Detecting Entailment in Code-Mixed Hindi-English Conversations

Sharanya Chakravarthy* Anjana Umapathy* Alan W Black

Abstract

Code-mixing is the intertwined usage of multiple languages commonly seen in informal conversations among polyglots. With the rising importance of dialog agents, understanding code-mixing is imperative, but the scarcity of code-mixed Natural Language Understanding datasets has precluded research in this area. We tackle the task of detecting conversational entailment in code-mixed Hindi-English text. We investigate language modeling, data augmentation, and architectural approaches to address the code-mixed, conversational, low-resource aspects of the dataset. We obtain a test set accuracy of 62.41%, an increase of 8.09% over the current state-of-the-art (mBERT).

Task - NLI

- **Input (In Hinglish):** (Premise, Hypothesis)
- **Output Label:** Entailment / Contradiction
- **Premise Example:** RAHUL: Tumhara scooter aur ek joota security guard ko lobby mein mila . RIANA: Thank god!!
- **Premise (Translation)**: RAHUL: The security guard found your scooter and shoe in the lobby. RIANA: Thank god!!
- **Entailment Hypothesis (Translation):** Rahul told RIANA her shoe was found
- **Contradiction Hypothesis (Translation):** Riana told Rahul that security found his shoe
- **Dataset:** 2,240 examples

Methodology

Code-mixing

mod-mBERT:

- MLM fine-tuning of mBERT on code-mixed Bollywood movie scripts, other Hinglish datasets

Transliteration & Translation:

- Token level language ID
- Transliterate Hindi words to Devanagari
- Translate English phrases to Hindi

Low-resource

Data Augmentation:

- Stanford NLI (SNLI): English NLI
- Cross-lingual NLI (XNLI): Devanagari Hindi
- Multi-Premise Entailment (MPE): Longer premises

Conversational Premises

Data Augmentation:

- Additional contradiction examples by changing speaker in entailment hypotheses
- Additional examples by changing all speaker names in premise & hypothesis

Utterance Representations:

- mod-mBERT representations of each utterance passed through an LSTM network

Results and Analysis

Results on 8-fold cross-validation. Hi: Hindi

Model Name	Mean Acc.	Std. Dev.
FINE-TUNING PRE-TRAINED MODELS		
BERT	61.11%	3.38
mBERT	60.94%	3.16
mod-mBERT	61.28%	2.08
TRANSLITERATION & TRANSLATION (MBERT)		
Transliteration of CS-NLI	62.17%	2.00
Hi translation of CS-NLI	60.04%	3.71
CS-NLI & its Hi translation	63.30%	3.05
AUGMENTATION OF CS-NLI		
mod-mBERT on 3k XNLI	**63.69%**	**1.58**
mod-mBERT on 4k SNLI & 4k XNLI	63.35%	2.53
mod-mBERT on 4k MPE	62.19%	3.11
XLM-R on 4k SNLI & 4k XNLI	63.52%	1.85
CONVERSATIONAL APPROACHES (MOD-MBERT)		
CS-NLI & Speaker Name Augmentation	62.85%	2.00
CS-NLI & Speaker Name, Contradiction Augmentation	61.39%	1.87
biLSTM	54.83%	1.72

Analysis

- Fine-tuning mBERT on Hinglish text helped the model understand contextual code-mixing
- Translation errors cascade, but augmenting the code-mixed data with translated versions helps
- Additional NLI examples help the model learn the task, despite language and domain mismatch

Future Work

- Data selection strategies, Speaker / conversation aware architectures, Pre-trained models for code-mixed data

Carnegie Mellon University
Language Technologies Institute