Motivation

- Indio-Aryan languages are widely spoken by roughly billion speakers.
- Many Indo-Aryan languages lack large amounts of monolingual corpora or labelled data for various NLP tasks.
- Indo-Aryan languages exhibit high language relatedness which should alleviate labelled data sparsity problems.
- For example, the verb alleviates labelled data sparsity problem.

Aryan languages exhibit high language relatedness which should alleviate labelled data sparsity problem. Aryan languages are widely spoken by roughly billion speakers.

Task from IndicGLUE

Textual Entailment task on copa-translated dataset (Hi, Gu, Mr, Pr, Pa)

Named Entity Recognition task on wikian-ner dataset (Hi, Bn, Gu, Mr, Or, Pa)

Title Prediction task on wiki-section-title dataset (Hi, Bn, Gu, Mr, Or, Pa)

Part-of-Speech Tagging task on Universal Dependency datasets (Hi, Mr, Ut)

Analysis

Research Question 1

Does multilingual fine-tuning with a set of related languages yield improvements over monolingual fine-tuning (FT) on downstream tasks?

Within the set of related languages, likely, there exists a subset of languages that yields the best performance.

Research Question 2

Starting from monolingual FT, as each related language is gradually added to the training set, FT to ultimately a multilingual FT with all related languages.

Tuning on multiple Indo-Aryan languages significantly enhanced when coupled with translated on the combined labelled data of Aryan languages.

Research Question 3

Does having a common script benefit better language transfer for languages with high lexical similarity but different scripts?

Effectiveness of multilingual fine-tuning is significantly enhanced when coupled with common-script representation via transliteration.