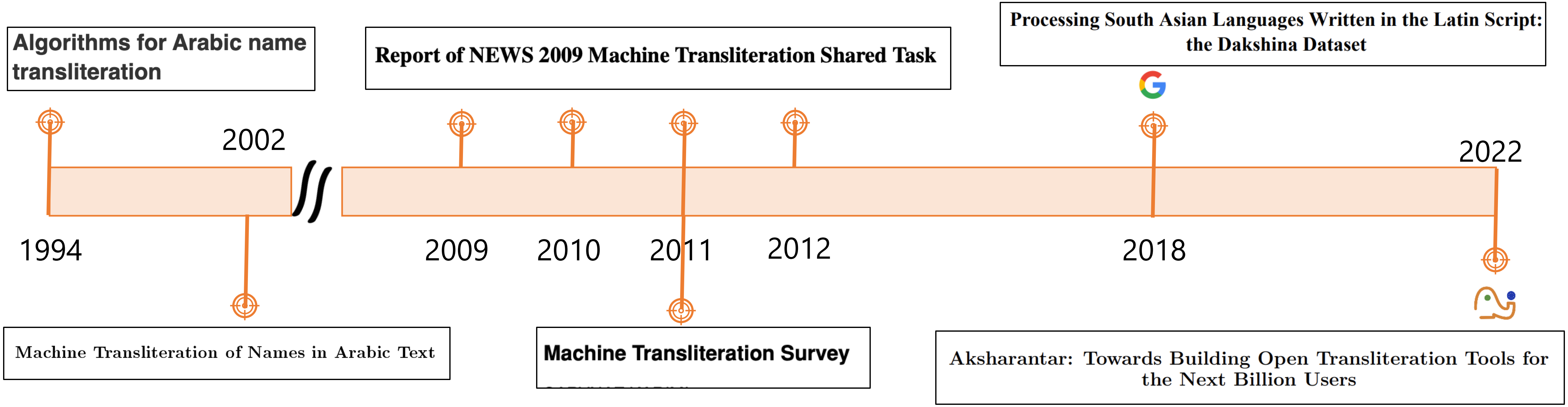
Machine Transliteration

https://ai4bharat.iitm.ac.in/transliteration

Workshop on 28th July 2022, IIT Madras

Machine Transliteration: A brief history



1994

Algorithms for Arabic name transliteration

2002

Machine Transliteration of Names in Arabic Text

2009

Report of NEWS 2009 Machine Transliteration Shared Task

2010

2011

Machine Transliteration Survey

2012

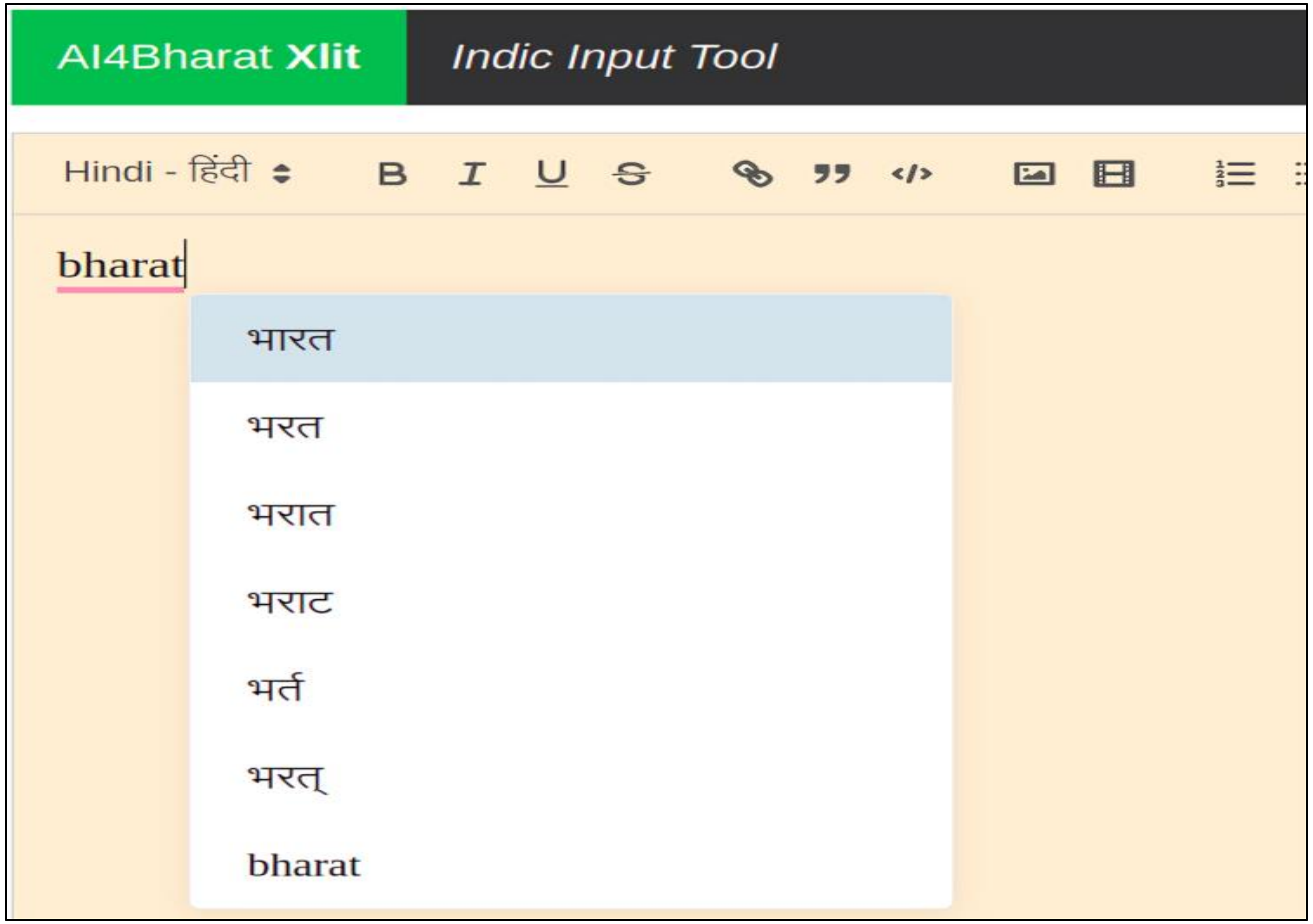
2018

Processing South Asian Languages Written in the Latin Script: the Dakshina Dataset

2022

Aksharantar: Towards Building Open Transliteration Tools for the Next Billion Users

The Problem



bharat

- भारत
- भरत
- भरात
- भराट
- भर्त
- भरत्
- bharat

Enable romanised typing in Indian languages

Our contributions

DATA

En हि

A large number of parallel sentences words En and Indic languages mined from monolingual and parallel corpora

MODELS

Large scale joint models with innovations specific to Indic languages

EVALUATION

Robust evaluation with diverse benchmarks for 20 languages

Mining data at scale

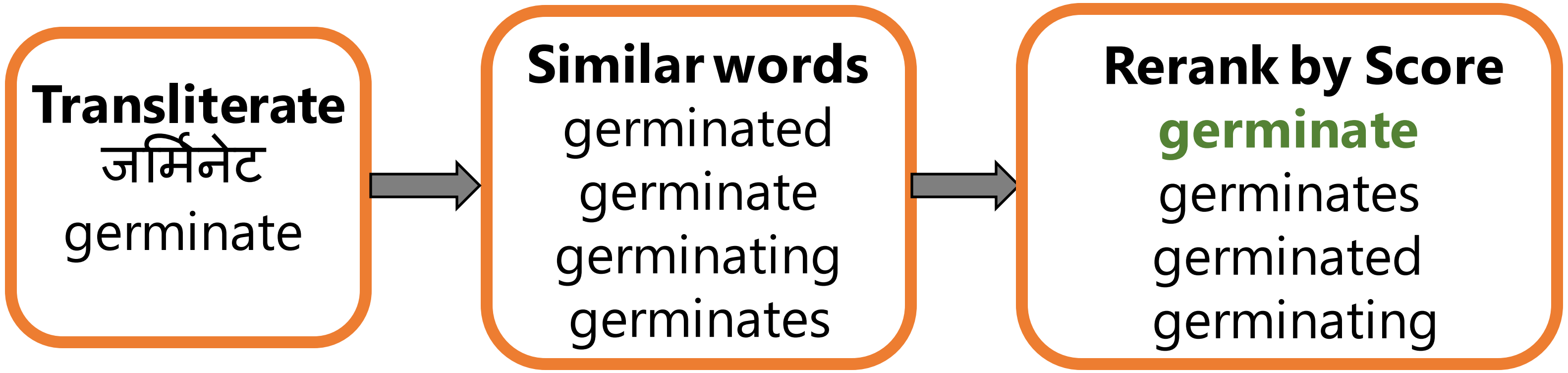
Samanantar

India will wear the orange **jersey** in **match** against **England** on **June** 30

⟺

टीम **इंडिया** 30 **जून** को विश्व कप **मैच** में **इंग्लैंड** के खिलाफ नारंगी **जर्सी** में खेलेगी

IndicCorp



Transliterate
जर्मिनेट
germinate

→

Similar words
germinated
germinate
germinating
germinates

→

Rerank by Score
germinate
germinates
germinated
germinating

Constitutes 90% of total mined data

Large scale manual data collection

Collected 103k pairs in 19 languages

2-step maker-checker process

Different types of Indic words Romanized

68 language experts across India

Built-in Automatic Transliteration validator

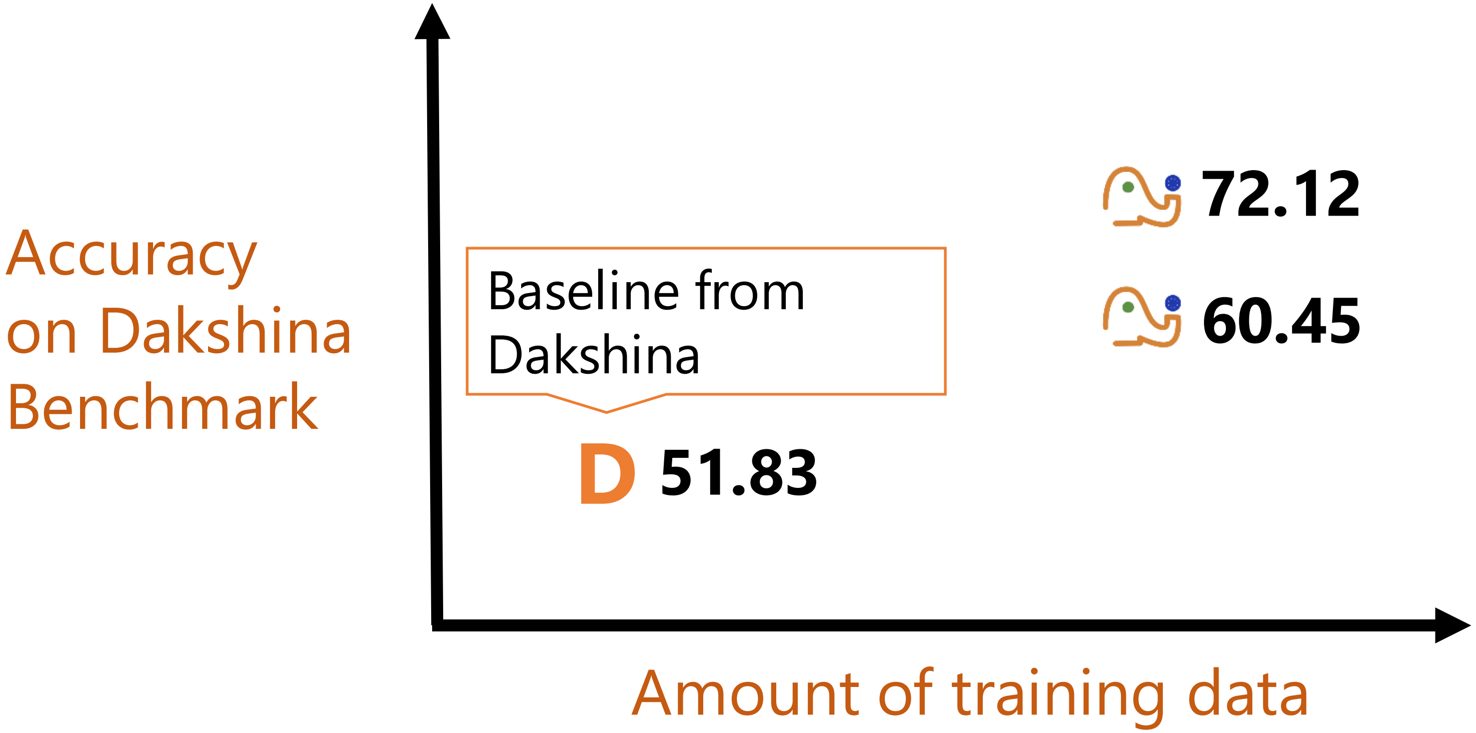
Model

A transformer based multilingual model

Reranking using frequency in IndicCorp

Covers 21 Indian languages

Results



Accuracy on Dakshina Benchmark

Baseline from Dakshina

D 51.83

72.12

60.45

Amount of training data