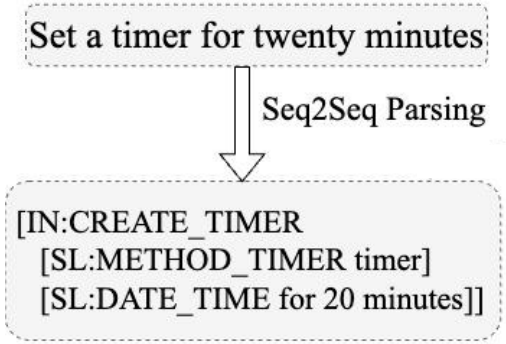
Evaluating Inter-Bilingual Semantic Parsing for Indian Languages

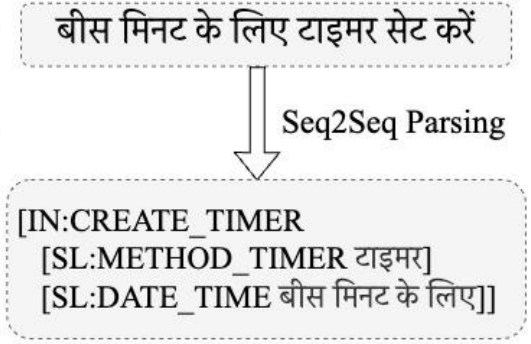
Divyanshu Aggarwal[1], Vivek Gupta[2],
Anoop Kunchukkutan[3,4]

[1]Amex AI Labs; [2]University of Utah; [3]AI4Bharat; [4]Microsoft India

Inter-Bilingual Semantic Parsing Task

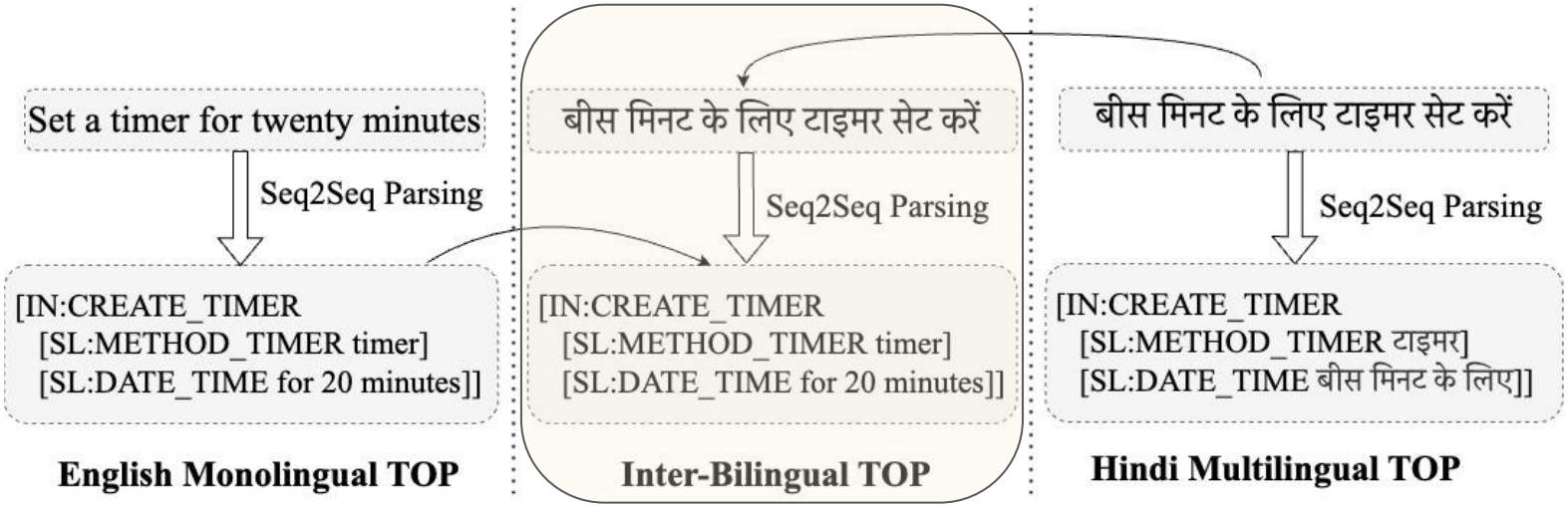


English Monolingual TOP



Hindi Multilingual TOP

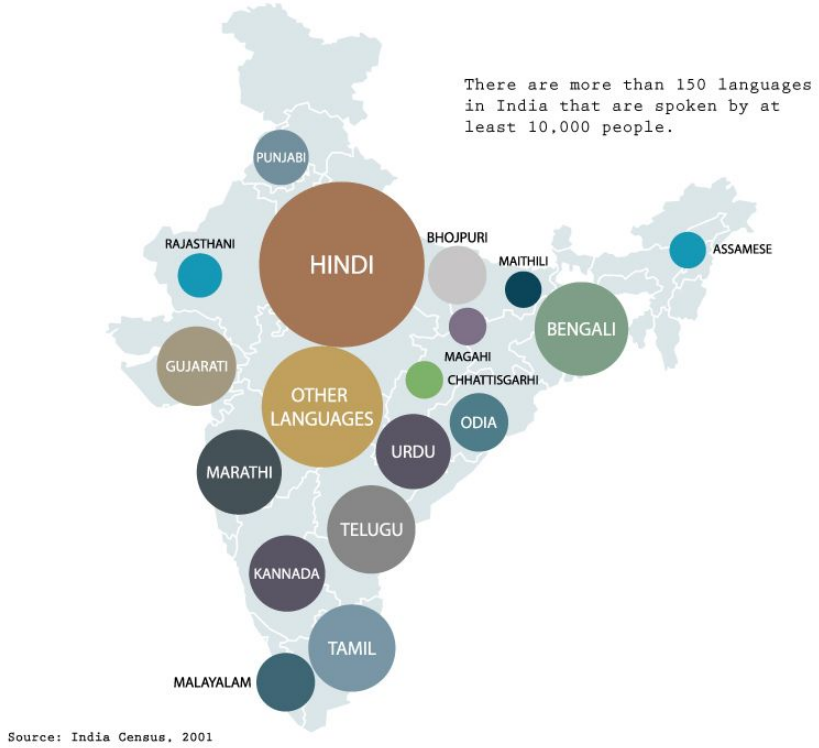
Inter-Bilingual Semantic Parsing Task



Utterance in Indic and Logical Form Slot Values in English

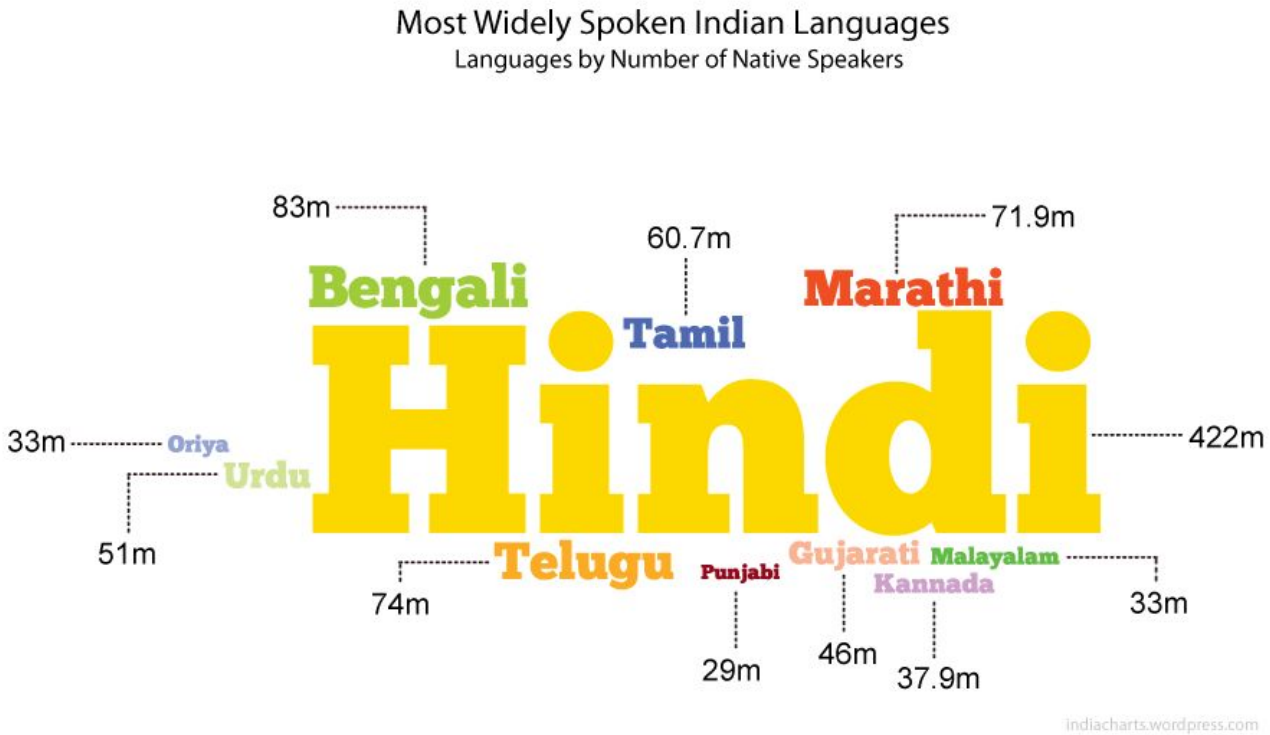
Motivation

- Indian Languages are a diverse set of languages which are spoken by more than 1.5 billion people mainly in the south asian region.
- They are also one of the largest set of internet users in the world.
- Such demographic can heavily benefit by leveraging the current advances in conversational AI.
- Yet, there is no semantic parsing dataset available due to the difficulty in process of generating logical forms for these languages.



Reference: Wikipedia

Speakers of Indian Languages



Reference: India charts

Challenges and Solutions

Challenges

- Lack of semantic parsing data in Indian Languages.
- Lack of skilled annotators who can annotate utterances to their corresponding logical forms
- Lack of Seq2seq models and extensive benchmarking of these models on semantic parsing task

Challenges and Solutions

Challenges	Solution
● Lack of semantic parsing data in Indian Languages.	● Translation of utterances of existing semantic parsing datasets using state of the art translation model.
● Lack of skilled annotators who can annotate utterances to their corresponding logical forms	● Preserve the logical form of english utterance with english slot values
● Lack of Seq2seq models and extensive benchmarking of these models on semantic parsing task	● Perform extensive analysis on existing multilingual and indic seq2seq models.

Our Contributions

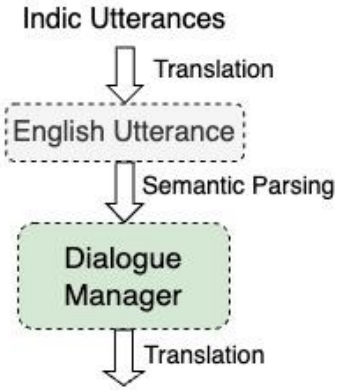
- A novel task called **Inter-Bilingual TOP**, involving **Indic utterances** as input and **logical forms with English slot values** as output.

- A dataset suite called **IE-SEMPARSE consisting of 3 semantic parsing datasets**, which encompasses 11 Indo-Dravidian languages, representing approximately 22% of the world's population.

- Exploration of various seq2seq models and different train-test strategies.

- Extensive Analysis across datasets and languages on the basis of utterance diversity, Logical form complexity and frame rareness

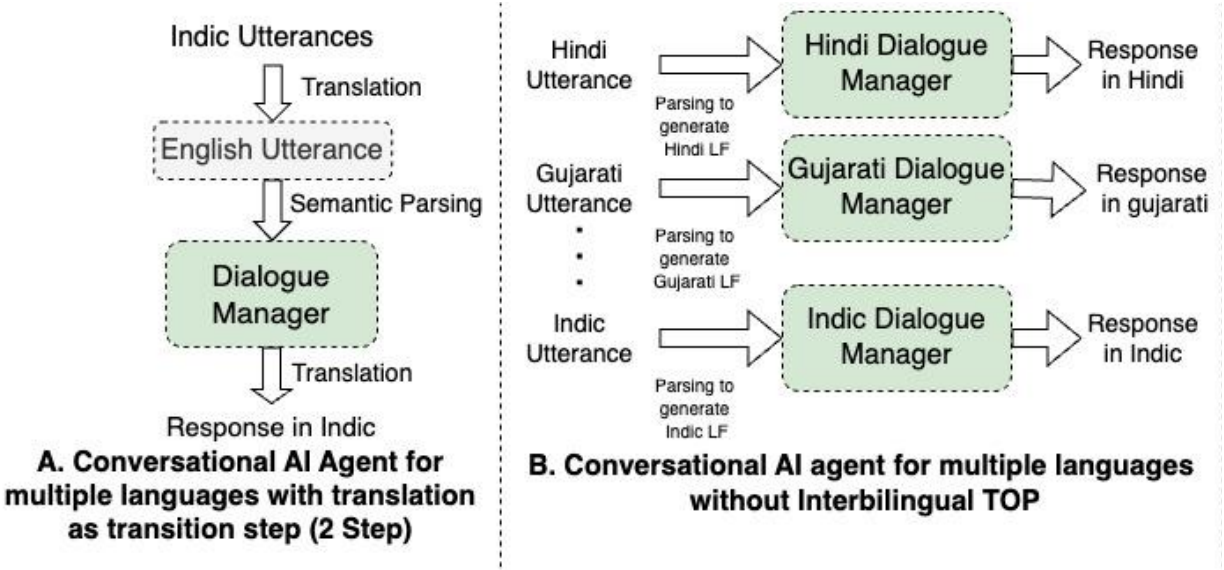
Why Inter-bilingual Semantic Parsing?



A. Conversational AI Agent for multiple languages with translation as transition step (2 Step)

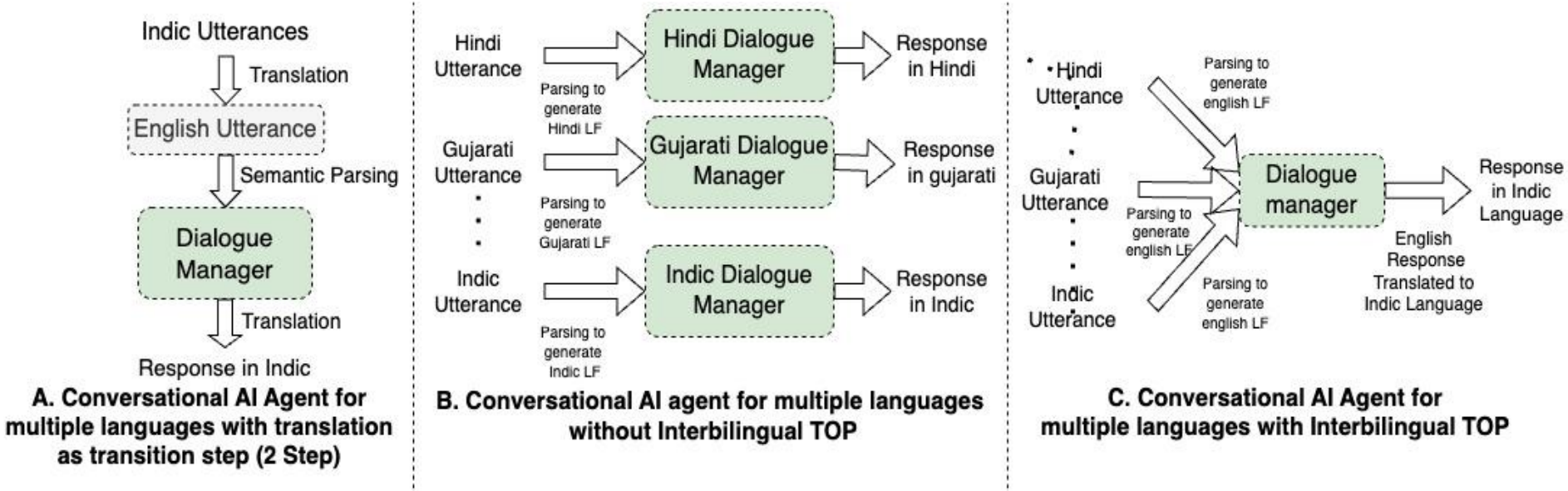
1. **Approach A:** Translate to English then parse to logical form.

Why Inter-bilingual Semantic Parsing?



1. **Approach A:** Translate to English then parse to logical form.
2. **Approach B:** Separate parser and dialogue manager for each language

Why Inter-bilingual Semantic Parsing?



1. **Approach A:** Translate to English then parse to logical form.
2. **Approach B:** Separate parser and dialogue manager for each language
3. **Approach C:** Inter-bilingual Semantic Parsing.

Inter-bilingual Semantic Parsing is a good middle ground approach to enhance model's multilingual semantic parsing ability and reduce system latency and redundancy.

Automatic Evaluation

English Translated (Round Trip)

- Capture similarity between Back translated english sentence and original english sentence.
- We used BT_BertScore to compare back translated and original english sentence.
- We compared scores on IE-mTOP, IE-multilingualTOP and IE-multiATIS++.

Automatic Evaluation

English Translated (Round Trip)

- Capture similarity between Back translated english sentence and original english sentence.
- We used BT_BertScore to compare back translated and original english sentence.
- We compared scores on IE-mTOP, IE-multilingualTOP and IE-multiATIS++.

Multilingual (Single Trip)

- Capture similarity between forward translated indic sentence and original english sentence.
- We used BertScore with XLM-R as base model and Comet Score to compare forward translated Indic sentence and original english sentence.
- We compared scores on IE-mTOP, IE-multilingualTOP and IE-multiATIS++.

Automatic Evaluation Scores

Score	Samanantar	IE-mTOP	IE-multilingualTOP	IE-multiATIS++
BertScore	0.85	0.86	0.98	0.86
CometScore	0.12	0.13	0.14	0.13
BT_BertScore	0.96	0.93	0.92	0.92

Table 1: Automatic Evaluation Scores Using BertScore, CometScore and BT_BertScore ($X10^{-2}$)

Human Evaluation

Problem

It is both time consuming and expensive to get all samples evaluated.

Furthermore, it require expert fluent speakers in all 11 Indic languages and English.

Human Evaluation

Problem	Solution
It is both time consuming and expensive to get all samples evaluated. Furthermore, it require expert fluent speakers in all 11 Indic languages and English.	Sample a relatively small diverse set (~ 2% of each test set of the dataset) of examples with maximum coverage in the test set.

Human Evaluation

It is both time consuming and expensive to get all samples evaluated.

Furthermore, it require expert fluent speakers in all 11 Indic languages and English.

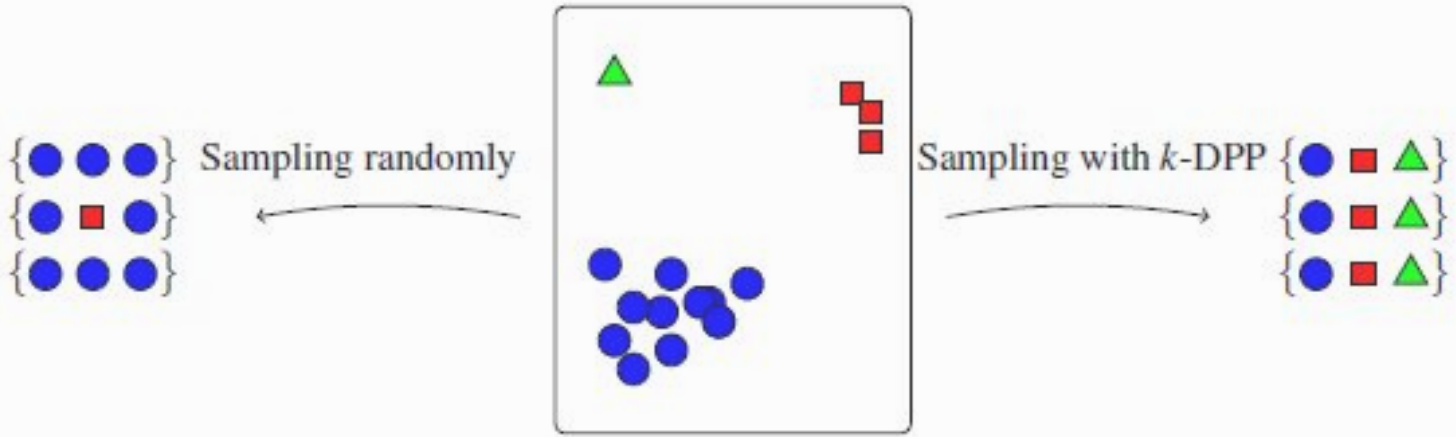
Sample a relatively small diverse set (~ 2% of each test set of the dataset) of examples with maximum coverage in the test set.

Sampled 2% of sentences from the bert embeddings of the test set of every dataset using dppy library[1] i.e. **DPP**

The diversity was also preserved by domain and intents distribution of the dataset obtained from DPP Sampling.

[1] https://github.com/guilgautier/DPPy

Diverse Sampling: What and Why?



K-DPP Process (Reference: Disney Research Studios)

Human Score Labelling

- 22 evaluators (2 for each language),
- fluent in both english and Indic mother tongue
- use Semeval-2016 Task-I guidelines[2].
- 5 Indian Rupees per sentence.

[2] https://web.eecs.umich.edu/~mihalcea/papers/agirre.semeval16.pdf

Human Evaluation Scores

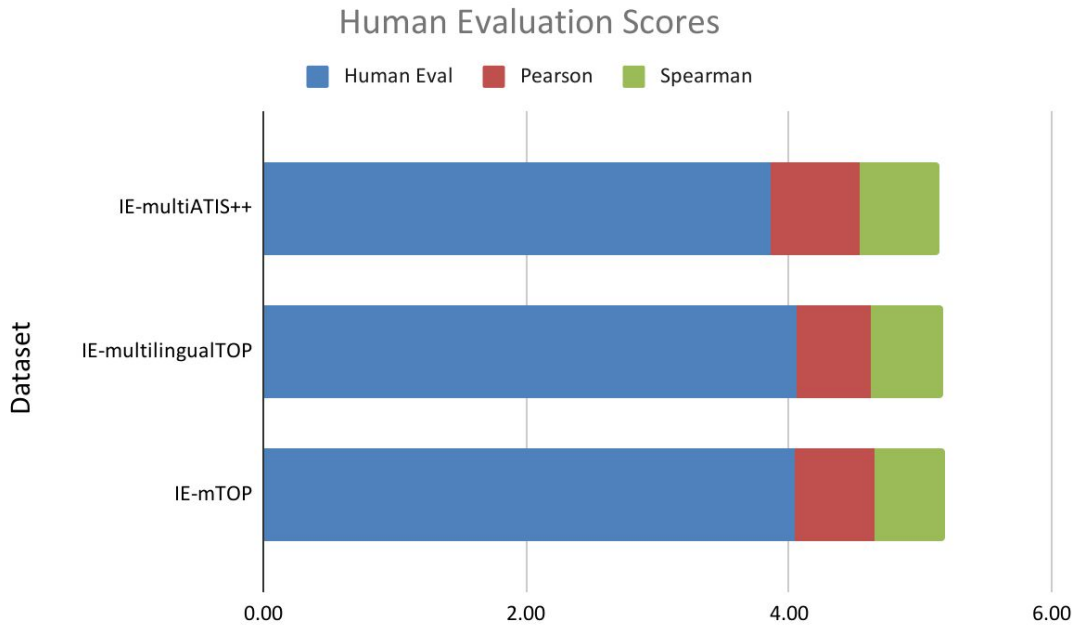
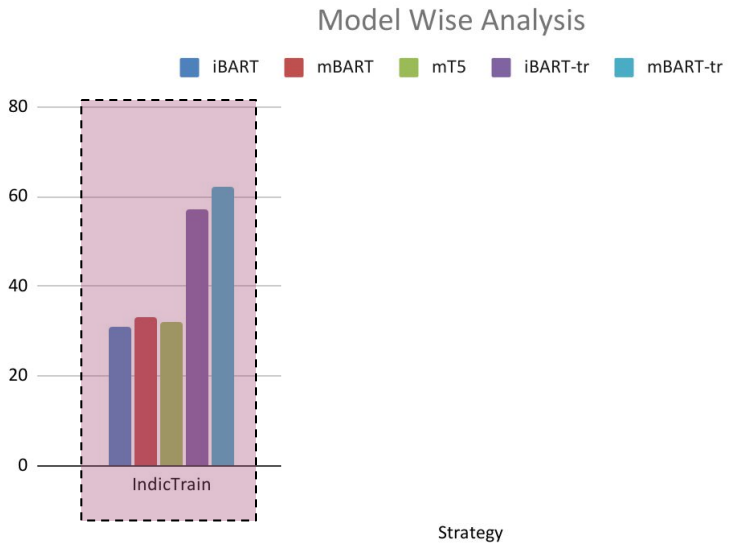


Table 2: Human Validation Score ($X10^{-2}$)

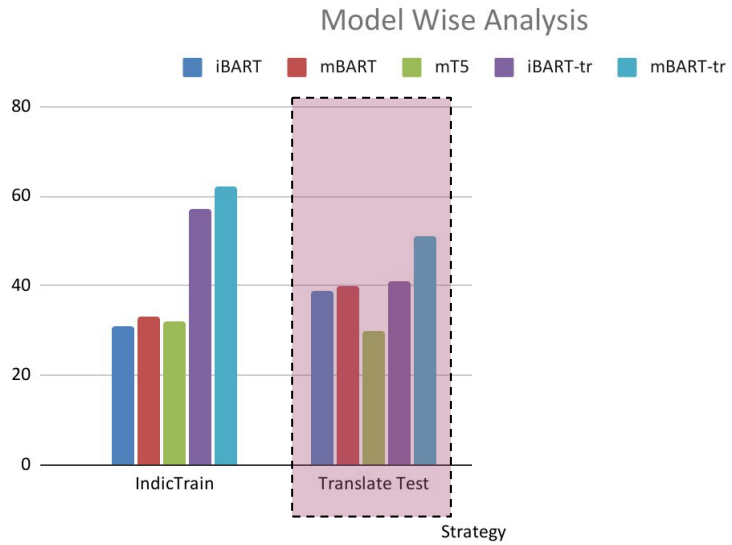
Results and Analysis



Indic Train

- Models are trained in indic language
- Models are tested in Indic Language
- This is similar to in-language training

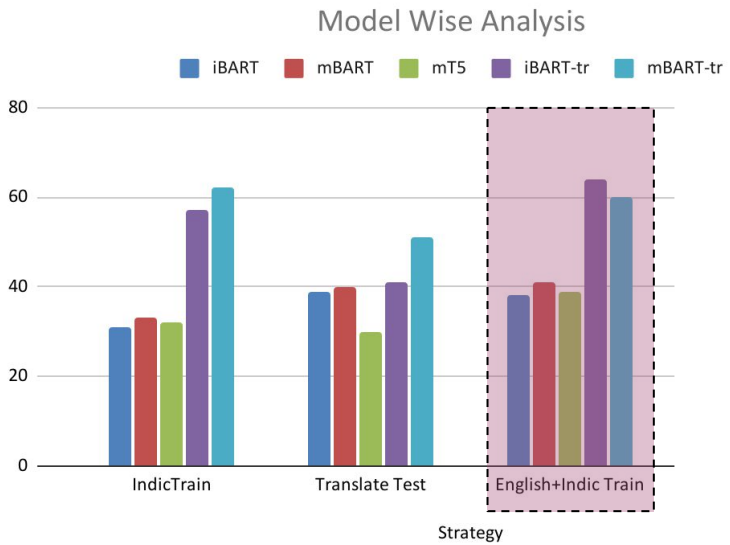
Results and Analysis



Translate Test

- The model are trained on original English semantic parsing dataset.
- The model is evaluated on back translated semantic parsing dataset.
- We also evaluated monolingual english seq2seq models described further in the paper

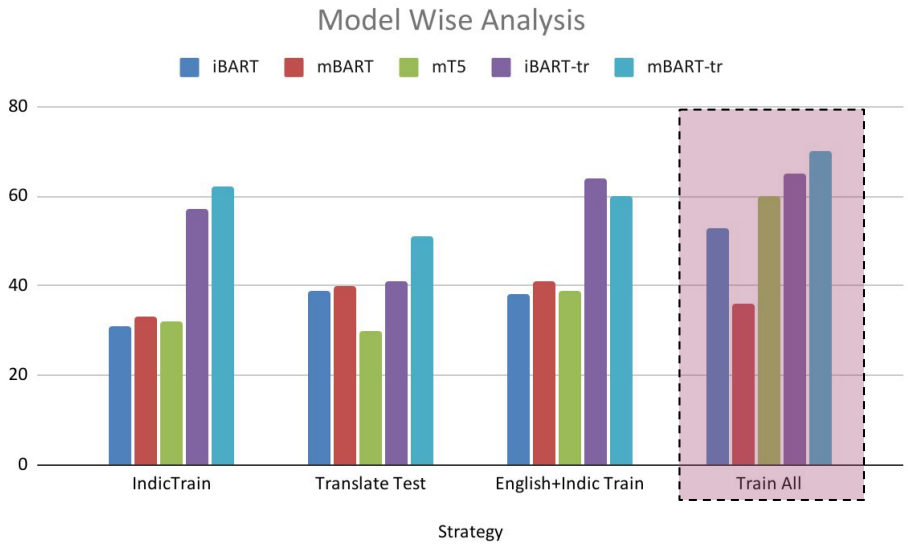
Results and Analysis



English+Indic Train

- The models are trained on first English dataset and then Indic train set data.
- The model is evaluated on Indic test set data.

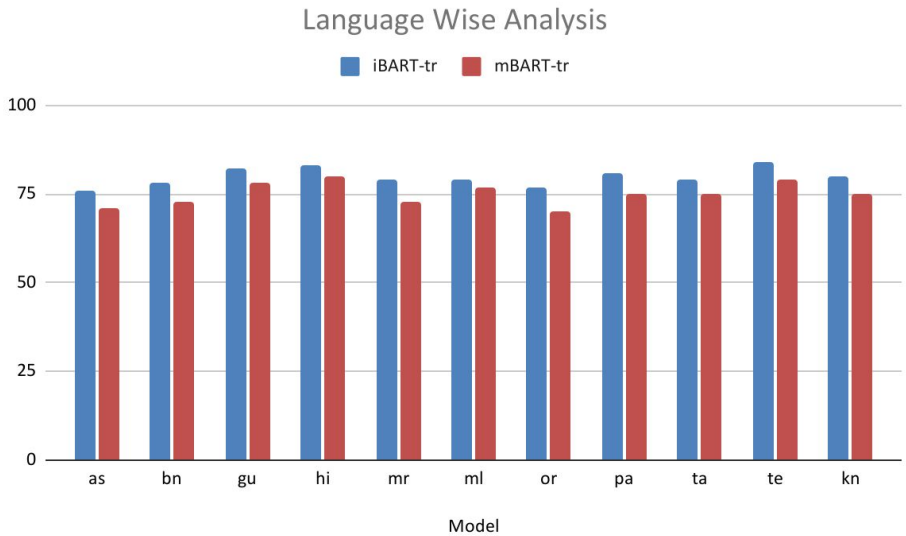
Results and Analysis



Train All

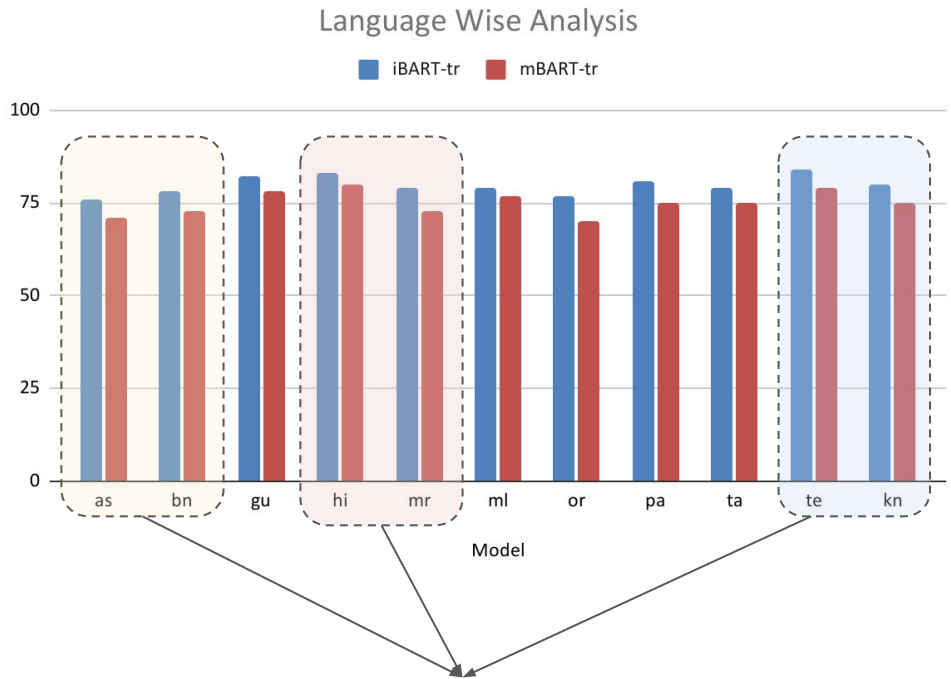
- The model is first finetuned on English data and then finetuned on Indic data of all indic languages.
- The model is evaluated on Indic test set data.
- Translation Finetuned models performs better on our tasks as compared to pretrained only models.

Language Wise Comparison



- iBart-tr > mBART-tr
- High Resource language > Mid resource language >> Low Resource Language
- Low resource languages with similar script to high resource languages perform well.
- Translation Finetuned models perform better than pretrained only models on our tasks.

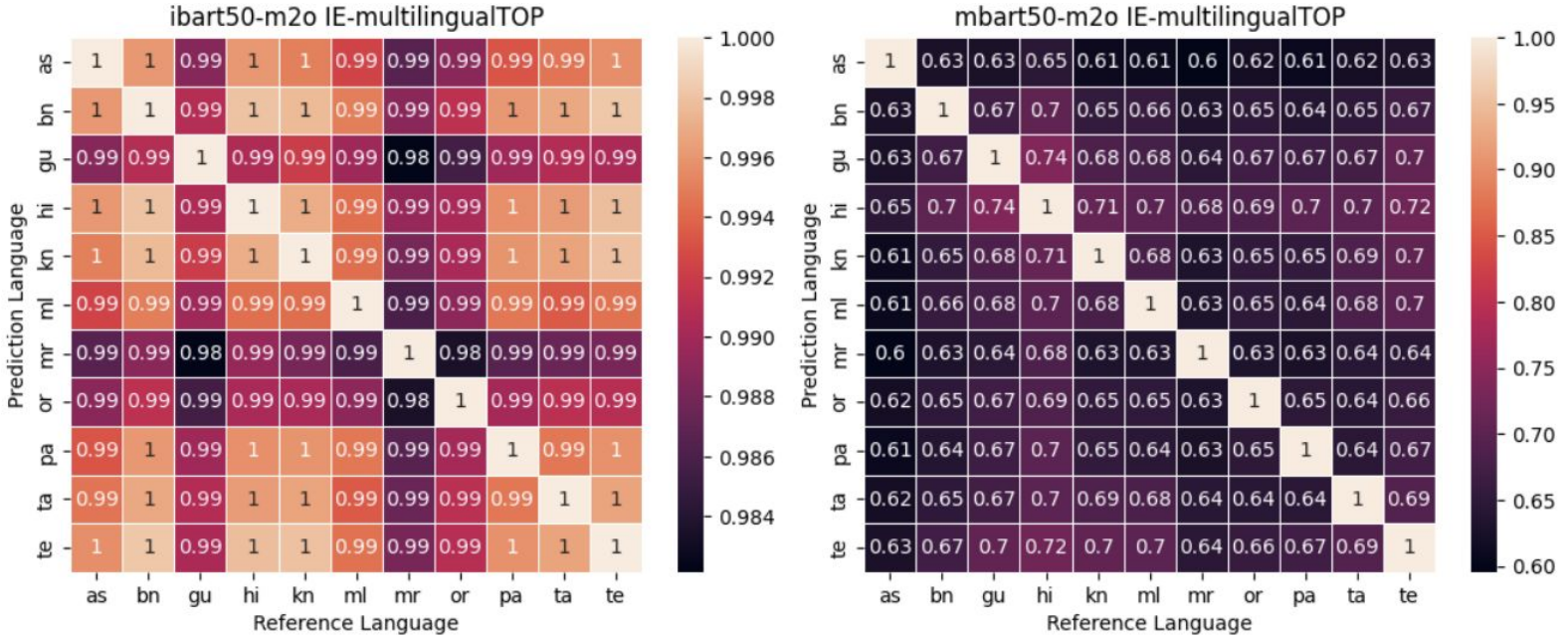
Language Wise Comparison



Pairs of languages with similar script

- iBart-tr > mBART-tr
- High Resource language > Mid resource language >> Low Resource Language
- Low resource languages with similar script to high resource languages perform similarly.
- Translation Finetuned models perform better than pretrained only models on our tasks.
- iBART-tr is able to leverage indic specific pretraining and perform better than mBART-tr

Similarity in Predictions across Languages

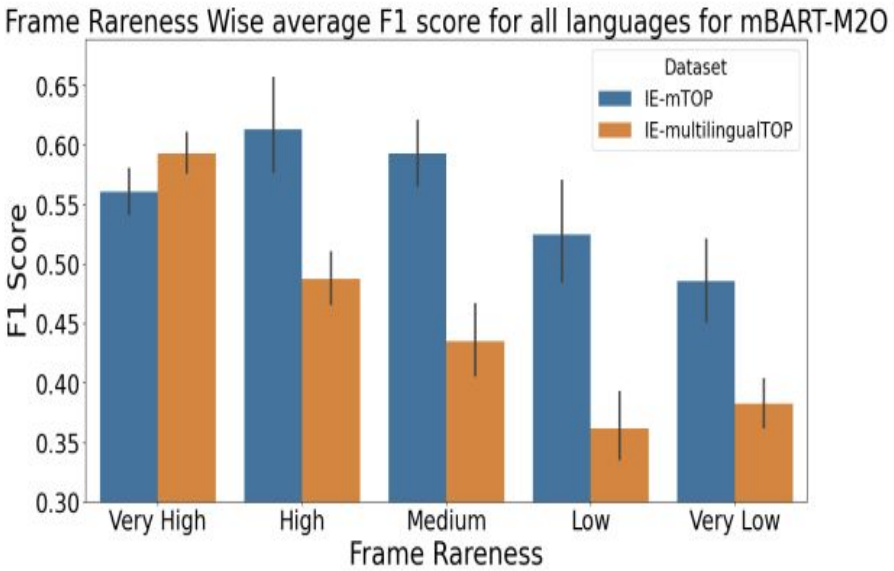


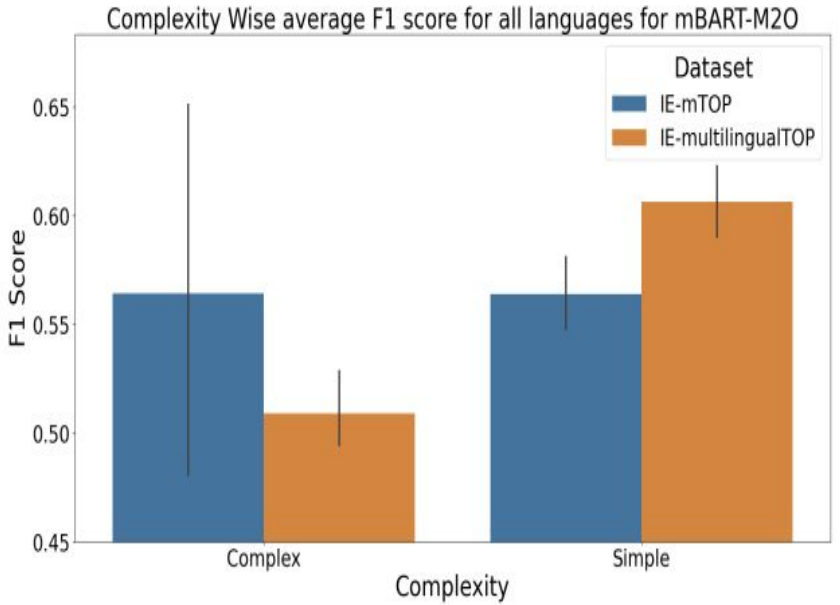
(a) IndicBART-M2O

(b) mBART-large-50-M2O

IndicBART seems to perform consistently across language, i.e. it produces the same output for an utterance regardless of the language unlike mBART

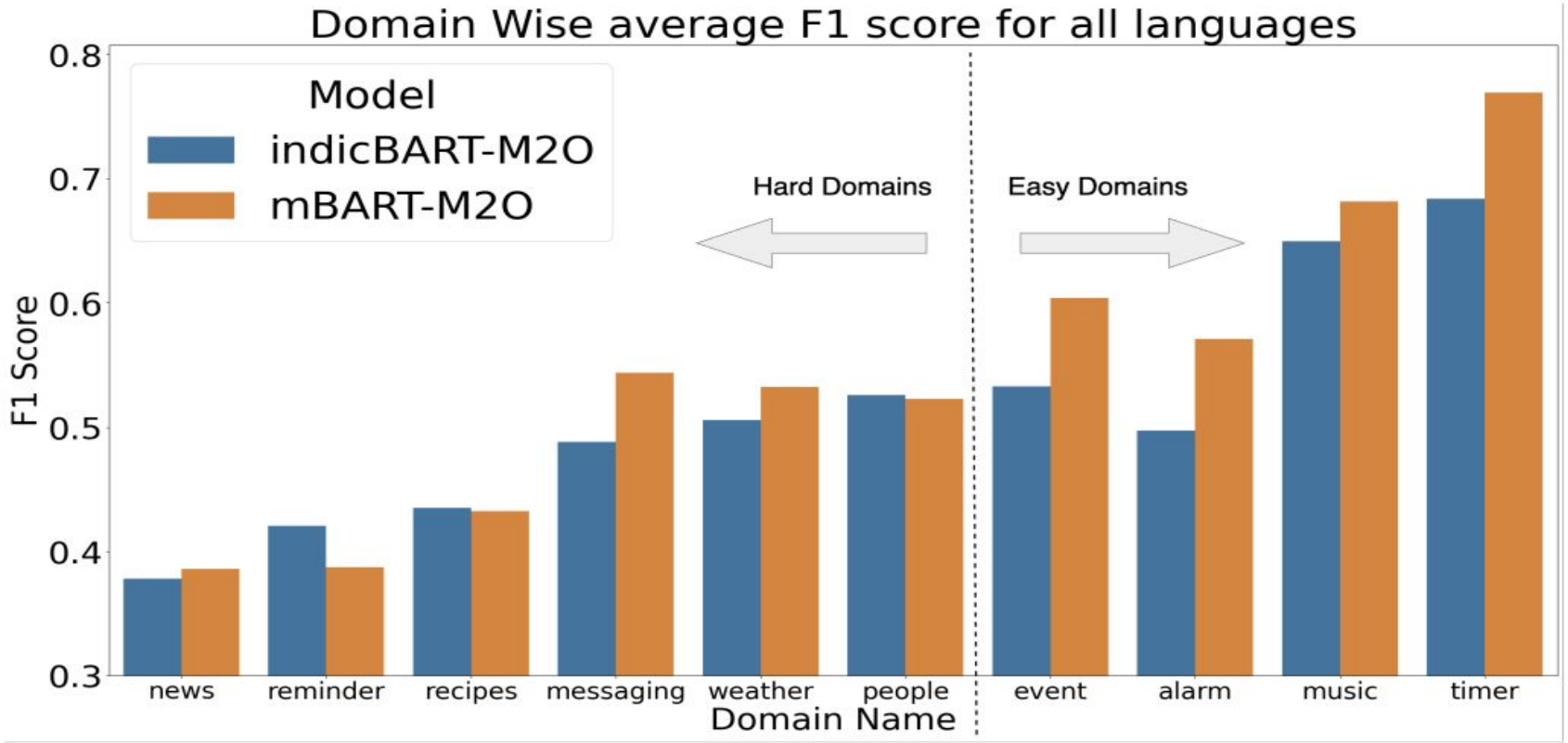
Analysis Across Datasets





IE-mTOP is a more diverse dataset in terms of domains and logical form structures.

Domain Wise Performance for IE-mTOP



General functions like timer, and music are easier to asses in Inter Bilingual Setting while functions like recipe and news are harder to parse for the model.

Key Takeaways

- With **IE-SemParse** we contribute a semantic parsing suite of 3 semantic parsing datasets for **Indic languages family** Namely **IE-mTOP, IE-multilingualTOP, IE-multiATIS++.**
- We Also introduce a novel **Inter-Bilingual Semantic Parsing Task** and discuss the benefits of that task in **multilingual conversational AI systems.**
- We Evaluate the quality of our dataset with various **automatic** and **human evaluation** techniques which are **less expensive and time consuming**.
- We benchmark **IE-SemParse** with several **multi-lingual models** using various **train-test strategies**.
- We also extensively study the behaviour of model performance a**cross languages and datasets.**
- Furthermore,
    - **Future Work:** To Contribute a dataset with Logical form slot values in Indic Language using techniques like **LINGUIST** or **Translate & Fill**.