Evaluating Inter-Bilingual Semantic Parsing for Indian Languages

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

[1]American Express, AI Labs; [2]University of Utah; [3]Microsoft; [4]AI4Bharat

divyanshu.aggarwal1@aexp.com; vgupta@cs.utah.edu ; ankunchu@microsoft.com

Abstract

Despite significant progress in Natural Language Generation for Indian languages (Indic-NLP), there is a lack of datasets around complex structured tasks such as semantic parsing. One reason for this imminent gap is the complexity of the logical form, which makes English to multilingual translation difficult. The process involves alignment of logical forms, intents and slots with translated unstructured utterance. To address this, we propose an Interbilingual Seq2seq Semantic parsing dataset IE-SEMPARSE for 11 distinct Indian languages. We highlight the proposed task's practicality, and evaluate existing multilingual seq2seq models across several train-test strategies. Our experiment reveals a high correlation across performance of original multilingual semantic parsing datasets (such as mTOP, multilingual TOP and multiATIS++) and our proposed IE-SEMPARSE suite.

1 Introduction

Task-Oriented Parsing (TOP) is a Sequence to Sequence (seq2seq) Natural Language Understanding (NLU) task in which the input utterance is parsed into its logical sequential form. Refer to Figure 1 where logical form can be represented in form of a tree with intent and slots as the leaf nodes (Gupta et al., 2018; Pasupat et al., 2019). With the development of seq2seq models with self-attention (Vaswani et al., 2017), there has been an upsurge in research towards developing *generation* models for complex TOP tasks. Such models explore numerous training and testing strategies to further enhance performance (Sherborne and Lapata, 2022; Gupta et al., 2022). Most of the prior work focus on the English TOP settings.

However, the world is largely multilingual, hence new conversational AI systems are also expected to cater to the non-English speakers. In that regard works such as mTOP (Li et al.,

Figure 1: TOP vs Bilingual TOP.

2021), multilingual-TOP (Xia and Monti, 2021), multi-ATIS++ (Xu et al., 2020; Schuster et al., 2019), MASSIVE dataset (FitzGerald et al., 2022) have attempted to extend the semantic parsing datasets to other multilingual languages. However, the construction of such datasets is considerably harder since mere translation does not provide high-quality datasets. The logical forms must be aligned with the syntax and the way sentences are expressed in different languages, which is an intricate process.

Three possible scenarios for parsing multilingual utterances exists, as described in Figure 1. For English monolingual TOP, we parse the English utterance to it's English logical form, where the slot values are in the English language. Seq2Seq models (Raffel et al., 2019; Lewis et al., 2020) tuned on English TOP could be utilized for English specific semantic parsing. Whereas, for multi lingual setting, a *Indic* multilingual TOP (e.g. Hindi Multilingual TOP in Figure 1) is used to parse Indic utterance to it's respective Indic logical form. Here, the slot values are also Indic (c.f. Figure 1).[1]

The English-only models, with their limited input vocabulary, produce erroneous translations as it requires utterance translation. The multilingual models on the other side require larger multilingual vocabulary dictionaries (Liang et al., 2023; Wang et al., 2019). Although models with large vocabulary sizes can be effective, they may not perform equally well in parsing all languages, resulting in

---

*Equal Contribution

[1] In both English and Indic Multilingual TOP, the utterance and it's corresponding logic form are in same language, English or Indic respectively.

overall low-quality output. Moreover, managing multilingual inputs can be challenging and often requires multiple dialogue managers, further adding complexity. Hence, we asked ourselves: *"Can we combine the strengths of both approaches?"*

Therefore, we explore a third distinct setting: Inter-bilingual TOP. This setting involves parsing Indic utterances and generating corresponding logical forms with English slot values (in comparison, multilingual top has non-english multilingual slot values). For a model to excel at this task, it must accurately parse and translate simultaneously. The aim of inter-bilingual semantic parsing is to anticipate the translation of non-translated logical forms into translated expressions, which presents a challenging reasoning objective. Moreover, many scenarios, such as e-commerce searches, music recommendations, and finance apps, require the use of English parsing due to the availability of search vocabulary such as product names, song titles, bond names, and company names, which are predominantly available in English. Additionally, APIs for tasks like alarm or reminder setting often require specific information in English for further processing. Therefore, it is essential to explore inter-bilingual task-oriented parsing with English slot values.

In this spirit, we establish a novel task of Inter-Bilingual task-Oriented Parsing (Bi-lingual TOP) and develop a semantic parsing dataset suite a.k.a IE-SEMPARSE for Indic languages. The utterances are translated into eleven Indic languages while maintaining the logical structures of their English counterparts.[2] We created inter-bilingual semantic parsing dataset IE-SEMPARSE Suite (IE represents Indic to English). IE-SEMPARSE suite consists of three Interbilingual semantic datasets namely IE-mTOP, IE-multilingualTOP, IE-multiATIS++ by machine translating English utterances of mTOP, multilingualTOP and multiATIS++ (Li et al., 2021; Xia and Monti, 2021; Xu et al., 2020) to eleven Indian languages described in §3. In addition, §3 includes the meticulously chosen automatic and human evaluation metrics to validate the quality of the machine-translated dataset.

We conduct a comprehensive analysis of the performance of numerous multilingual seq2seq models on the proposed task in §4 with various input combinations and data enhancements. In our experiments, we demonstrate that interbilingual parsing is more complex than English and multilingual parsing, however, modern transformer models with translation fine-tuning are capable of achieving results comparable to the former two. We also show that these results are consistent with those obtained from semantic parsing datasets containing slot values in the same languages as the utterance. Our contributions to this work are the following:

1. We proposed a novel task of Inter-Bilingual TOP with multilingual utterance (input) and English logical form (output). We introduced IE-SEMPARSE, an Inter-Bilingual TOP dataset for 11 Indo-Dravidian languages representing about 22% of speakers of the world population.

2. We explore various seq2seq models with several train-test strategies for this task. We discuss the implications of an end-to-end model compared to translation followed by parsing. We also compare how pertaining, pre-finetuning and structure of a logical form affect the model performance.

The IE-SEMPARSE suite along with the scripts will be available at https://iesemparse.github.io/.

2 Why Inter Bilingual Parsing?

In this section, we delve deeper into the advantages of our inter-bilingual parsing approach and how it affects the dialogue management and response generation. We will address the question: *"Why preserve English slot values in the logical form?"*.

Limited Decoder Vocabulary: Using only English logical forms simplifies the seq2seq model decoder by reducing its vocabulary to a smaller set. This will make the training process more stable and reduce the chances of hallucination which often occurs in decoders while decoding long sequences with larger vocabulary size (Raunak et al., 2021).

Multi-lingual Models Evaluation: In this work, we explore the unique task of translating and parsing spoken utterances into logical forms. We gain valuable insights into the strengths and weaknesses of current multilingual models on this task. Specifically, we investigate how multilingual models compare to monolingual ones, how translation finetuning affects performance, and how the performance of Indic-specific and general multilingual models

---

[2] Like previous scenarios, the slot tags and intent operators such as METHOD_TIMER and CREATE_TIMER are respectively preserved in the corresponding English languages.
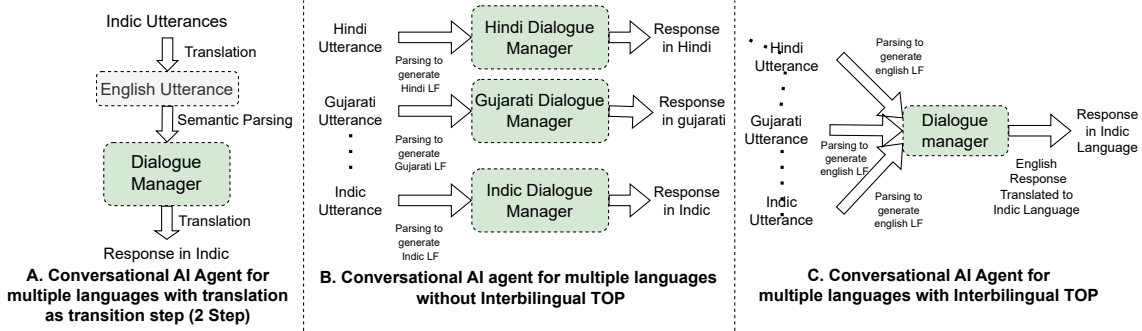
Figure 2: Conversational AI Agents comparisons with (w/o) inter-bilingual parsing. LF refers to logical form.

differ. We also analyze the predictions of the two best models across languages in §4.2, which is a novel aspect of our task. These insights enhance our understanding of existing multilingual models on IE-SEMPARSE.

Improved Parsing Latency: In figure 2, we illustrate three multilingual semantic parsing scenarios:

1. In **scenario A**, the Indic utterance is translated to English, parsed by an NLU module, and then a dialogue manager delivers an English response, which is translated back to Indic language.

2. In **scenario B**, language-specific conversational agents generate a logical form with Indic slot values, which is passed to a language-specific dialogue manager that delivers an Indic response.

3. In **scenario C**, a multilingual conversation agent generates a logical form with English slot values, which is passed to an English Dialogue Manager that delivers an English response, which is translated back into Indic language.

We observe that our approach scenario C is 2x faster than A. We further discuss the latency gains and the performances differences in appendix §A. Scenario B, on the other hand, has a significant developmental overhead owing to multilingual language, as detailed below.

Handling System Redundancy: We argue that IE-SEMPARSE is a useful dataset for developing dialogue managers that can handle multiple languages without redundancy. Unlike existing datasets such as mTOP (Li et al., 2021), multilingual-TOP (Schuster et al., 2019), and multi-ATIS++ (Xu et al., 2020), which generate logical forms with English intent functions and slot tags but multilingual slot values, our dataset generates logical forms with English slot values as well. This

avoids the need to translate the slot values or to create separate dialogue managers for each language, which would introduce inefficiencies and complexities in the system design. Therefore, our approach offers a practical trade-off between optimizing the development process and minimizing the inference latency for multilingual conversational AI agents. Finally, the utilization of a multilingual dialogue manager fails to adequately adhere to the intricate cultural nuances present in various languages (Jonson, 2002).

3 IE-SEMPARSE **Creation and Validation**

In this section, we describe the IE-SEMPARSE creation and validation process in details.

IE-SEMPARSE **Description:** We create three inter-bilingual TOP datasets for eleven major *Indic* languages that include Assamese ('as'), Gujarat ('gu'), Kannada ('kn'), Malayalam ('ml'), Marathi ('mr'), Odia ('or'), Punjabi ('pa'), Tamil ('ta'), Telugu ('te'), Hindi ('hi'), and Bengali ('bn'). Refer to the appendix §A, for additional information regarding the selection of languages, language coverage of models, and the selection of translation model. The three datasets mentioned are described below:

1. **IE-mTOP:** This dataset is a translated version of the multi-domain TOP-v2 dataset. English utterances were translated to Indic languages using IndicTrans (Ramesh et al., 2021), while preserving the logical forms.

2. **IE-multilingualTOP**: This dataset is from the multilingual TOP dataset, where utterances were translated and logical forms were decoupled using the pytext library.[3]

3. **IE-multiATIS++**: This dataset comes from the multi-ATIS++, where utterances were translated and the logical forms were generated from labelled dictionaries and decoupled, as described in appendix §3.
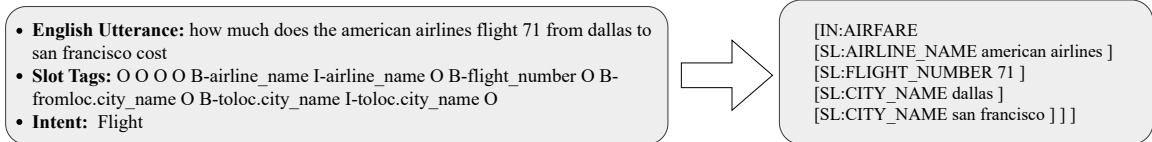
---

[3] https://github.com/facebookresearch/pytext

```
[IN:AIRFARE
 [SL:AIRLINE_NAME american airlines ]
 [SL:FLIGHT_NUMBER 71 ]
 [SL:CITY_NAME dallas ]
 [SL:CITY_NAME san francisco ] ] ]
```

Figure 3: IE-multiATIS++ Logical Form Generation

Score	Dataset	as	bn	gu	hi	kn	ml	mr	or	pa	ta	te
BertScore	**Samanantar**	0.83	0.83	0.85	0.87	0.86	0.85	0.85	0.84	0.87	0.87	0.87
	IE-mTOP	0.83	0.85	0.85	0.87	0.86	0.85	0.86	0.85	0.87	0.87	0.87
	IE-multilingualTOP	0.98	0.98	0.98	0.96	0.98	0.98	0.99	0.98	0.97	0.98	0.98
	IE-multiATIS++	0.83	0.85	0.86	0.87	0.86	0.85	0.85	0.85	0.86	0.87	0.87
CometScore	**Samanantar**	0.12	0.12	0.11	0.12	0.12	0.12	0.13	0.13	0.12	0.12	0.12
	IE-mTOP	0.12	0.13	0.12	0.12	0.12	0.13	0.13	0.13	0.14	0.12	0.12
	IE-multilingualTOP	0.13	0.14	0.14	0.13	0.14	0.14	0.14	0.14	0.14	0.14	0.14
	IE-multiATIS++	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13
BT_BertScore	**Samanantar**	0.95	0.96	0.96	0.97	0.96	0.96	0.96	0.96	0.97	0.96	0.96
	IE-mTOP	0.92	0.94	0.93	0.94	0.94	0.93	0.94	0.93	0.93	0.93	0.93
	IE-multilingualTOP	0.93	0.93	0.89	0.93	0.92	0.96	0.93	0.9	0.92	0.91	0.91
	IE-multiATIS++	0.91	0.92	0.92	0.93	0.93	0.92	0.92	0.91	0.92	0.92	0.92

Table 1: Automatic scores on IE-SEMPARSE and Benchmark Dataset Samanantar.

IE-multiATIS++ Logical Form Creation The logical forms are generated from the label dictionaries, where the Intent was labeled with 'IN:' tag and Slots were labelled with 'SL:' Tags and decoupled like IE-multilingualTOP dataset. The process of generating logical forms out of intent and slot tags from the ATIS dataset is illustrated in figure 3.

IE-SEMPARSE Processing: To construct IE-SEMPARSE we perform extensive pre and post processing, as described below:

Pre-processing We extensively preprocess IE-SEMPARSE. We use Spacy NER Tagger[4] to tag date-time and transform them into their corresponding lexical form. E.g. tag date time *"7:30 pm on 14/2/2023."* is transformed to *"seven thirty pm on fourteen february of 2023."*

Post-processing For many languages some words are commonly spoken and frequently. Therefore, we replace frequently spoken words in IE-SEMPARSE with their transliterated form, which often sounds more fluent, authentic, and informal than their translated counterparts.

To accomplish this, we replace commonly spoken words with their transliterated form to improve understanding. We created corpus-based transliteration token dictionaries by comparing Hindi mTOP, translated mTOP, and transliterated mTOP datasets. We utilize the human-translated Hindi set of mTOP dataset to filter frequently transliterated phrases and repurpose the same Hindi dictionary to post-process the text for all other Indic languages.

3.1 IE-SEMPARSE Validation

As observed in past literature, machine translation can be an effective method to generate high quality datasets (K et al., 2021; Aggarwal et al., 2022; Agarwal et al., 2022b). However, due to inherent fallibility of the machine translation system, translations may produce incorrect utterance instances for the specified logical form. Consequently, making the task more complicated and generalizing the model more complex. Thus, it is crucial to examine the evaluation dataset quality and alleviate severe limitations accurately. Early works, including Bapna et al. (2022); Huang (1990); Moon et al. (2020a,b), has established that quality estimation is an efficacious method for assessing machine translation systems in the absence of reference data a.k.a the low-resource settings.

Using Quality Estimation: In our context, where there is a dearth of reference data for the IE-SEMPARSE translated language, we also determined the translation quality of IE-SEMPARSE using a (semi) automatic quality estimation technique. Most of recent works on quality estimation compare the results with some reference data and then prove the correlation between reference scores and referenceless quality estimation scores (Fomicheva et al., 2020; Yuan and Sharoff, 2020; Cuong and Xu, 2018). Justifying and interpreting quality estimation metrics, however, remains a stiff challenge for real-world referenceless settings.

IE-SEMPARSE Automatic Benchmarking: When a parallel corpus in both languages is

---

[4] https://spacy.io/api/entityrecognizer

Dataset	Statistics	as	bn	gu	hi	kn	ml	mr	or	pa	ta	te
	Human Eval	3.15	3.07	3.65	4.1	3.7	4.12	4	4.4	4.45	4.03	3.83
IE-multiATIS++	Pearson	0.66	0.85	0.69	0.61	0.76	0.62	0.56	0.72	0.61	0.71	0.68
	Spearman	0.71	0.86	0.42	0.57	0.49	0.51	0.59	0.59	0.59	0.65	0.6
	Human Eval	3.06	3.21	3.92	4.46	4.33	4.13	4.24	4.74	4.47	4.22	3.84
IE-multilingualTOP	Pearson	0.55	0.79	0.56	0.53	0.45	0.5	0.65	0.42	0.67	0.58	0.59
	Spearman	0.57	0.74	0.54	0.53	0.45	0.46	0.62	0.63	0.51	0.5	0.49
	Human Eval	3.1	3.39	4	4.42	4.28	3.99	4	4.61	4.42	4.16	4.13
IE-mTOP	Pearson	0.66	0.74	0.64	0.55	0.61	0.63	0.73	0.45	0.51	0.5	0.62
	Spearman	0.67	0.7	0.6	0.45	0.4	0.64	0.67	0.41	0.5	0.45	0.5

Table 2: Human Evaluation Results: **Human Eval** represents the average score of 3 annotators for each language for each dataset. **Pearson** is the average pearson correlation of 1st and 2nd, 1st and 3rd and 2nd and 3rd annotators and similarly for **Spearman** which is spearman correlation.

not available, it is still beneficial to benchmark the data and translation model. In our context, we conducted an evaluation of the Samanantar corpus, which stands as the most comprehensive publicly accessible parallel corpus for Indic languages (Ramesh et al., 2021). The purpose of this assessment was to emulate a scenario wherein the Samanantar corpus serves as the benchmark reference parallel dataset, allowing us to provide a rough estimate of the scores produced by quality estimation models when evaluated in a referenceless setting on a gold standard parallel translation corpus.

We use two approaches to compare English and translated text directly. For direct quality estimation of English sentences and translated sentences in a reference-less setting, we utilize Comet Score (Rei et al., 2020) and BertScore (Zhang* et al., 2020) with XLM-RoBERTa-Large (Conneau et al., 2020) backbone for direct comparison of translated and english utterances. We also calculate BT BertScore (Agrawal et al., 2022; Moon et al., 2020a; Huang, 1990), which has shown to improve high correlation with human judgement (Agrawal et al., 2022) for our three datasets and Samanantar for reference. In this case, we translate the Indic sentence back to English and compare it with the original English sentence using BertScore (Zhang* et al., 2020). The scores for the Samanantar subset on a random subset of filtered 100k phrases and our datasets IE-SEMPARSE are provided in the table 1.

Original vs Machine Translated Hindi: As the human (translated) reference was available in mTOP and multi-ATIS for Hindi language, we leveraged that data to calculate Bert and Comet score to evaluate the translation quality of our machine translation model. We notice a high correlation between both datasets' referenceless and reference scores. Thus suggesting good translation quality for Hindi and other languages.

Dataset	Referenceless Score	Score
	Comet Score	0.83
IE-mTOP	Bert Score	0.96
	BT Bert Score	0.88
	Comet Score	0.81
IE-multiATIS++	Bert Score	0.85
	BT Bert Score	0.87

Table 3: Comet Score, BertScore and BT BertScore of Hindi dataset and translated Hindi dataset for IE-mTOP and IE-multiATIS++

In table 3 comet scores and Bert scores are scores keeping original English sentence as source, original Hindi sentence as reference and translated Hindi sentence as hypothesis. For the BT BertScore, the translated Hindi sentence and the original (human-translated) Hindi sentence are back-translated (BT) back onto English and their correlation is assessed using the Bert Score.

IE-SEMPARSE **Human Evaluation:** In our human evaluation procedure, we employ three annotators for each language [5]. We used determinantal point processes[6] (Kulesza, 2012) to select a highly diversified subset of English sentences from the test set of each dataset. We select 20 sentences from IE-multiATIS++, 120 from IE-multilingualTOP and 60 from IE-mTOP. For each dataset, this amounts to more than 1% of the total test population. We then got them scored between 1-5 from 3 fluent speakers of each Indic English and Indic language by providing them with a sheet with parallel data of English sentences and subsequent translation.

Analysis. We notice that the scores vary with resource variability where languages like "as" and "kn" have the lowest scores. However, most scores are within the range of 3.5-5 suggesting the high quality of translation for our dataset. Detailed scores are reported in Appendix §B table 7.

---

[5] every annotator was paid 5 INR for each sentence annotation each   [6] https://github.com/guilgautier/DPPy

4 Experimental Evaluation

For our experiments, we investigated into the following five train-test strategies: **1. Indic Train:** Models are both finetuned and evaluated on Indic Language. **2. English+Indic Train:** Models are finetuned on English language and then Indic Language and evaluated on Indic language data. **3. Translate Test:** Models are finetuned on English data and evaluated on back-translated English data. **4. Train All:** Models are finetuned on the compound dataset of English + all other 11 Indic languages and evaluated on Indic test dataset. **5. Unified Finetuning:** IndicBART-M2O and mBART-large-50-M2O models are finetuned on all three datasets for all eleven languages creating unified multi-genre (multi-domain) semantic parsing models for all 3 datasets for all languages. This can be considered as data-unified extension of 4th Setting.

Models: The models utilized can be categorized into four categories as follows: (a.) MULTILINGUAL such as **mBART-large-50**, **mT5-base** such as (b.) INDIC SPECIFIC such as **IndicBART** (c.) TRANSLATION PREFINETUNED such as **IndicBART-M2O**, **mBART-large-50-M2O**, which are pre finetuned on XX-EN translation task (d.) MONOLINGUAL (ENGLISH) such as **T5-base**, **T5-large**, **BART-large**, **BART-base** used only in **Translate Test** Setting. The models are specified in the table's §8 *"Hyper Parameter"* column, with details in the appendix §C. Details of the fine-tuning process with hyperparameters details and the model's vocabulary augmentation are discussed in the appendix §D and §E respectively.

Evaluation Metric: For Evaluation, we use tree labelled F1-Score for assessing the performance of our models from the original TOP paper (Gupta et al., 2018). This is preferred over an exact match because the latter can penalize the model's performance when the slot positions are out of order. This is a common issue we observe in our outputs, given that the logical form and utterance are not in the same language. However, exact match scores are also discussed in appendix §F.5.

4.1 Analysis across Languages, Models and Datasets

We report the results of **Train All** and **Unified Finetuning** settings for all datasets in table 4 and 5 in the main paper as these were the best technique out of all. The scores for other train-test strategies such as translate test, Indic Train, English+Indic

Train for all 3 datasets are reported in appendix §F.1 table 9, 10 and 11 respectively. However, we have discussed the comparison between train-test settings in the subsequent paragraphs.

Across Languages: Models perform better on high-resource than medium and low-resourced languages for **Train All** setting. This shows that the proposed inter-bilingual seq2seq task is challenging. In addition to linguistic similarities, the model performance also relies on factors like grammar and morphology (Pires et al., 2019). For other settings such as **Translate Test**, **Indic Train**, and **English+Indic**, similar observations were observed.

Across Train-Test Strategies: Translate Test method works well, however end-to-end English+Indic and Train All models perform best; due to the data augmentation setting, which increases the training size.[7] However, the benefits of train data enrichment are much greater in **Train All** scenario because of the larger volume and increased linguistic variation of the training dataset. We also discuss the comparisons in inference latency for a 2-step vs end-to-end model in §2.

Across Datasets: We observe that IE-multilingualTOP is the simplest dataset for models, followed by IE-mTOP and IE-multiATIS++. This may be because of the training dataset size, since IE-multilingualTOP is the largest of the three, followed by IE-mTOP and IE-multiATIS++. In addition, IE-multilingualTOP is derived from TOP(v1) dataset which have utterances with more simpler logical form structure (tree depth=1). IE-mTOP, on the other hand, is based on mTOP, which is a translation of TOP(v2), with more complex logical form having (tree depth>=2). We discuss the performance of models across logical form complexity in §4.2. For **Unified Finetuning** we observe an average performance gain of 0.2 in the tree labelled F1 score for all languages for all datasets as reported in table 5 in appendix.

Across Models: We analyse the performance across various models based on three criteria, language coverage, model size and translation finetuning, as discussed in detail below:

(a.) *Language Coverage:* Due to its larger size, mBART-large-50-M2O performs exceptionally well on high-resource languages, whereas IndicBART-M2O performs uniformly across all the languages due to its indic specificity. In addition, translation-optimized models perform better than

---

[7] By 2x (English + Indic) and 12x (1 English + 11 Indic).

Dataset	Model	Train All													ModAvg
		as	bn	gu	hi	kn	ml	mr	or	pa	ta	te	hi$_{IE}$	hi$_O$	
IE-mTOP	IndicBART	50	56	49	56	45	54	**67**	44	56	56	58	52	60	50
	mBART-large-50	51	53	51	**62**	51	55	51	32	53	48	52	58	66	51
	mT5-base	46	53	56	58	53	55	50	45	53	**58**	**58**	54	62	53
	IndicBART-M2O	54	57	57	61	59	58	58	57	59	57	**61**	59	63	58
	mBART-large-50-M2O	56	59	61	65	60	63	59	59	59	64	**65**	63	67	**61**
	Language Average	51	56	55	**60**	54	57	57	47	56	57	59	57	64	55
IE-multilingualTOP	IndicBART	44	50	57	**80**	43	42	50	37	67	70	77	–	–	56
	mBART-large-50	44	57	66	77	29	28	46	17	47	48	48	–	–	46
	mT5-base	49	54	57	60	56	55	52	50	53	53	**58**	–	–	54
	IndicBART-M2O	74	75	**79**	78	70	70	75	75	75	76	77	–	–	**75**
	mBART-large-50-M2O	54	57	60	**63**	58	58	53	56	57	57	61	–	–	58
	Language Average	51	56	55	**60**	54	57	57	47	56	57	59	–	–	55
IE-multiATIS++	IndicBART	51	58	52	**70**	50	41	63	25	50	39	56	66	76	54
	mBART-large-50	54	**86**	54	58	54	53	53	45	57	51	55	54	63	57
	mT5-base	67	**87**	73	73	72	78	64	59	70	68	74	70	77	72
	IndicBART-M2O	70	**90**	80	80	79	79	73	69	78	73	82	78	82	**78**
	mBART-large-50-M2O	73	**91**	83	81	77	79	75	65	78	73	79	79	83	**78**
	Language Average	63	82	68	72	66	66	66	53	67	61	69	69	76	68

Table 4: $Tree\_Labelled\_F1 * 100$ scores for the **Train All** setting. The bold numbers in the table indicate the row-wise maximum, i.e. the model's best language performance in the given context. The numbers in bold in the **ModAvg** (Model Average) column indicate the model with the best performance for the train-test strategy specified in the table's heading. Similarly, the numbers in bold in the **Language Average** row indicate the language with the best performance. Subsequently, hi$_O$ refers to the original Hindi dataset from the dataset and hi$_{IE}$ refers to the inter-bilingual dataset constructed by picking Hindi utterances and English logical form and joining them.

Dataset	Model	Unified Finetuning													ModAvg
		as	bn	gu	hi	kn	ml	mr	or	pa	ta	te	hi$_{IE}$	hi$_O$	
IE-mTOP	IndicBART-M2O	74	77	77	**81**	79	78	78	77	79	77	81	79	83	78
	mBART-large-50-M2O	76	79	81	**85**	80	83	79	79	79	84	85	83	87	**82**
	Language Average	75	78	79	**83**	80	81	79	78	79	81	83	81	85	80
IE-multilingualTOP	IndicBART-M2O	75	76	80	**79**	71	71	76	76	76	77	78	–	–	**76**
	mBART-large-50-M2O	55	58	61	**64**	59	59	54	57	58	58	62	–	–	59
	Language Average	65	67	71	**72**	65	65	65	67	67	68	70	–	–	67
IE-multiATIS++	IndicBART-M2O	80	80	90	**90**	89	89	83	79	88	83	92	88	92	84
	mBART-large-50-M2O	83	82	93	**91**	87	89	85	75	88	83	89	89	93	84
	Language Average	82	82	92	**91**	88	89	84	77	88	83	**91**	89	93	84

Table 5: $Tree\_Labelled\_F1 * 100$ scores of **IndicBART-M2O** and **mBART-large-50** model trained on all languages and all datasets. Other notations similar to that of Table 4.

those that are not. mBART-large-50 outperforms mT5-base despite its higher language coverage, while mBART-large-50's superior performance can be ascribed to its denoising pre-training objective, which enhances the model's ability to generalize for the *"intent"* and *"slot"* detection task. In section §4.2 we discuss more about the complexity of the logical forms.

(b.) *Model Size:* While model size has a significant impact on the Translate Test setting for monolingual models, we find that pre-training language coverage and Translation fine-tuning are still the most critical factors. For example, despite being a smaller model, IndicBART outperforms mT5-base on average for similar reasons. Another reason for better performance for IndicBART and mBART-large-50 denoising based seq2seq pre-training vs multilingual multitask objective of mT5-base.

(c.) *Translation Finetuning:* The proposed task is a mixture of semantic parsing and translation. We also observe this empirically, when models finetuned for translation tasks perform better. This result can be attributed to fact that machine translation is the most effective strategy for aligning phrase embeddings by multilingual seq2seq models (Voita et al., 2019), as emphasized by Li et al. (2021). In addition, we observe that the models perform best in the **Train All** setting, indicating that data augmentation followed by fine-tuning enhances performance throughout all languages on translation fine-tuned models.

Original vs Translated Hindi: We also evaluated the performance of Hindi language models on original datasets (hi$_O$) and (hi$_{IE}$) which combine Hindi utterances with logical forms of English of mTOP and multi-ATIS++ datasets, as shown in ta-

ble 4. Inter-bilingual tasks pose a challenge and result in lower performance, but translation-finetuned models significantly reduce this gap. Model performance is similar for both 'hi' and 'hi$_{IE}$', indicating the quality of translations. Additional details can be refered in Appendix §G.

Domain Wise Comparison: IE-mTOP dataset contains domain classes derived from mTOP. We compare the average F1 scores for different domains in IE-mTOP dataset for IndicBART-M2O and mBART-large-50-M2O in the **Train All** setting, as shown in Figure 4. We observe that mBART-large-50-M2O outperforms IndicBART-M2O for most domains except for people and recipes, where both perform similarly well due to cultural variations in utterances.
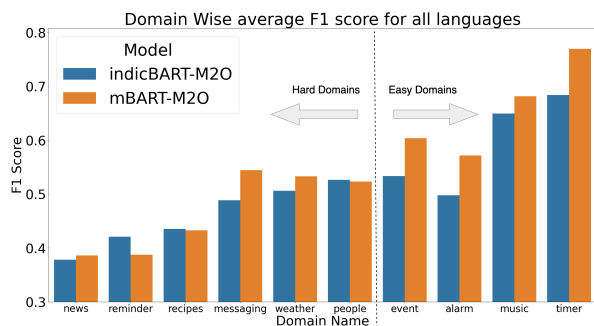

Figure 4: Domain Wise all language average F1 score in IE-mTOP dataset for IndicBART-M2O and mBART-large-50-M2O.

4.2 Analysis on Logical Forms

In this paper, we maintain the slot values in the English language and ensure consistency in the logical form across languages for each example in every dataset. This can be useful in assessing the model performance across language and datasets on the basis of logical form structure which we have analysed in this section. Previous works have shown a correlation between model performance and logical form structures (Gupta et al., 2022).

Logical Form Complexity: We evaluate the performance of the mBART-large-50-M2O model on utterances with simple and complex logical form structures in the Train All setting for IE-mTOP and IE-multilingualTOP datasets. Simple utterances have a flat representation with a single intent, while complex utterances have multiple levels [8] of branching in the parse tree with more than one intent. In IE-multiATIS++, instances are only attributed to simple utterances since they have a single unique intent. Figure 5 shows, that mBART-

---

[8] depth $>= 2$

large-50-M2O performs better for complex utterances in IE-mTOP, while there is better performance for simple utterances in IE-multilingualTOP due to its larger training data size and a higher proportion of simple logical forms in training data.
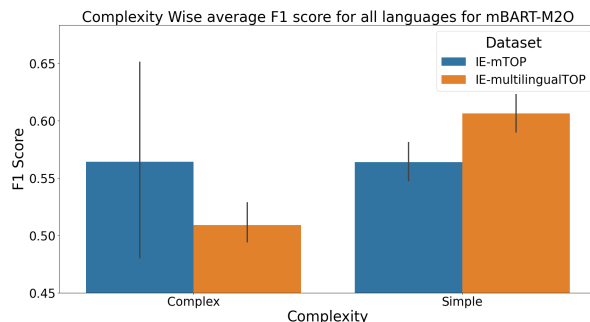

Figure 5: Complexity Wise all language average F1 score in IE-mTOP dataset for IE-mTOP and IE-multilingualTOP for mBART-large-50-M2O.

Effect of Frame Rareness: We compared mBART-large-50-M2O and IE-multilingualTOP on the Train All setting by removing slot values from logical forms and dividing frames into five frequency buckets[9]. A shown in figure 6, F1 scores increase with frame frequency, and IE-mTOP performs better for smaller frequencies while IE-multilingualTOP performs better for very large frequencies. This suggests that IE-mTOP has more complex utterances, aiding model learning with limited data, while IE-multilingualTOP's larger training size leads to better performance in very high frequency buckets.
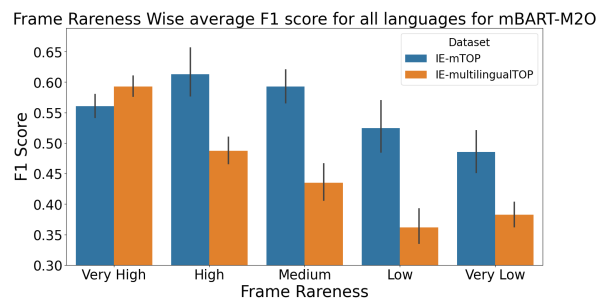
Figure 6: Frame Rareness Wise all language average F1 score in IE-mTOP dataset for IE-mTOP and IE-multilingualTOP for mBART-large-50-M2O.

Post Translation of Slot Values: We translate slot values from Hindi to English using IndicTrans for the logical forms of 'hi' mTOP and 'hi' multiATIS++ datasets in the Train All setting. Table 6 compares the F1 scores of models for IE-mTOP and IE-multiATIS++ datasets, which only had the original Hindi dataset available. Despite minor decreases in scores and visible translation errors, our

---

[9] namely very high, high, medium, low and very low.

approach yields accurate translations due to the short length of slot values and the high-resource nature of Hindi. However, we argue that our proposed task or multilingual TOP task is superior in terms of latency and performance, as discussed in §2 and §4.1.

Dataset	Model	F1
IE-mTOP	IndicBART	49
	mBART-large-50	55
	mT5-base	50
	IndicBART-M2O	56
	mBART-large-50-M2O	58
IE-multiATIS++	IndicBART	55
	mBART-large-50	67
	mT5-base	41
	IndicBART-M2O	68
	mBART-large-50-M2O	70

Table 6: Tree Labelled F1 scores of hindi dataset with post translation of slot values to english for IE-mTOP and IE-multiATIS++

Language Wise Correlation: We compared the logical form results of each language by calculating the average tree labelled F1 score between the datasets of one language to the other. We then plotted correlation matrices[10] and analysed performance on all datasets using IndicBART-M2O and mBART-large-50-M2O in **Train All** setting, as described in Figure 7, 8, and 9 in Appendix §F.4.

Our analysis shows that IndicBART-M2O has more consistent predictions than mBART-large-50-M2O. We also observed that models perform most consistently for the IE-multiATIS++ dataset. Additionally, related languages, such as 'bn' and 'as', 'mr' and 'hi', and 'kn' and 'te', have high correlation due to script similarity.

5 Related Work

Multi-Lingual Semantic Parsing: Recently, TOP has attracted a lot of attention due to the development of state-of-the-art seq2seq models such as BART (Lewis et al., 2020) and T5 (Raffel et al., 2019). Moreover, several works have extended TOP to the multilingual setting, such as mTOP, multilingual-TOP, and multi-ATIS++. The recent MASSIVE dataset (FitzGerald et al., 2022) covers six Indic languages vs eleven in our work, and only contains a flat hierarchical structure of semantic parse. Furthermore, the logical form annotations in MASSIVE are not of a similar format to those in the standard TOP dataset.

---
[10] for 11 x 11 pairs

IndicNLP: Some works have experimented with code-mixed Hindi-English utterances for semantic parsing tasks, such as CST5 (Agarwal et al., 2022a). In addition to these advances, there have been significant contributions to the development of indic-specific resources for natural language generation and understanding, such as IndicNLG Suite Kumar et al. (2022), IndicBART Dabre et al. (2022), and IndicGLUE Kakwani et al. (2020). Also, some studies have investigated the intra-bilingual setting for multilingual NLP tasks, such as IndicXNLI (Aggarwal et al., 2022) and EI-InfoTabs (Agarwal et al., 2022b). In contrast to prior works, we focus on the complex structured semantic parsing task.

LLMs and Zero Shot: Our work is also related to zero-shot cross-lingual (Sherborne and Lapata, 2022) and cross-domain (Liu et al., 2021) semantic parsing, which aims to parse utterances in unseen languages or domains. Moreover, recent methods use scalable techniques such as automatic translation and filling (Nicosia et al., 2021) and bootstrapping with LLMs (Awasthi et al., 2023; Rosenbaum et al., 2022; Scao, 2022) to create semantic parsing datasets without human annotation. Unlike previous methods such as Translate-Align-Project (TAP) (Brown et al., 1993) and Translate and Fill (TAF) (Nicosia et al., 2021), which generate semantic parses of translated sentences, they propose a novel approach that leverages LLMs to generate semantic parses of multilingual utterances.

6 Conclusion and Future Work

We present a unique inter-bilingual semantic parsing task, and publish the IE-SEMPARSE suite, which consists of 3 inter-bilingual semantic parsing datasets for 11 Indic languages. Additionally, we discuss the advantages of our proposed approach to semantic parsing over prior methods. We also analyze the impact of various models and train-test procedures on IE-SEMPARSE performance. Lastly, we examine the effects of variation in logical forms and languages on model performance and the correlation between languages.

For future work, we plan to release a SOTA model, explore zero-shot parsing (Sherborne and Lapata, 2022), enhance IE-SEMPARSE with human translation (NLLB Team et al., 2022), explore zero-shot dataset generation (Nicosia et al., 2021), leverage LLM for scalable and diverse dataset generation(Rosenbaum et al., 2022; Awasthi et al., 2023), and evaluate instruction fine-tuning models.

7 Limitations

One of the main limitations of our approach is the use of machine translation to create the IE-SEMPARSE suite. However, we showed that the overall quality of our dataset is comparable to Samanantar, a human-verified translation dataset. Furthermore, previous studies Bapna et al. (2022); Huang (1990); Moon et al. (2020a,b) have shown the effectiveness of quality estimation in reference-less settings. Lastly, we have also extensively evaluated our dataset with the help of 3 human evaluators for each language as described in §3. We can further take help of GPT4 in future to evaluate the translations in a scaled manner (Gilardi et al., 2023).

The second point of discussion focuses on the motivation for preserving logical form slot values in English. We explore the use cases where querying data in English is crucial, and how this approach can enhance models by reducing latency, limiting vocabulary size, and handling system redundancy. While open-source tools currently cannot achieve this, it would be valuable to evaluate the effectiveness of this task by comparing it with the other two discussed approaches. To accomplish this, we suggest using a dialogue manager and scoring the performance of its responses on the three TOP approaches outlined in the paper.

Another potential limitation of our dataset is that it may contain biases and flaws inherited from the original TOP datasets. However, we contend that spoken utterances are generally simpler and more universal than written ones, which mitigates the risk of cultural mismatches in IE-SEMPARSE dataset. Furthermore, our work is confined only to the Indo-Dravidian Language family of Indic languages due to our familiarity with them and the availability of high-quality resources from previous research. Nonetheless, our approach is easily extendable to other languages with effective translation models, enabling broader applications in various languages worldwide. In the future, we plan to improve our datasets by publicly releasing them through initiatives like NLLB or IndicTransV2, and by collaborating with larger organizations to have the test sets human-translated.

8 Acknowledgements

References

Anmol Agarwal, Jigar Gupta, Rahul Goel, Shyam Upadhyay, Pankaj Joshi, and Rengarajan Aravamudhan. 2022a. Cst5: Data augmentation for code-switched semantic parsing.

Chaitanya Agarwal, Vivek Gupta, Anoop Kunchukuttan, and Manish Shrivastava. 2022b. Bilingual tabular inference: A case study on indic languages. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4018–4037, Seattle, United States. Association for Computational Linguistics.

Divyanshu Aggarwal, Vivek Gupta, and Anoop Kunchukuttan. 2022. IndicXNLI: Evaluating multilingual inference for Indian languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10994–11006, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sweta Agrawal, Nikita Mehandru, Niloufar Salehi, and Marine Carpuat. 2022. Quality estimation via back-translation at the wmt 2022 quality estimation task. In *Proceedings of the Seventh Conference on Machine Translation*, pages 593–596, Abu Dhabi. Association for Computational Linguistics.

Abhijeet Awasthi, Nitish Gupta, Bidisha Samanta, Shachi Dave, Sunita Sarawagi, and Partha Talukdar. 2023. Bootstrapping multilingual semantic parsers using large language models.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, Theresa Breiner, Vera Axelrod, Jason Riesa, Yuan Cao, Mia Xu Chen, Klaus Macherey, Maxim Krikun, Pidong Wang, Alexander Gutkin, Apurva Shah, Yanping Huang, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. 2022. Building machine translation systems for the next thousand languages.

Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Yvette Graham, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, and Matteo Negri, editors.

2020. *Proceedings of the Fifth Conference on Machine Translation*. Association for Computational Linguistics, Online.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

Hoang Cuong and Jia Xu. 2018. Assessing quality estimation models for sentence-level prediction. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1521–1533, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh Khapra, and Pratyush Kumar. 2022. IndicBART: A pre-trained model for indic natural language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1849–1863, Dublin, Ireland. Association for Computational Linguistics.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2022. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised Quality Estimation for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 8:539–555.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks.

Sonal Gupta, Rushin Shah, Mrinal Mohit, Anuj Kumar, and Mike Lewis. 2018. Semantic parsing for task oriented dialog using hierarchical representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2787–2792, Brussels, Belgium. Association for Computational Linguistics.

Vivek Gupta, Akshat Shrivastava, Adithya Sagar, Armen Aghajanyan, and Denis Savenkov. 2022. RetroNLU: Retrieval augmented task-oriented semantic parsing. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 184–196, Dublin, Ireland. Association for Computational Linguistics.

Barry Haddow and Faheem Kirefu. 2020. Pmindia – a collection of parallel corpora of languages of india.

Xiuming Huang. 1990. A machine translation system for the target language inexpert. In *COLING 1990 Volume 3: Papers presented to the 13th International Conference on Computational Linguistics*.

Rebecca Jonson. 2002. Multilingual nlp methods for multilingual dialogue systems.

Karthikeyan K, Aalok Sathe, Somak Aditya, and Monojit Choudhury. 2021. Analyzing the effects of reasoning types on cross-lingual transfer performance. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 86–95, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

Alex Kulesza. 2012. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2-3):123–286.

Aman Kumar, Himani Shrotriya, Prachi Sahu, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, Amogh Mishra, Mitesh M. Khapra, and Pratyush Kumar. 2022. Indicnlg suite: Multilingual datasets for diverse nlg tasks in indic languages.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Haoran Li, Abhinav Arora, Shuohui Chen, Anchit Gupta, Sonal Gupta, and Yashar Mehdad. 2021. MTOP: A comprehensive multilingual task-oriented semantic parsing benchmark. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2950–2962, Online. Association for Computational Linguistics.

Davis Liang, Hila Gonen, Yuning Mao, Rui Hou, Naman Goyal, Marjan Ghazvininejad, Luke Zettlemoyer, and Madian Khabsa. 2023. Xlm-v: Overcoming the vocabulary bottleneck in multilingual masked language models.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Zihan Liu, Genta Indra Winata, Peng Xu, and Pascale Fung. 2021. X2Parser: Cross-lingual and cross-domain framework for task-oriented compositional semantic parsing. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 112–127, Online. Association for Computational Linguistics.

Jihyung Moon, Hyunchang Cho, and Eunjeong L. Park. 2020a. Revisiting round-trip translation for quality estimation. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 91–104, Lisboa, Portugal. European Association for Machine Translation.

Jihyung Moon, Hyunchang Cho, and Eunjeong L. Park. 2020b. Revisiting round-trip translation for quality estimation. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 91–104, Lisboa, Portugal. European Association for Machine Translation.

Massimo Nicosia, Zhongdi Qu, and Yasemin Altun. 2021. Translate & Fill: Improving zero-shot multilingual semantic parsing with synthetic data. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3272–3284, Punta Cana, Dominican Republic. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Panupong Pasupat, Sonal Gupta, Karishma Mandyam, Rushin Shah, Mike Lewis, and Luke Zettlemoyer. 2019. Span-based hierarchical semantic parsing for task-oriented dialog. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1520–1526, Hong Kong, China. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2021. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183, Online. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Andy Rosenbaum, Saleh Soltan, Wael Hamza, Yannick Versley, and Markus Boese. 2022. Linguist: Language model instruction tuning to generate annotated utterances for intent classification and slot tagging. In *COLING 2022*.

Teven Le Scao. 2022. Bloom: A 176b-parameter open-access multilingual language model. *ArXiv*, abs/2211.05100.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3795–3805, Minneapolis, Minnesota. Association for Computational Linguistics.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface.

Tom Sherborne and Mirella Lapata. 2022. Zero-shot cross-lingual semantic parsing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4134–4153, Dublin, Ireland. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. The bottom-up evolution of representations in the transformer: A study with machine translation and language modeling objectives. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4396–4406, Hong Kong, China. Association for Computational Linguistics.

Elena Voita, Rico Sennrich, and Ivan Titov. 2021. Analyzing the source and target contributions to predictions in neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1126–1140, Online. Association for Computational Linguistics.

Hai Wang, Dian Yu, Kai Sun, Jianshu Chen, and Dong Yu. 2019. Improving pre-trained multilingual model with vocabulary expansion. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 316–327, Hong Kong, China. Association for Computational Linguistics.

Menglin Xia and Emilio Monti. 2021. Multilingual neural semantic parsing for low-resourced languages. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 185–194, Online. Association for Computational Linguistics.

Weijia Xu, Batool Haider, and Saab Mansour. 2020. End-to-end slot alignment and recognition for cross-lingual NLU. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5052–5063, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Yu Yuan and Serge Sharoff. 2020. Sentence level human translation quality estimation with attention-based neural networks. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1858–1865, Marseille, France. European Language Resources Association.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A   Further Discussions

Why Indic Languages?:   Indic languages are a set of Indo-Aryan languages spoken mainly in the Indian subcontinent. These languages combined are spoken by almost 22% of the total world population in monolingual, bilingual, or multilingual ways. these speakers also are the 2nd largest population of smartphone users, and almost everyone interacts with AI through chatbots. Hence it poses an excellent opportunity for NLP researchers to push state-of-the-art further for standard NLU tasks in these languages to benefit the digital business perspective and make technology more accessible to people through AI. However, most NLU benchmarks lack datasets in those languages despite some being high resource (such as 'hi,' 'bn,' and 'pa'). Moreover, with the introduction of various NLU models like IndicBERT (Kakwani et al., 2020), indicCorp, indicBART (Kumar et al., 2022), and state-of-the-art NMT module IndicTrans (Ramesh et al., 2021) that has opened new opportunities for researchers to innovate and contribute benchmark datasets which support building NLU models for Indic languages.

Lastly, discourse in languages other than English helps society understand more diverse perspectives and leads to a more inclusive society. As the world is mainly multilingual, various studies have proven that multilingual people can contribute more diverse societal perspectives through digital discourse.

Why IndicTrans translation?   Furthermore we use IndicTrans because of the following three reasons, (a.) **Lightweight**: IndicTrans is an extremely lightweight yet state of the art machine translation model for Indic languages. (b.) **Indic Coverage**: IndicTrans covers the widest variety of Indic languages as compared to other models like mBART, mT5 and google translate and azure translate are not free for research. (c.) **Open Source:** IndicTrans is open source and free for research purposes, more on this is elaborated in Aggarwal et al. (2022).

Why Inter-Bilingual TOP task? Task-Oriented Parsing has seen significant advances in recent years with the rise of attention models in deep learning. There have been significant extensions of this dataset in the form of mTOP (Li et al., 2021) and multilingual-TOP (Xia and Monti, 2021). However, they remain limited in terms of language coverage, only covering a few major global languages and only Hindi in the Indic category.

These datasets are especially difficult to expand to other languages due to the fact that each language has a unique word order and the logical form of each sentence should be modified accordingly. They cannot be altered using a simple dictionary lookup or alignment technique to generate a high-quality dataset. In keeping with this, we propose an inter bilingual TOP task in which only input utterances are translated. As current computers continue to employ English to make decisions and interact with the outside world, modern dialogue managers can work with the logical forms of the English counterparts, construct a response, and translate it back to the input utterance's language.

This resolves the latency issue where the model must first convert the statement to English before parsing it with another seq2seq model. This was mentioned in section §4.1 which demonstrates that end to end models perform better than translate + parsing models in certain instances. Despite the difficulties of learning translation and parsing in a single set of hyper parameters, our research demonstrates that this is feasible with existing seq2seq models, especially models that have being pretrained with translation task.

Task Oriented Parsing in the era of ChatGPT: With the rising popularity of chatGPT [11] in open-domain conversational AI. It is still a challenge to actually use these large language models in a task-oriented manner. Moreover, these open domain models may not understand the intent of the user correctly or they may take incorrect actions provided a user utterance. These LLMs also have the risk of being biased and toxic. Recent works like HuggingGPT (Shen et al., 2023) have also shown that while these models may have outstanding language understanding capabilities, it is still better to use task specific models to execute tasks in a narrow scope.

---

[11] https://openai.com/blog/chatgpt

Model Coverages: Listed below is the language coverage for all employed multilingual models.

1. **mBART-large-50:** 'bn', 'gu', 'hi', 'ml', 'mr', 'ta', 'te'

2. **mT5-base:** 'bn', 'gu', 'hi', 'kn', 'ml', 'mr', 'pa', 'ta', 'te'

3. **IndicBART**: ''as', bn', 'gu', 'hi', 'kn', 'ml', 'mr', 'or', 'pa', 'ta', 'te'

4. **IndicBART-M2O**: ''as', bn', 'gu', 'hi', 'kn', 'ml', 'mr', 'or', 'pa', 'ta', 'te'

5. **mBART-large-50-M2O**: 'bn', 'gu', 'hi', 'ml', 'mr', 'ta', 'te'

Two-step vs End2End parsing: We measure the translation time of IndicTrans (Ramesh et al., 2021) on an NVIDIA T4 GPU and find that it takes 0.015 seconds on average to translate a single utterance from one language to another. In scenario A, this adds 0.03 seconds of latency per utterance, while our approach only adds 0.015 seconds ($\approx \frac{1}{2}$). In scenario B, where the logical form has slot values in Indic, there is no latency overhead for either approach, but there are significant development challenges due to multilingualism as discussed below.

B   Details: Human Evaluation

In table 7 we show the detailed scores of human evaluation process discussed in the main paper §3.

C   Details: Multilingual Models

1. **Generic Multilingual (Multilingual):** these models are generic Seq2Seq multilingual models, we used mBART-large-50, mT5-base (Liu et al., 2020; Xue et al., 2021) for experiments for this category.

2. **Indic Specific (Indic):** These seq2seq models are specifically pretrained on Indic data, we uexplore IndicBART for experiments (Dabre et al., 2022) in this category.

3. **Translation Finetuned (Translation):** These pretrained seq2seq models are finetuned on the translation task with a single target language i.e. English. The models we explored form this category areIndicBART-M2O and mBART-large-50-M2O (Dabre et al., 2022; Tang et al., 2021).

Dataset	Score	as	bn	gu	hi	kn	ml	mr	or	pa	ta	te
	$Score_1$	3.1	3	3.8	4.3	3.9	4.2	4.1	4.9	4.6	3.8	4.4
	$Score_2$	3	3	3.1	3.7	3.8	3.7	3.5	4	4.5	4.5	3.5
	$Score_3$	3.4	3.3	4.1	4.4	3.4	4.5	4.5	4.4	4.3	3.9	3.6
	$Pearson_{1,2}$	0.8	0.8	0.9	0.8	0.8	0.7	0.6	0.8	0.6	0.7	0.1
IE-multiATIS++	$Pearson_{1,3}$	0.6	0.9	0.2	0.5	0.8	0.7	0.4	0.6	0.7	0.7	0
	$Pearson_{2,3}$	0.6	0.8	0.1	0.5	0.6	0.5	0.6	0.7	0.6	0.8	0.7
	$Spearman_{1,2}$	0.8	0.8	0.8	0.7	0.4	0.5	0.6	0.6	0.3	0.7	0.1
	$Spearman_{1,3}$	0.7	0.9	0.2	0.5	0.8	0.8	0.5	0.6	0.5	0.7	0.1
	$Spearman_{2,3}$	0.6	0.9	0.2	0.6	0.3	0.3	0.7	0.6	0.1	0.6	0.7
	$Score_1$	2.9	3	4	4.6	4.4	4.4	4.3	4.9	4.7	4.1	4.4
	$Score_2$	3.1	3.2	3.7	4.2	4.3	4.2	4.2	4.7	4.5	4.1	3.6
	$Score_3$	3.2	3.5	4	4.6	4.3	3.8	4.3	4.7	4.3	4.5	3.5
	$Pearson_{1,2}$	0.7	0.8	0.5	0.7	0.5	0.7	0.6	0.6	0.7	0.6	0.4
IE-multilingualTOP	$Pearson_{1,3}$	0.6	0.7	0.4	0.5	0.3	0.4	0.7	0.4	0.7	0.4	0.5
	$Pearson_{2,3}$	0.4	0.8	0.7	0.4	0.6	0.4	0.6	0.2	0.6	0.8	0.9
	$Spearman_{1,2}$	0.7	0.8	0.4	0.5	0.4	0.5	0.6	0.5	0.5	0.6	0.4
	$Spearman_{1,3}$	0.6	0.7	0.4	0.3	0.3	0.4	0.7	0.3	0.5	0.3	0.4
	$Spearman_{2,3}$	0.4	0.8	0.8	0.3	0.6	0.4	0.6	0.1	0.5	0.6	0.7
	$Score_1$	2.9	3.2	4.2	4.3	4.5	4.3	4.1	4.8	4.7	4.2	4.5
	$Score_2$	2.8	3.5	3.8	4.2	4	3.9	3.9	4.4	4.2	4	4.3
	$Score_3$	3.2	3.6	4	4.7	4.3	3.8	4	4.6	4.4	4.3	3.6
	$Pearson_{1,2}$	0.8	0.7	0.6	0.7	0.5	0.6	0.8	0.4	0.4	0.4	0.3
IE-mTOP	$Pearson_{1,3}$	0.6	0.8	0.5	0.4	0.8	0.6	0.7	0.3	0.2	0.4	0.3
	$Pearson_{2,3}$	0.5	0.7	0.7	0.5	0.5	0.7	0.7	0.6	0.1	0.7	0.6
	$Spearman_{1,2}$	0.9	0.7	0.6	0.6	0.4	0.6	0.8	0.4	0.3	0.3	0.3
	$Spearman_{1,3}$	0.6	0.7	0.5	0.3	0.5	0.7	0.6	0.4	0.2	0.3	0.5
	$Spearman_{2,3}$	0.5	0.7	0.7	0.5	0.3	0.6	0.6	0.7	0.3	0.5	0.4

Table 7: Detailed Human Evaluation Scores. $Score_x$ refers to the average score of the column language given by x annotator. $Pearson_{x,y}$ refers to the person correlation between the scores of annotators x and y for the column language and similarly for $Spearman_{x,y}$

4. **Monolingual (Monolingual):** These seq2seq models are pretrained on English data only. They were utilize only in the Translate Test setting. The models we explored form this category are T5-large, T5-base (Raffel et al., 2019) and BART-base, BART-large (Lewis et al., 2020).

D   Hyperparameters Details

In Table 8 the hyperparamaters are abbreviated as mentioned below:

1. **PO:** Pre-training Objective.

2. **PD:** Pretraining Dataset,

3. **LR:** Learning Rate,

4. **BS:** Batch Size,

5. **NE:** Maximum Number of Epochs,

6. **WD:** Weight Decay,

7. **MSL:** Maximum Sequence Length,

8. **MS:** Model Size described as a number of parameters in millions,

9. **WS**: Warm-up Step.

All the experiments were run on RTX A5000 GPUs in Jarvis labs [12]. The code was written in PyTorch and Huggingface accelerate library [13]. We used early stopping callback in training process with patience of 2 epochs for each setting.

The Average runtime for each for T5-base, BART-base, IndicBART, IndicBART-M2O was 3 minutes for IE-mTOP, 1 minute for IE-multiATIS++ and 5 minutes for IE-multilingualTOP. The Average runtime for each for T5-large, BART-large, mT5-base,mBART-large-50, mBART-large-50-M2O was 5 minutes for IE-mTOP, 3 minute for IE-multiATIS++ and 10 minutes for IE-multilingualTOP.

E   Vocabulary Augmentation

Unique Intents and slots from each dataset (IE-mTOP, IE-multilingualTOP, IE-multiATIS++) were extracted and added to the tokenizer and model vocabulary so that the models could predict them more accurately. In a typical slot and intent tagging task, these tags would have been treated as classes in the classification model. However, since our models are trained to not predict the entire word but only subwords (Raffel et al., 2019; Lewis et al., 2020) as usually done in modern self-attention architecture (Vaswani et al., 2017), we

---

[12] https://jarvislabs.ai/

[13] https://huggingface.co/docs/accelerate/index

Hyper Parameter	MS	LR	WD	MSL	BS	NE	PO	PD
BART-base	139	3.00e-3	0.001	64	128	50	Deniosing Autoencoder	Wikepedia Data (Lewis et al., 2020)
BART-large	406	3.00e-5	0.001	64	16	50	Deniosing Autoencoder	Wikepedia Data
T5-base	222	3.00e-3	0.001	64	256	50	Multi task Pretraining	C4 (Raffel et al., 2019)
T5-large	737	3.00e-5	0.001	64	16	50	Multi task Pretraining	C4
IndicBART	244	3.00e-3	0.001	64	128	50	Deniosing Autoencoder	Indic Corp (Kakwani et al., 2020)
mBART-large-50	610	1.00e-4	0.001	64	16	50	Deniosing Autoencoder	CC25(Liu et al., 2020)
mT5-base	582	3.00e-4	0.001	64	16	50	Multi task Pretraining	mC4 (Xue et al., 2021)
IndicBART-M2O	244	3.00e-3	0.001	64	128	50	Deniosing Autoencoder	PM India (Haddow and Kirefu, 2020)
mBART-large-50-M2O	610	1.00e-4	0.001	64	16	50	Deniosing Autoencoder	WMT16 (Barrault et al., 2020)

Table 8: Hyper Parameters and Pretraining Details

decided to include them in the vocabulary so that they can be generated easily during prediction run-time. This also contributed to the reduction of the maximum sequence length to 64 tokens, which improved generalisation as seq2seq models generalise better on shorter sequences (Voita et al., 2021). The Excel spreadsheet containing unique slots and intents will be made accessible alongside the code and supplemental materials.

F Additional Results

F.1 Other Train Test Settings

We include the results of all other settings except Train All (Already discussed in main paper) in table 9 till 15. We have discussed the comparisons of these settings in main paper §4.1.

F.2 Translate Test vs End2End models

While the performance of Monolingual models in the Translate Test setting is adequate, the performance of models in the end-to-end Train All setting outperform. Translation is prone to error, and the acquired logical form in English cannot be guaranteed to be precise. Moreover, a two-step approach to translation followed by parsing will incur greater execution time than a unified model.

F.3 Unified Models Results

In unified models, we observe a gain of atleast 0.15 in all languages for all datasets for both IndicBART-M2O and mBART-large-50-M2O.

F.4 Language verses Language

From figure 7, 8, 9 we observe that IndicBART-M2O is a more consistent than mBART-large-50-M2O.

F.5 Exact Match Results

We calculated modified exact match scores as inspired by Awasthi et al. (2023) which are agnostic of the positions of the slot tokens in the logical form. These scores are presented in tables 12, 13,

14, 15. We observed that exact match is a stricter metric as compared to tree labelled F1 (Gupta et al., 2018). We also observe that exact match scores are consistent with tree labelled F1 scores across languages, datasets and models.

G Original verses Interbilingual Hindi

As demonstrated by figure 1, we have data accessible in Hindi for all three settings. To produce Hindi bilingual TOP data, we utilize mTOP and multi-ATIS++ to internally combine Hindi and English data tables by unique id (uid). To construct our dataset, we filter the Hindi utterances column and the English logical form columns; we refer to these datasets as hi_{IE} in table 4. Furthermore, we conduct tests using original Hindi datasets (slot values in Hindi in logical form) and compare their performance to that of other languages. In the table 4, we refer to these datasets as hi_{O} for the mTOP dataset and multi-ATIS++ dataset both.

Analysis. We see a decline in F1 score for all models for hi_{IE} in both IE-mTOP and IE-multiATIS++. This might be due to data loss when hindi and english data are combined, as not all utterances of english data are included in both datasets. Furthermore, the hindi utterances in the original dataset may be more complex. The results for hi_{O} and hi_{O} enhances because the tokens were copied from the utterance and the model does not have to transform the tokens to English.

117

| Dataset | Model | Translate Test | | | | | | | | | | | ModAvg |
		as	bn	gu	hi	kn	ml	mr	or	pa	ta	te	
IE-mTOP	BART-base	28	37	35	**42**	35	38	39	35	36	41	33	36
	BART-large	30	41	38	44	38	41	41	39	38	**46**	36	39
	T5-base	31	44	41	**49**	41	43	43	41	42	47	41	42
	T5-large	29	43	39	**47**	39	42	42	40	40	44	38	40
	IndicBART	30	40	36	42	36	40	39	38	37	**42**	33	38
	mT5-base	34	43	40	**48**	40	43	43	38	40	45	38	41
	mBART-large-50	18	20	20	**23**	20	19	23	16	21	**23**	21	20
	IndicBART-M2O	35	44	43	51	44	46	44	41	42	**49**	41	44
	mBART-large-50-M2O	36	45	45	50	45	47	46	41	46	**53**	43	**45**
	Language Average	30	40	37	**44**	38	40	40	37	38	43	36	38
IE-multilingualTOP	BART-base	11	15	**16**	**16**	13	14	13	14	14	14	16	14
	BART-large	12	18	19	**20**	16	16	15	16	16	16	19	17
	T5-base	8	11	12	**13**	11	11	11	11	11	11	**13**	11
	T5-large	7	9	10	**11**	8	8	8	9	9	8	10	9
	IndicBART	20	29	31	**32**	27	29	25	26	27	25	31	27
	mT5-base	20	26	26	**28**	25	25	24	23	25	24	27	25
	mBART-large-50	26	34	35	**38**	34	35	33	30	34	32	36	33
	IndicBART-M2O	20	27	29	**30**	27	28	25	25	26	25	29	26
	mBART-large-50-M2O	30	42	45	**46**	41	44	41	38	41	39	45	**41**
	Language Average	17	23	25	**26**	22	23	22	21	23	22	25	23
IE-multiATIS++	BART-base	15	**20**	14	18	17	18	14	18	17	16	18	17
	BART-large	15	20	14	15	19	19	14	**21**	16	17	20	17
	T5-base	46	**70**	52	62	61	65	47	51	58	51	66	57
	T5-large	49	**74**	58	66	62	70	48	52	63	53	70	60
	IndicBART	44	**66**	46	56	54	63	47	46	58	49	63	54
	mT5-base	25	25	18	26	24	26	19	**27**	25	20	24	24
	mBART-large-50	55	70	58	70	66	**71**	60	56	68	59	68	64
	IndicBART-M2O	44	61	48	55	52	**68**	48	53	56	47	59	54
	mBART-large-50-M2O	53	70	68	**76**	67	73	63	62	69	56	71	**66**
	Language Average	38	**53**	42	49	47	**53**	40	43	48	41	51	46

Table 9: $Tree\_Labelled\_F1 * 100$ scores for the all the dataset for **Translate Test** settings. **ModAvg** is shorthand for Model Average. The bold numbers in the table indicate the row-wise maximum, i.e. the model's best language performance in the given context. The numbers in bold in the **ModAvg** column indicate the model with the best performance for the train-test strategy specified in the table's heading. Similarly, the numbers in bold in the **Language Average** row indicate the language with the best performance for that train-test strategy.

| Dataset | Model | Indic Train | | | | | | | | | | | Model Average |
		as	bn	gu	hi	kn	ml	mr	or	pa	ta	te	
IE-mTOP	IndicBART	19	**55**	35	53	33	30	50	15	31	45	44	37
	mBART-large-50	41	51	14	**60**	22	25	25	4	44	0	57	31
	mT5-base	30	22	28	52	50	**54**	36	8	36	53	15	35
	IndicBART-M2O	50	55	45	61	55	58	58	53	13	56	**59**	51
	mBART-large-50-M2O	55	59	61	**66**	56	63	57	52	53	59	63	**59**
	Language Average	39	48	37	**58**	43	46	45	26	35	43	48	43
IE-multilingualTOP	IndicBART	36	29	24	**65**	48	9	56	30	37	42	40	38
	mBART-large-50	51	55	35	55	55	54	54	50	34	55	**57**	50
	mT5-base	45	**56**	56	20	23	49	47	47	10	37	**56**	41
	IndicBART-M2O	50	56	60	**63**	60	20	55	15	57	57	62	50
	mBART-large-50-M2O	52	60	62	**65**	60	59	57	57	51	58	64	**59**
	Language Average	47	51	47	**54**	49	38	**54**	40	38	50	56	48
IE-multiATIS++	IndicBART	12	16	8	**25**	15	19	22	22	23	22	18	19
	mBART-large-50	16	18	10	30	10	10	18	13	**33**	20	15	18
	mT5-base	15	**39**	16	18	24	18	25	6	11	35	28	22
	IndicBART-M2O	34	**86**	63	68	73	74	57	63	64	63	71	68
	mBART-large-50-M2O	71	**92**	82	81	69	80	72	4	66	74	82	**70**
	Language Average	30	**50**	36	44	38	40	39	22	39	43	43	39

Table 10: $Tree\_Labelled\_F1 * 100$ scores for the all the dataset for **Indic Train** setting. The numbers in bold in the **Model Average** column indicate the model with the best performance for the train-test strategy specified in the table's heading. Similarly, the numbers in bold in the **Language Average** row indicate the language with the best performance for that train-test strategy.

Dataset	Model	English+Indic Train											Model Average
		as	bn	gu	hi	kn	ml	mr	or	pa	ta	te	
IE-mTOP	**IndicBART**	34	37	42	**58**	41	35	54	10	42	44	43	40
	mBART-large-50	50	52	58	56	54	51	55	0	42	**59**	57	49
	mT5-base	31	25	45	**60**	48	36	44	21	6	46	48	37
	IndicBART-M2O	51	54	57	60	57	58	54	57	57	55	**62**	57
	mBART-large-50-M2O	57	60	60	65	62	**66**	58	55	58	65	64	**61**
	Language Average	45	46	52	**60**	52	49	53	29	41	54	55	49
IE-multilingualTOP	**IndicBART**	43	45	52	53	47	40	**57**	30	47	38	49	46
	mBART-large-50	0	35	35	39	0	56	48	22	58	0	**60**	32
	mT5-base	14	53	**56**	50	53	50	50	48	52	51	**56**	48
	mBART-large-50-M2O	56	60	63	**66**	61	60	57	57	60	60	64	60
	IndicBART-M2O	54	56	60	**63**	60	58	54	57	24	57	**63**	**55**
	Language Average	33	50	53	**54**	44	53	53	43	48	41	58	48
IE-multiATIS++	**IndicBART**	34	12	12	**58**	25	21	65	12	30	16	37	29
	mBART-large-50	43	22	69	**78**	14	54	58	12	36	10	66	42
	mT5-base	25	36	28	38	33	**44**	23	23	35	30	35	32
	mBART-large-50-M2O	21	**86**	78	74	73	76	56	64	72	65	75	**67**
	IndicBART-M2O	71	**87**	77	77	71	82	74	54	45	71	82	72
	Language Average	39	49	53	**65**	43	55	55	33	44	38	59	48

Table 11: $Tree\_Labelled\_F1 * 100$ scores for the all the dataset for **English+Indic Train** setting. The numbers in bold in the **Model Average** column indicate the model with the best performance for the train-test strategy specified in the table's heading. Similarly, the numbers in bold in the **Language Average** row indicate the language with the best performance for that train-test strategy.

(a) IndicBART-M2O

(b) mBART-large-50-M2O
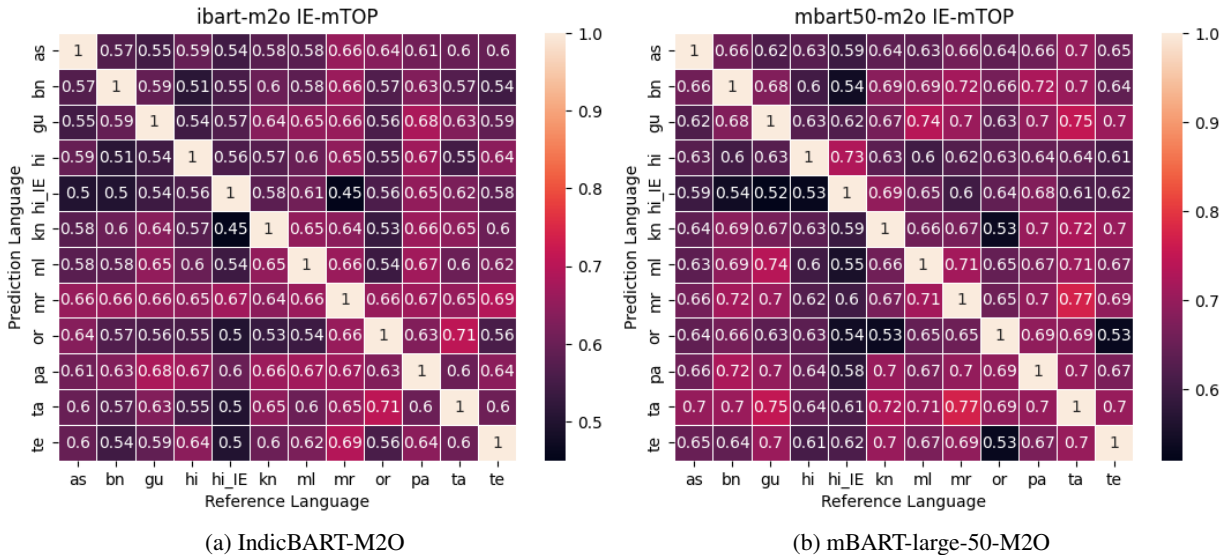
Figure 7: Language wise f1 score of predictions of 2 languages for **IE-mTOP dataset** for **Train All** setting

(a) IndicBART-M2O

(b) mBART-large-50-M2O
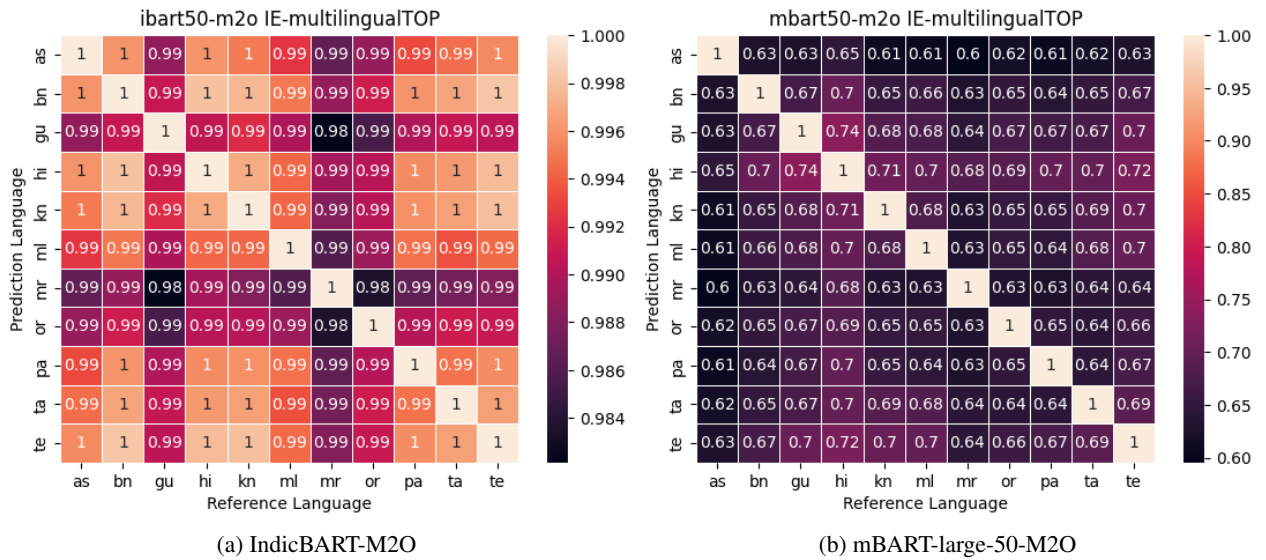
Figure 8: Language wise f1 score of predictions of 2 languages for **IE-multilingualTOP Dataset** for **Train All** settings

(a) IndicBART-M2O

(b) mBART-large-50-M2O
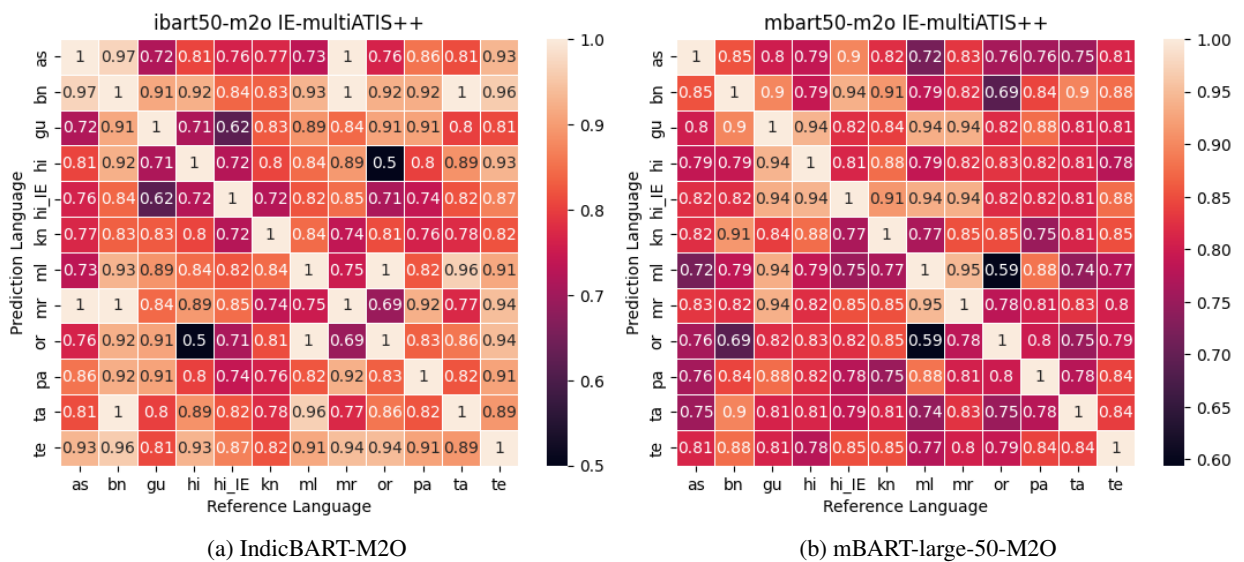
Figure 9: Language wise f1 score of predictions of 2 languages for **IE-multiATIS++ Dataset** for **Train All** settings

Table 12:

| Dataset | Model | Train All | | | | | | | | | | | | | ModAvg |
		as	bn	gu	hi	kn	ml	mr	or	pa	ta	te	hi$_O$	hi$_{IE}$	
IE-mTOP	IndicBART	31	32	29	**42**	29	32	42	20	28	30	31	64	49	35
	IndicBART-M2O	42	40	46	48	46	52	47	47	48	48	**50**	68	53	49
	mBART-large-50	37	33	40	**48**	39	42	38	43	36	42	35	62	51	42
	mBART-large-50-M2O	48	45	50	50	50	53	49	50	47	**53**	51	67	54	**51**
	mT5-base	43	47	51	**52**	50	51	50	50	47	51	**52**	59	55	**51**
	Language Average	40	39	43	**46**	43	**46**	45	42	41	45	44	61	50	45
IE-multilingualTOP	IndicBART	35	38	42	**56**	39	37	47	22	38	36	43	–	–	39
	IndicBART-M2O	45	47	47	55	46	46	52	45	53	50	**57**	–	–	49
	mBART-large-50	37	41	43	**48**	41	41	36	40	40	41	47	–	–	41
	mBART-large-50-M2O	49	53	55	**60**	53	53	48	52	52	53	59	–	–	**53**
	mT5-base	43	49	52	**56**	52	50	47	45	49	48	54	–	–	50
	Language Average	28	31	33	**37**	32	31	31	27	30	30	34	–	–	31
IE-multiATIS++	IndicBART	37	20	23	**41**	32	23	37	13	39	38	19	34	16	29
	IndicBART-M2O	43	45	40	**59**	53	44	58	34	45	46	40	55	37	46
	mBART-large-50	60	85	73	**76**	75	76	60	59	67	66	72	36	18	63
	mBART-large-50-M2O	67	80	71	**73**	71	71	66	58	72	66	68	49	31	**65**
	mT5-base	45	70	58	**61**	60	**61**	45	44	52	51	57	34	16	50
	Language Average	50	60	53	**62**	58	55	53	42	55	53	51	42	24	51

Table 12: $Exact\_Match*100$ scores for the all the dataset for **Train All** settings. **ModAvg** is shorthand for Model Average. The bold numbers in the table indicate the row-wise maximum, i.e. the model's best language performance in the given context. The numbers in bold in the **ModAvg** column indicate the model with the best performance for the train-test strategy specified in the table's heading. Similarly, the numbers in bold in the **Language Average** row indicate the language with the best performance for that train-test strategy.

Table 13:

| Dataset | Model | Translate Test | | | | | | | | | | | Model Average |
		as	bn	gu	hi	kn	ml	mr	or	pa	ta	te	
IE-mTOP	IndicBART	29	40	38	**47**	38	40	41	39	37	43	34	**39**
	IndicBART-M2O	28	37	36	**46**	37	39	39	39	35	43	35	38
	BART-base	18	28	28	**35**	27	29	29	29	28	33	24	28
	BART-large	23	35	33	40	33	36	36	36	33	**41**	30	34
	mBART-large-50	13	14	15	**17**	15	13	18	15	16	16	14	15
	mBART-large-50-M2O	29	38	39	44	38	39	39	36	38	**46**	36	38
	mT5-base	26	36	33	**42**	33	36	36	33	32	38	31	34
	T5-base	21	33	31	**40**	30	31	33	35	31	37	32	32
	T5-large	20	33	29	**38**	29	31	32	35	30	35	29	31
	Language Average	23	33	31	**39**	31	33	34	33	31	37	29	32
IE-multilingualTOP	IndicBART	16	24	26	**28**	21	24	20	21	22	20	26	23
	IndicBART-M2O	13	20	23	**24**	20	21	18	19	19	19	22	20
	BART-base	12	13	13	**14**	11	12	11	11	12	11	13	12
	BART-large	10	15	16	**17**	13	14	12	14	13	14	16	14
	mBART-large-50	22	30	31	**35**	30	31	29	26	29	28	32	29
	mBART-large-50-M2O	26	38	40	**43**	36	38	36	33	35	34	40	**36**
	mT5-base	15	20	21	**23**	19	20	18	18	20	19	21	19
	T5-base	12	13	12	**15**	10	12	13	9	11	14	14	12
	T5-large	22	23	22	25	26	26	25	26	26	26	**27**	25
	Language Average	16	22	23	**25**	21	22	20	20	21	21	23	21
IE-multiATIS++	IndicBART	30	49	34	41	41	**51**	34	33	43	33	44	39
	IndicBART-M2O	32	51	39	44	40	**59**	37	42	43	35	46	43
	BART-base	31	**32**	**32**	30	31	30	30	30	30	30	30	31
	BART-large	31	**32**	**32**	30	31	30	30	30	31	31	31	31
	mBART-large-50	41	56	54	62	61	**66**	54	50	60	47	56	55
	mBART-large-50-M2O	40	60	66	**69**	62	66	57	58	60	47	59	59
	mT5-base	24	29	28	**35**	28	24	26	27	22	25	24	27
	T5-base	34	53	44	48	**55**	61	34	42	42	43	56	47
	T5-large	38	**60**	51	57	56	68	34	42	50	44	57	51
	Language Average	33	47	42	46	45	**51**	37	39	42	37	45	42

Table 13: $Exact\_Match*100$ scores for the all the dataset for **Translate Test** settings. The bold numbers in the table indicate the row-wise maximum, i.e. the model's best language performance in the given context. The numbers in bold in the **Model Average** column indicate the model with the best performance for the train-test strategy specified in the table's heading. Similarly, the numbers in bold in the **Language Average** row indicate the language with the best performance for that train-test strategy.

Dataset	Model	**Indic Train**											**Model Average**
		as	bn	gu	hi	kn	ml	mr	or	pa	ta	te	
IE-mTOP	IndicBART	24	26	29	33	28	24	**44**	12	25	23	23	26
	IndicBART-M2O	43	48	49	56	48	**53**	52	47	6	49	50	46
	mBART-large-50	34	44	43	**55**	40	44	45	27	36	0	50	38
	mBART-large-50-M2O	48	53	55	**62**	50	58	53	48	46	54	57	**53**
	mT5-base	22	29	21	**45**	42	46	29	24	28	25	24	30
	Language Average	34	40	39	**50**	42	45	45	32	28	30	41	39
IE-multilingualTOP	IndicBART	30	24	20	**61**	43	37	51	25	31	37	32	36
	IndicBART-M2O	45	54	56	**60**	56	15	51	20	54	53	59	48
	mBART-large-50	46	51	50	**57**	51	50	49	46	31	50	54	49
	mBART-large-50-M2O	49	56	59	**62**	56	55	53	53	46	54	60	**55**
	mT5-base	40	40	51	**61**	51	43	43	43	40	47	53	47
	Language Average	42	45	47	**60**	51	40	49	37	40	48	52	47
IE-multiATIS++	IndicBART	46	45	43	**54**	32	34	46	23	20	30	32	37
	IndicBART-M2O	56	56	54	**74**	44	55	68	47	40	50	52	54
	mBART-large-50	56	67	**76**	66	54	47	59	62	51	53	46	58
	mBART-large-50-M2O	66	**91**	81	81	60	65	72	78	69	65	60	**72**
	mT5-base	46	53	47	**56**	45	47	48	42	43	44	45	47
	Language Average	54	62	60	**66**	47	50	59	50	45	48	47	53

Table 14: $Exact\_Match*100$ scores for the all the dataset for **Indic Train** settings. The bold numbers in the table indicate the row-wise maximum, i.e. the model's best language performance in the given context. The numbers in bold in the **Model Average** column indicate the model with the best performance for the train-test strategy specified in the table's heading. Similarly, the numbers in bold in the **Language Average** row indicate the language with the best performance for that train-test strategy.

Dataset	Model	**English+Indic Train**											**Model Average**
		as	bn	gu	hi	kn	ml	mr	or	pa	ta	te	
IE-mTOP	IndicBART	27	29	36	**53**	34	28	49	17	34	37	36	35
	IndicBART-M2O	45	46	50	**54**	51	53	50	53	53	51	**54**	51
	mBART-large-50	43	46	50	50	47	45	50	0	37	**54**	50	43
	mBART-large-50-M2O	51	55	53	**61**	56	62	54	51	53	60	**61**	**56**
	mT5-base	23	30	37	**56**	41	27	38	16	27	38	39	34
	Langauge Average	38	41	45	**55**	46	43	48	27	41	48	48	44
IE-multilingualTOP	IndicBART	37	30	47	52	42	35	**53**	25	42	33	44	40
	IndicBART-M2O	48	52	56	59	56	54	50	53	16	53	**60**	51
	mBART-large-50	45	49	42	54	47	52	44	25	54	**56**	56	48
	mBART-large-50-M2O	51	56	59	**63**	57	56	53	53	56	57	61	**57**
	mT5-base	39	48	51	46	49	45	42	43	47	47	**52**	46
	Language Average	44	47	51	**55**	50	48	48	40	43	49	55	48
IE-multiATIS++	IndicBART	28	32	32	**63**	31	25	57	10	29	33	28	33
	IndicBART-M2O	74	**78**	76	**78**	72	80	40	54	64	53	68	**67**
	mBART-large-50	31	40	71	**83**	71	69	57	21	23	40	58	51
	mBART-large-50-M2O	64	84	73	78	70	**88**	71	46	66	70	76	71
	mT5-base	18	25	22	**35**	26	29	26	28	28	25	27	26
	Language Average	43	52	55	**67**	54	58	50	32	42	44	51	50

Table 15: $Exact\_Match*100$ scores for the all the dataset for **English+Indic Train** settings. The bold numbers in the table indicate the row-wise maximum, i.e. the model's best language performance in the given context. The numbers in bold in the **Model Average** column indicate the model with the best performance for the train-test strategy specified in the table's heading. Similarly, the numbers in bold in the **Language Average** row indicate the language with the best performance for that train-test strategy.