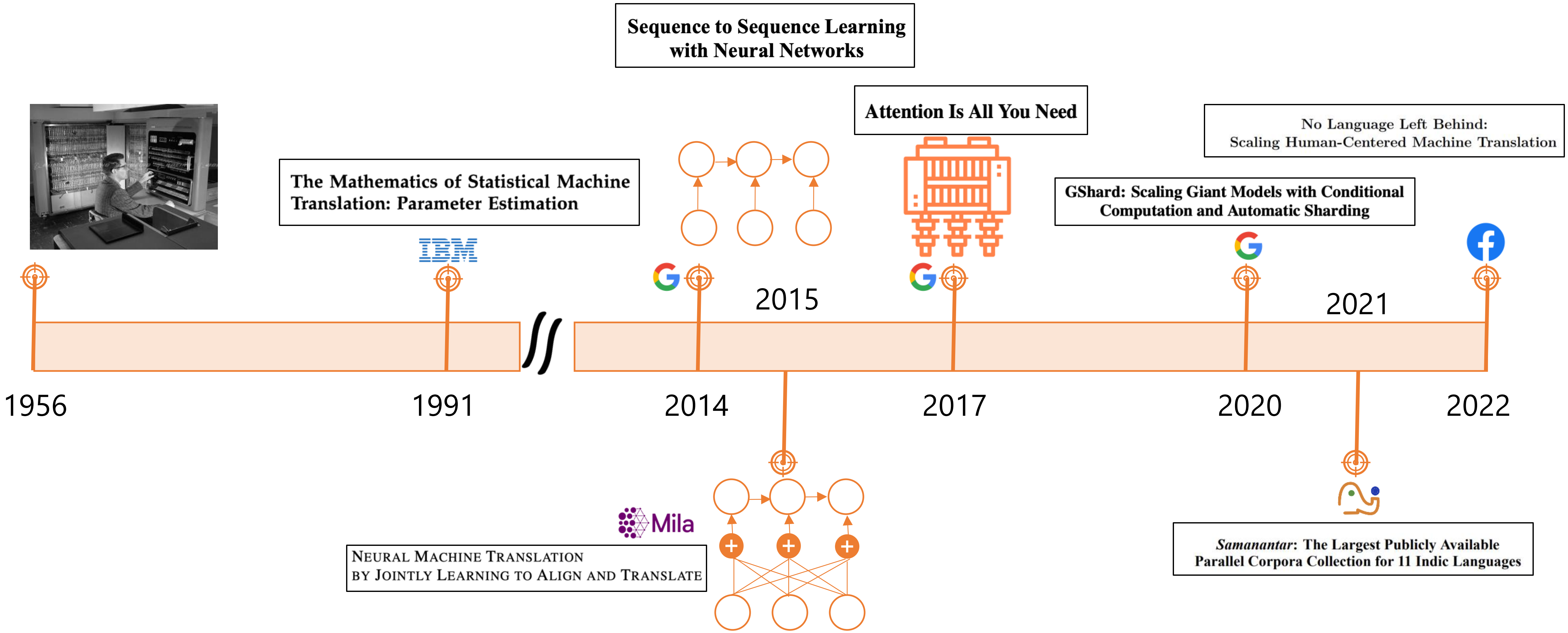
Machine Translation and Transliteration

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

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

Workshop on 28th July 2022, IIT Madras

Machine Translation: A brief history



1956

1991

The Mathematics of Statistical Machine Translation: Parameter Estimation

2014

Sequence to Sequence Learning with Neural Networks

2015

NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND TRANSLATE

2017

Attention Is All You Need

2020

GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding

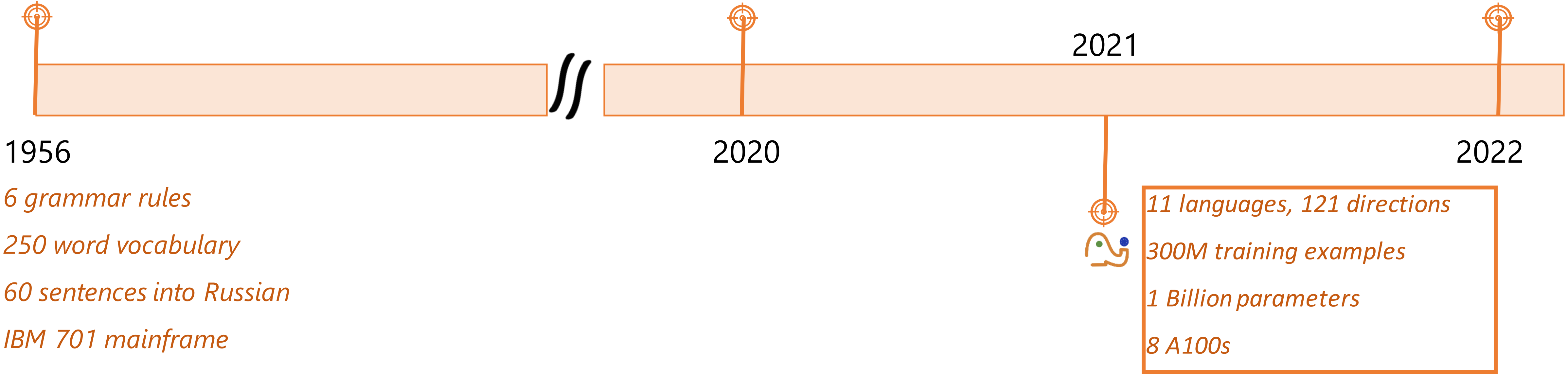
2021

***Samanantar*: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages**

2022

No Language Left Behind: Scaling Human-Centered Machine Translation

Machine Translation: A brief history



1956

6 grammar rules

250 word vocabulary

60 sentences into Russian

IBM 701 mainframe

2020

100 languages

13B training examples

1 Trillion parameters

*2048 TPUs**

2021

11 languages, 121 directions

300M training examples

1 Billion parameters

8 A100s

2022

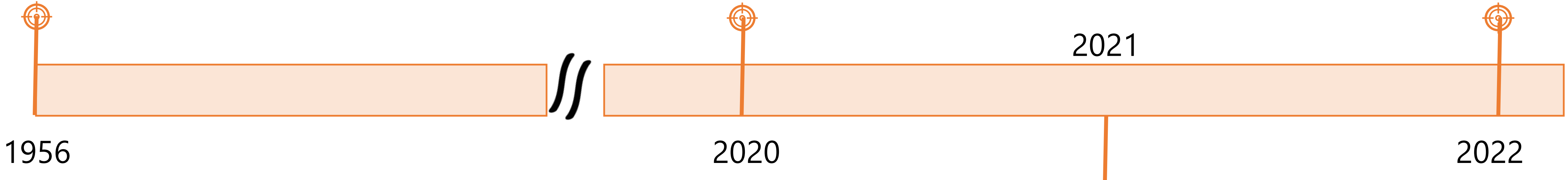
200 languages

XX training examples

YY parameters

ZZ A100s

Machine Translation: A brief history



6 grammar rules

250 word vocabulary

60 sentences into Russian

IBM 701 mainframe

100 languages

13B training examples

1 Trillion parameters

*2048 TPUs**

11 languages, 121 directions

300M training examples

1 Billion parameters

8 A100s

200 languages

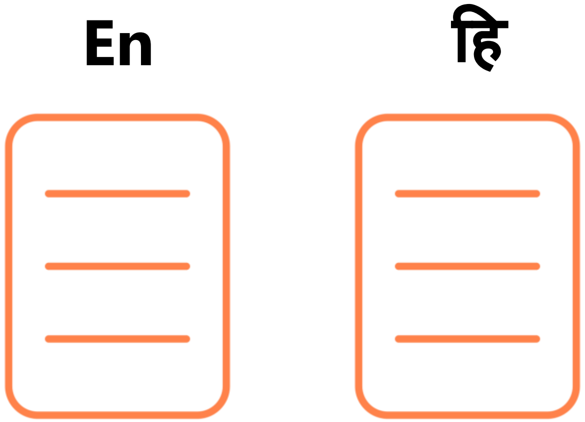
XX training examples

YY parameters

ZZ A100s

What is the recipe or modern NMT?

DATA



A large number of parallel sentences between En and Indic languages

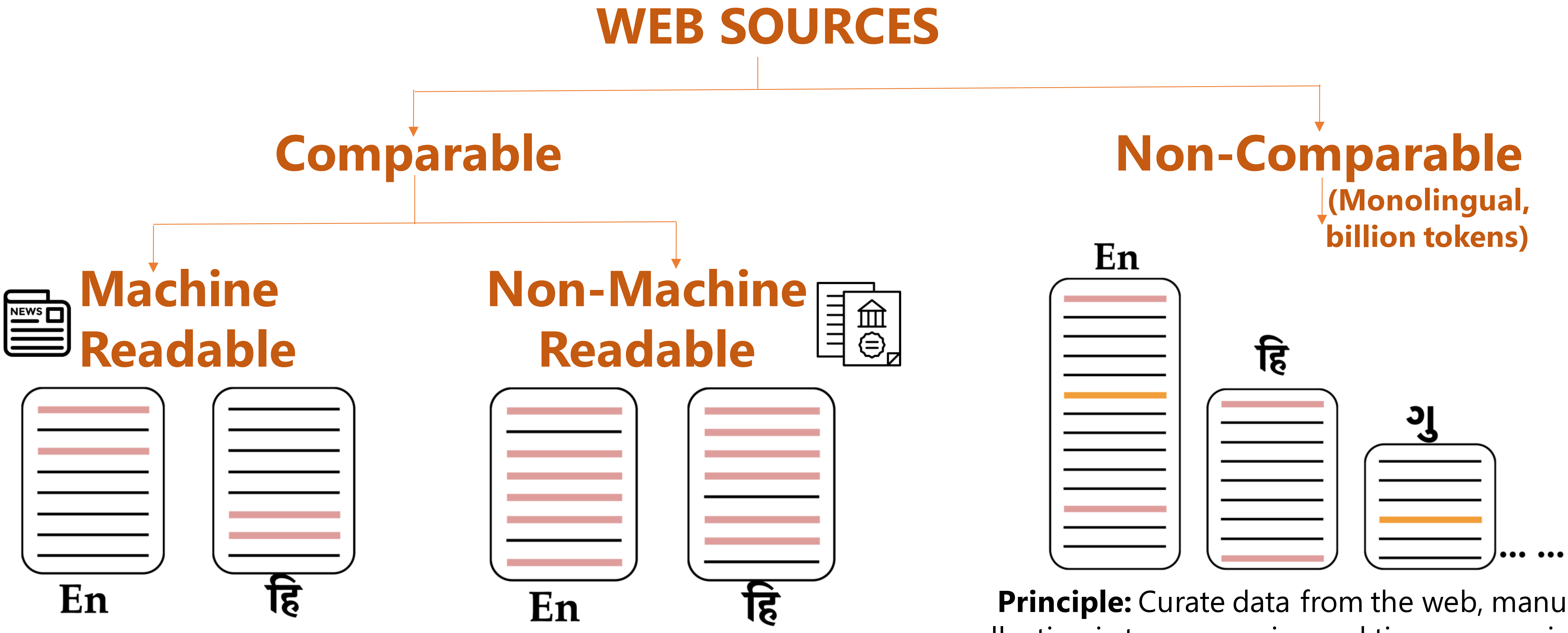
MODELS

Large scale models with innovations specific to Indic languages

EVALUATION

Robust evaluation with diverse benchmarks and reliable evaluation metrics

How did we solve the data problem?



WEB SOURCES

Comparable

Machine Readable

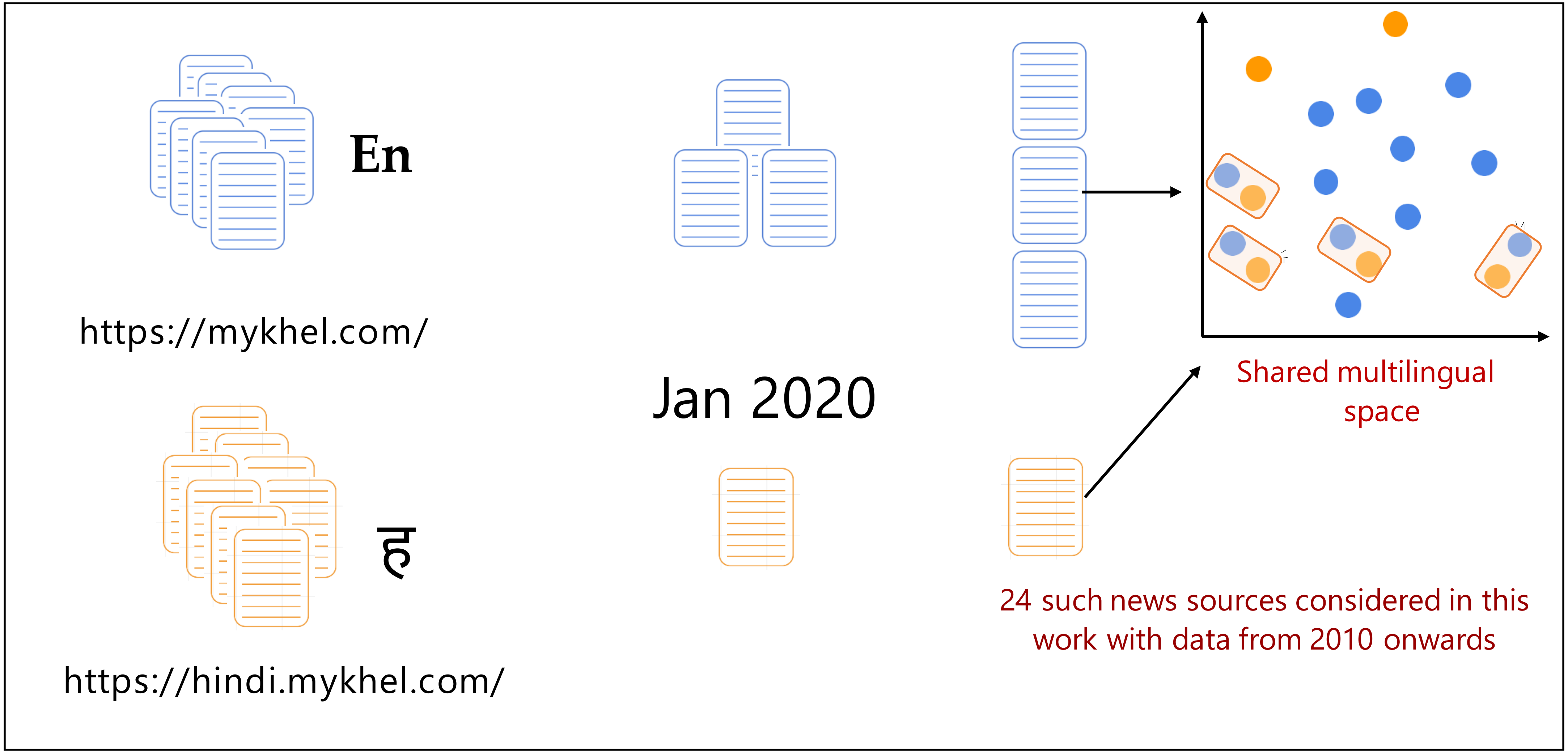
Non-Machine Readable

Non-Comparable

(Monolingual, billion tokens)

Principle: Curate data from the web, manual collection is too expensive and time consuming

How did we solve the data problem?



https://mykhel.com/

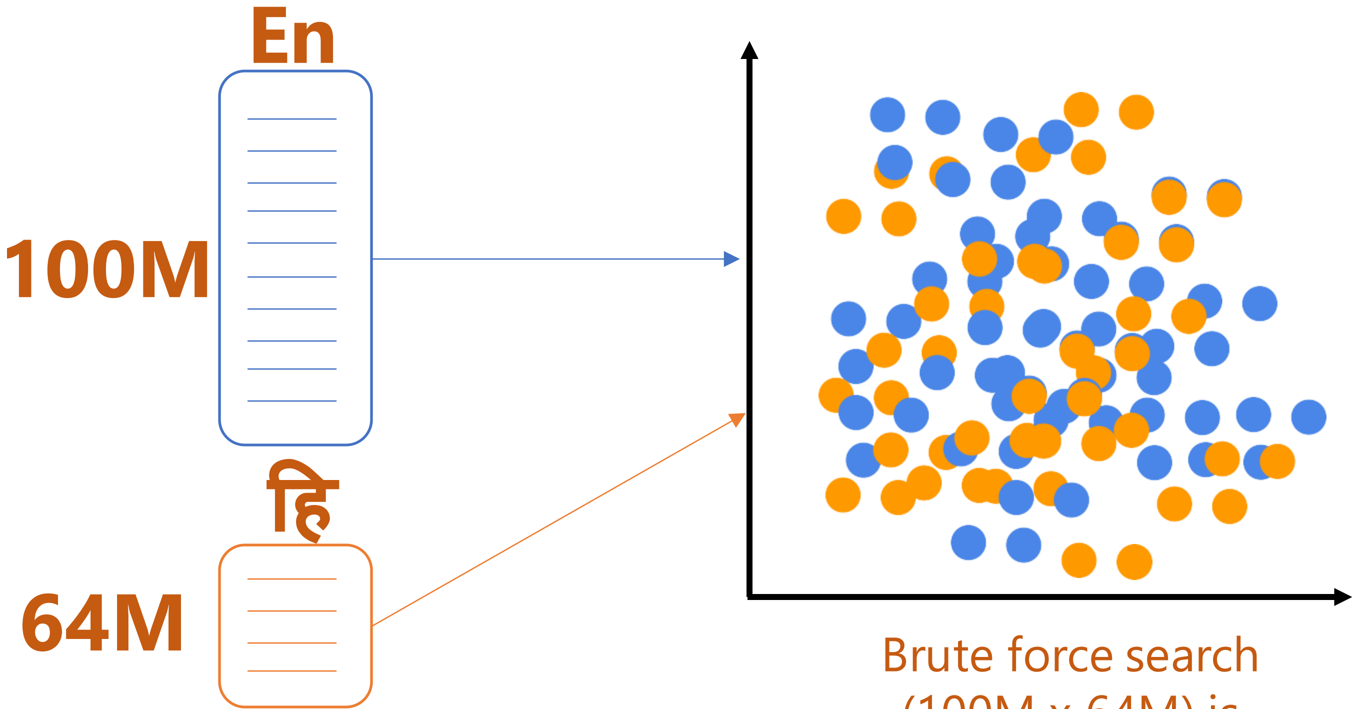
https://hindi.mykhel.com/

Jan 2020

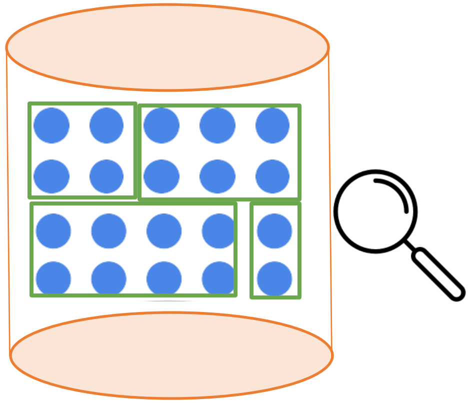
Shared multilingual space

24 such news sources considered in this work with data from 2010 onwards

How did we solve the data problem?

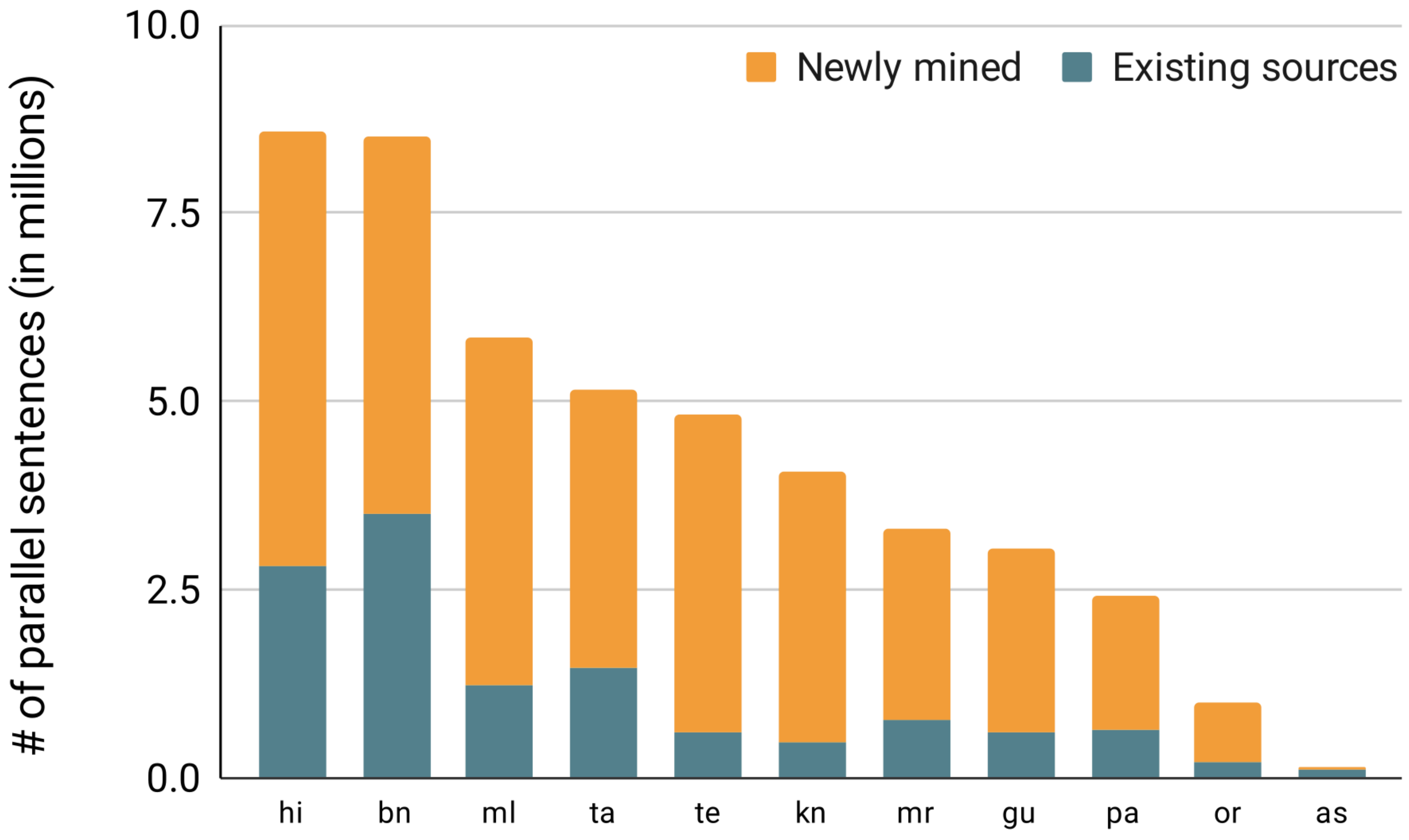


Brute force search (100M x 64M) is infeasible



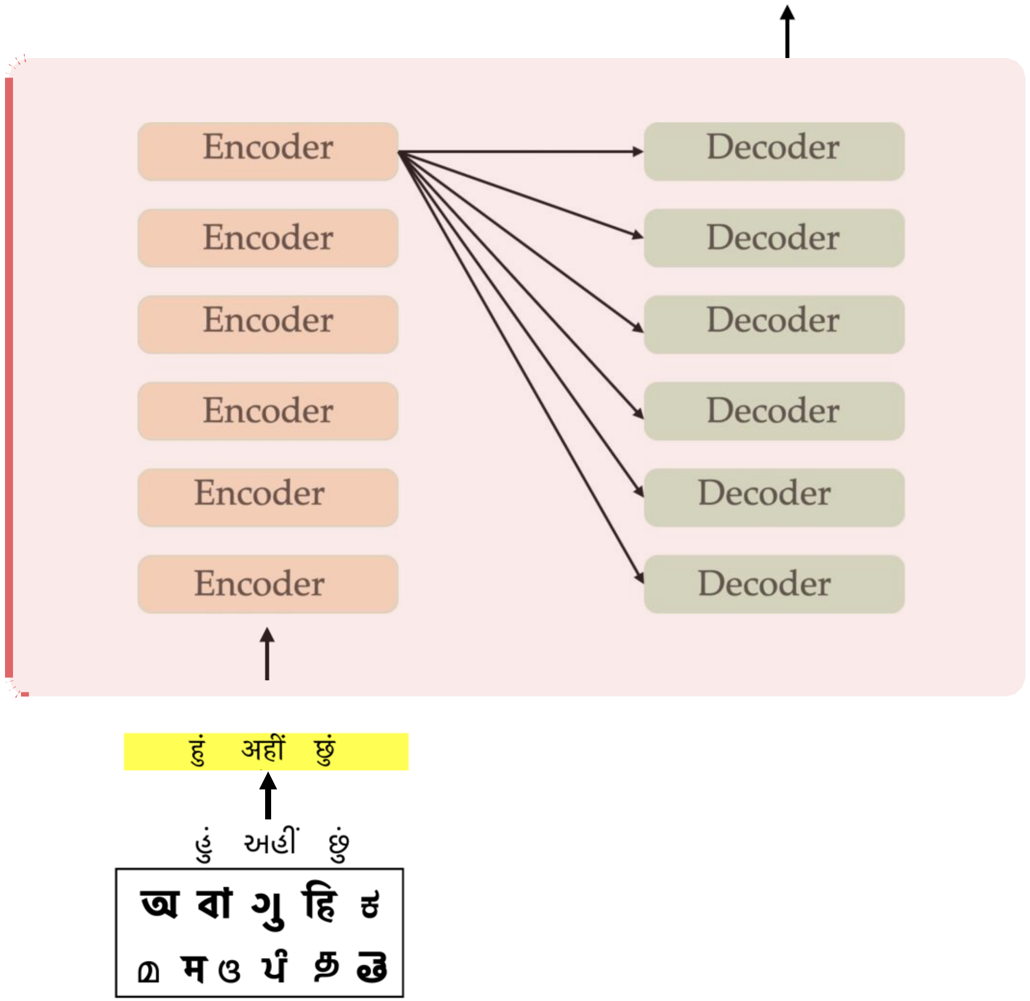
FAISS Index* for efficient indexing, clustering, semantic matching, and retrieval of dense vectors (1000 sent./sec)

How much data did we collect?



33M parallel sentences mined from the web (3X improvement)

What is the model that we use?



Highlights

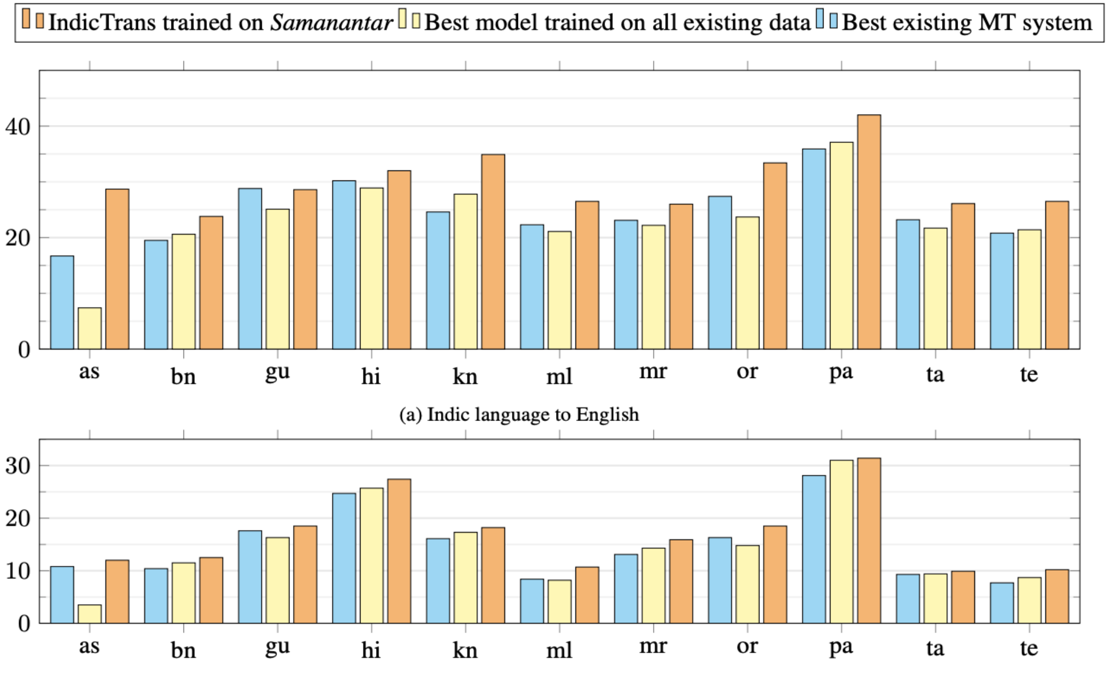
Joint multilingual model for 11 Indic languages

Single script (enables transfer, reduces vocabulary)

3 models: En-X, X-En, X-X

6 encoder layer, 6 decoder layers, 16 heads/layer

What is the model that we use?



(a) Indic language to English

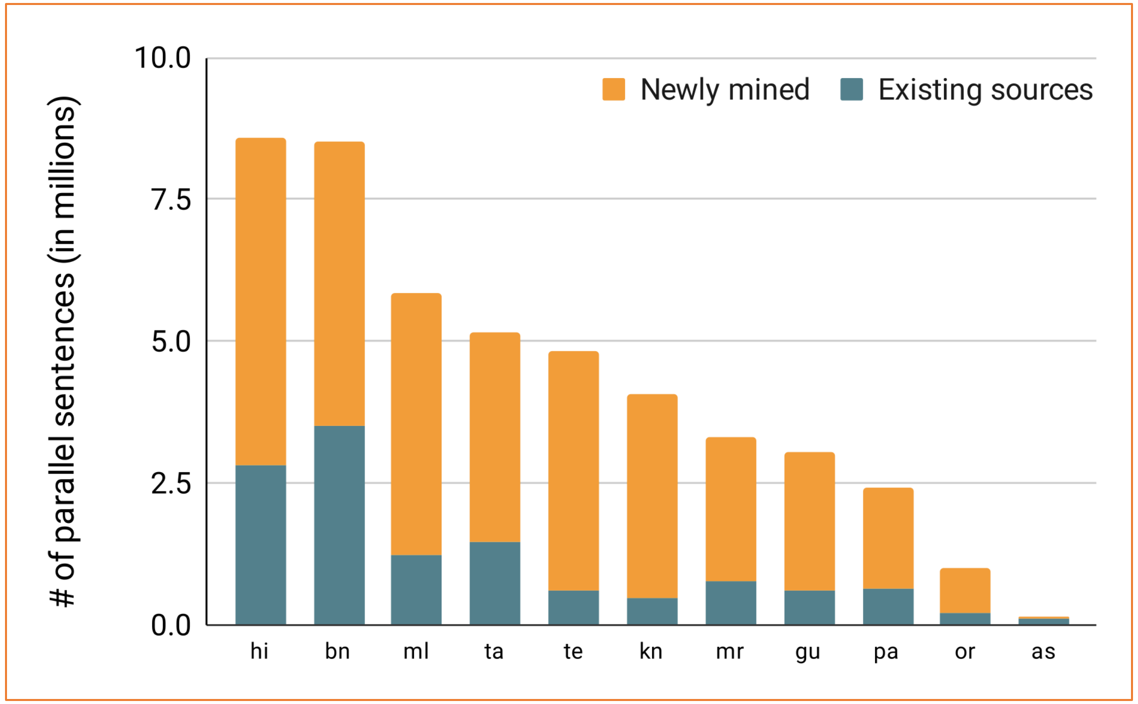
(b) English to Indic language

Highlights

State of the art performance*

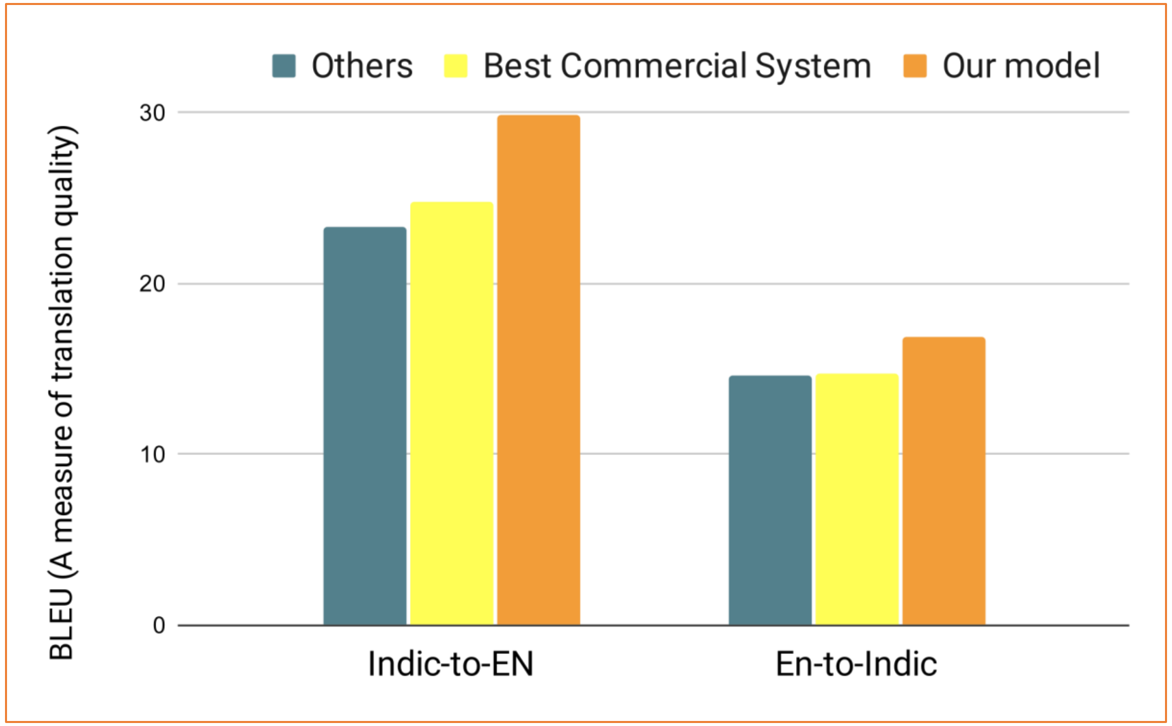
Gains are higher for low resource languages

Summary



Samanantar: Largest Parallel Corpus for Indic Languages

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



IndicTrans: State of the art translation models for En-X and X-En

https://ai4bharat.iitm.ac.in/indic-trans

What next?

DATA

Mined data from diverse sources

Manual training data

MODELS

Multilingual adapters

Simpler attention mechanisms for Indic-Indic translation

EVALUATION

Diverse Benchmark

Better evaluation metrics