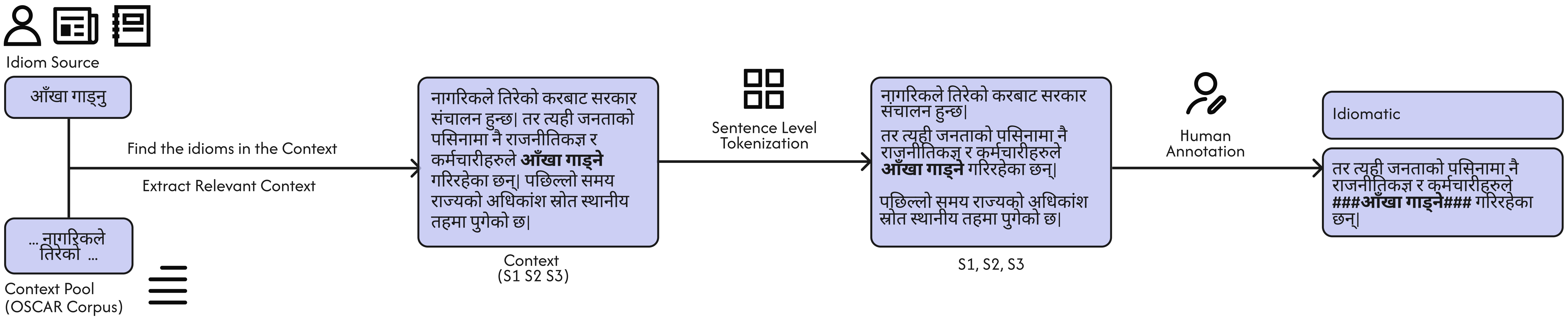
neDIOM: Dataset and Analysis of Nepali Idioms *CHiPSAL, COLING 2025, Abu Dhabi*

Rhitabrat Pokharel and Ameeta Agrawal
Portland State University

Data Collection Process



Idiom Source

आँखा गाड्नु

Find the idioms in the Context

Extract Relevant Context

... नागरिकले तिरेको ...

Context Pool (OSCAR Corpus)

नागरिकले तिरेको करबाट सरकार संचालन हुन्छ| तर त्यही जनताको पसिनामा नै राजनीतिकज्ञ र कर्मचारीहरुले **आँखा गाड्ने** गरिरहेका छन्| पछिल्लो समय राज्यको अधिकांश स्रोत स्थानीय तहमा पुगेको छ|

Context (S1 S2 S3)

Sentence Level Tokenization

नागरिकले तिरेको करबाट सरकार संचालन हुन्छ|

तर त्यही जनताको पसिनामा नै राजनीतिकज्ञ र कर्मचारीहरुले **आँखा गाड्ने** गरिरहेका छन्|

पछिल्लो समय राज्यको अधिकांश स्रोत स्थानीय तहमा पुगेको छ|

S1, S2, S3

Human Annotation

Idiomatic

तर त्यही जनताको पसिनामा नै राजनीतिकज्ञ र कर्मचारीहरुले **###आँखा गाड्ने###** गरिरहेका छन्|

Data Collection Challenges

1. Lack of comprehensive idiom repositories.
2. Identification of contexts that include idioms.
3. Idiom detection issue due to variations in their usage caused by inflection.

उसको सम्झाउने प्रयास गर्नु त **बालुवामा पानी हाले** जस्तै हो

बालुवामा पानी हाल्नु

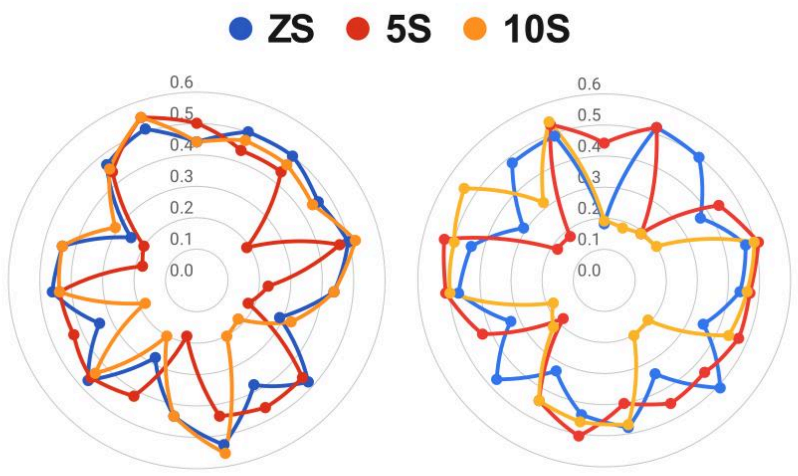
Inflection makes finding idioms challenging.

Dataset Statistics

Number of Unique Idioms: 191
Number of Samples: 1,216
Idiomatic Samples: 408; Literal Samples: 118
Discarded Samples: 690
Annotation: 3 native speakers were asked to mark I/L and the position of the idiom

Experimental Setting

Models: XLM-R, BLOOM-560m, BLOOM-1b1 , BLOOM-3b, BLOOM-7b, Llama2-7b, GPT-3.5
Task: Idiomaticity Detection
Training Setting: 0-shot, 5-shot, 10-shot; with and without idioms



Plot showing the average performance of the models under zero-shot (ZS), 5-shot (5S), and 10-shot (10S) settings with idiom (left) and without idiom (right).

Scan me to download the dataset.

Contact: pokharel@pdx.edu

Major Findings

1. It is challenging for the model to effectively distinguish between literal and idiomatic labels.
2. Despite LLMs being extensively trained on data from high-resource languages within the same language family, their performance in low-resource language contexts fell short of expectations, even after fine-tuning.
3. Larger models were not always more effective.
4. The effect of the presence of surrounding context depends on the model.