.edu/dataset.xhtml?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 be provided later

Data Creation Process

Collection:

- Scrape tweets and posts with Hindi-English hashtags and topics
- Collect public social media posts and comments from Indian pages

Filtering:

- Filter for sentences with substantial mixing (at least 20% of both languages)

Annotation:

- Manually annotate the dataset extracted from social media
- Dataset taken from Harvard is already annotated

Annotation Guidelines

Detailed guidelines will address:

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

Technical Approach: Preprocessing

Text Normalization:

- Handling of non-standard spellings
- Transliteration normalization using tools like indic-trans

Language Identification:

- Word-level language tagging using models like MultiLID
- Script identification (Roman vs. Devanagari)

Technical Approach: Feature Engineering for CRF

Lexical Features:

- Word identity, prefix/suffix n-grams
- Word shape (capitalization, digits, etc.)
- Language tag of word

Contextual Features:

- Previous/next words and their language tags
- Word n-grams

Dictionary Features:

- Gazetteers for both Hindi and English entities
- Transliteration dictionaries

Syntactic Features:

- POS tags from code-mixed POS taggers
- Chunking information if available

Technical Approach: Model Development

CRF Implementation:

- Using sklearn-crfsuite or CRF++
- Feature selection using grid search and cross-validation

Deep Learning Approach:

- BiLSTM-CRF with code-mixed word embeddings
- Using Glot500 or MBERT for representations

Transformer-based Models:

- Fine-tuning models like MuRIL (Multilingual Representations for Indian Languages)
- Code-mixed BERT variants using HuggingFace Transformers

Literature Review (1/3)

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, annotation guidelines can be adapted for Hindi-English
- Error analysis reveals person names are easier to detect than organizations across language pairs
- Deep learning methods outperform traditional CRF models
- Code-switching evaluation metrics can be applied to measure system performance on different mixing patterns

Literature Review (2/3)

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
- Approach to handling romanized Hindi words critical for preprocessing
- Language identification techniques can be incorporated into feature extraction
- Joint modeling approach suggests integrating language identification with NER rather than treating them as separate steps

Literature Review (3/3)

Named Entity Recognition in Code-Switched Data Using Neural Architecture by Winata et al. (2018)

- Character-level approach addresses challenges with non-standard spellings and transliterations
- Model architecture can be studied and adapted for Hindi-English NER system
- Strong results possible without explicit language identification
- Error analysis on entity boundaries in mixed-language contexts will inform annotation guidelines

Implementation Plan: Phase 1

Data Collection and Annotation

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

Implementation Plan: Phases 2 & 3

Phase 2: Model Development

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

Phase 3: Evaluation and Analysis

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

Evaluation Metrics

Standard NER Metrics:

- Precision, Recall, and F1-score for each entity type
- Overall micro and macro F1 scores

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

Required Resources and Tools

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

Annotation Tools:

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

Potential Challenges and Solutions

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

Conclusion

- Creation of valuable resource and system for NER in Hindi-English code-mixed text
- Addresses significant gap in current NLP capabilities for Indian languages
- Methods potentially extendable to other Indian language pairs
- Can inform approaches for other code-mixed NLP tasks
- Annotated corpus will serve as benchmark dataset for future research

Key Research Institutions:

- Language Technologies Research Lab at IIIT-Hyderabad
- LCS2 at IIT-Madras
- Microsoft Research India

Thank You

Questions?

LexiCoders
March 30, 2025