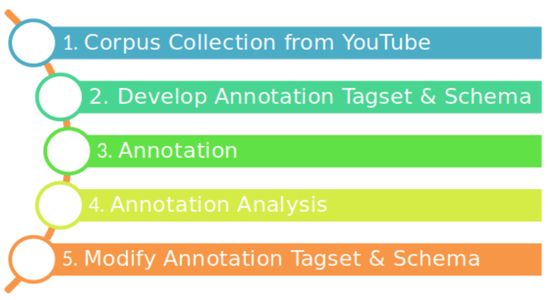
MHE: Code-Mixed Corpora for Similar Language Identification

Priya Rani John P McCrae Theodorus Fransen
Data Science Institute, National University of Ireland Galway

Objective

- We developed MHE - an annotated Magahi-Hindi-English code-mixed corpus for similar language identification.
- We present baseline scores for similar language identification on both sentence and word level identification tasks.

Corpus Creation Process



Corpus Statistics

Number	MHE
Sentences	16,784
Words	146,256
Unique words	15,348

Data Annotation

- Word level annotation examples:

Tag	Word	Translation
MAG	'हमहु'	I also
HIN	'आपसे'	From you
ENG	'Like'	---
H&M	'saal'	Year
OTH	'@', '#'	---
NAME	'Sushant', 'नरेश'	Sushant, Naresh
NUM	'1', '5'	---
ABV	'CM'	Chief Minister
UNK	'Jena'	This is not

- Sentence level annotation examples

Tag	Sentence	Translation
MAG	**'Bdi achha laglo Sr asne video bnayte rhiya'**	It was an awesome video, sir please make these types of videos more
HIN	**"Kya baat Bhaiya?'**	What's the matter brother?
ENG	**"Best commentry'**	—
UNK	**'Maithili Jena chhai'**	This is not Maithili

Annotation Analysis

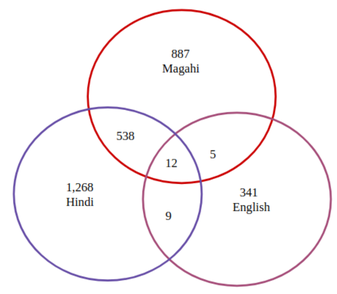
- Inter annotator agreement was calculated using Krippendorff's alpha.
- **Simple Data Annotation:** Annotation phase 1 & 2. One could infer the context of the tokens.
- **Reshuffled Data Annotation:** Annotation Phase 3 & 4. No one could infer the context of the tokens.

Phase	Word-level	Sentence-level
Phase-1	0.87	0.89
Phase-2	0.91	0.93
Phase-3	0.83	—
Phase-4	0.89	—

Code Mixing Analysis

The analysis helps us understand the relatedness between the languages and the dataset's characteristics. The analysis was carried out by studying:

- **Lexical Overlap:** To understand the similarities among the languages and the complexity of the code-mixed data.
  - Same words across the languages
  - Phonetic similarity



- **Code Mixing Index:** This is used to understand the dataset's complexity.

Language Pair	CMI
English-Bengali (Gambäck and Das, 2014)	24.48
Dutch-Turkish (Nguyen and Doğruöz, 2013)	22.65
Modern Arabic-Egyptian Arabic (Molina et al., 2016)	3.89
Spanish-English (Mave et al., 2018)	22.11
Hindi-English (Mave et al., 2018)	22.22
Nepali-English (Solorio et al., 2014)	20.32
Magahi-Hindi-English	51.54

Language Identification Baselines

Models	F1 Score (Word)	F1 Score (Sentence)
SVM (n-grams)	0.54	0.49
SVM (n-grams+TFIDF)	0.43	0.45
LSTM (Mave et al., 2018)	0.74	0.66
CNN (Zhang et al., 2015)	0.77	0.72
UDLDI (Goswami et al., 2020)	0.89	0.84
CLD2[8]	—	0.68

Conclusion

- A new Magahi-Hindi-English code-mixed annotated dataset for similar language identification.
- New detailed annotation guideline for closely related language identification.
- The CMI analysis shows the complexity of the dataset, which is substantially more complex compared to other datasets available.
- The baseline result for the dataset shows that attention-based models work best with such a dataset.

Acknowledgment

Financial support for this work is gratefully acknowledged from Science Foundation Ireland Grant No. 18/CRT/6223.