Bilingual Tabular Inference: A Case Study on Indic Languages

Chaitanya Agarwal[1*], **Vivek Gupta**[2*†], **Anoop Kunchukuttan**[3,4], **Manish Shrivastava**[1]
[1]LTRC, IIIT Hyderabad; [2]University of Utah; [3]AI4Bharat; [4]Microsoft
chaitanya.agarwal@research.iiit.ac.in; vgupta@cs.utah.edu;
ankunchu@microsoft.com; m.shrivastava@iiit.ac.in

Abstract

Existing research on Tabular Natural Language Inference (TNLI) exclusively examines the task in a monolingual setting where the tabular premise and hypothesis are in the same language. However, due to the uneven distribution of text resources on the web across languages, it is common to have the tabular premise in a high resource language and the hypothesis in a low resource language. As a result, we present the challenging task of bilingual Tabular Natural Language Inference (bTNLI), in which the tabular premise and a hypothesis over it are in two separate languages. We construct EI-INFOTABS: an English-Indic bTNLI dataset by translating the textual hypotheses of the English TNLI dataset INFOTABS into eleven major Indian languages. We thoroughly investigate how pre-trained multilingual models learn and perform on EI-INFOTABS. Our study shows that the performance on bTNLI can be close to its monolingual counterpart, with translate-train, translate-test and unified-train being strongly competitive baselines.

1 Introduction

Tabular Natural Language Inference (TNLI) is the task of classifying whether a textual hypothesis is an entailment, contradiction or a neutral extension of the given tabular premise. The task requires a broad range of reasoning abilities, including but not limited to the ability to make lexical, spatio-temporal, and semantic deductions. Recently published datasets, TabFact (Chen et al., 2020b) and INFOTABS (Gupta et al., 2020), have enabled the examination of the TNLI task. Moreover, sophisticated models based on deep contextual embeddings like BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), etc. trained under supervision on heuristic adaptations of these datasets perform adequately.

Joe Strummer	
Birth Name	John Graham Mellor
Born	1952-08-21 Ankara, Turkey
Died	2002-12-22 Broomfield, Somerset, England
Genres	Punk Rock, Post Punk
Occupation(s)	Musician, Songwriter, Radio Host, Actor
Instruments	Vocals, Guitar, Piano
Years Active	1970-2002
Labels	CBS, Sony, Hellcat, Mercury
Associated Acts	The 101ers, The Clash

H1: John Graham Mellor plays less instruments than the number of labels he has worked for.

H2: Joe Strummer changed his surname after he became a guitar player.

H3: Joe Strummer was active in the sports industry for over three decades.

$\mathrm{H1}_{hi-trl}$: jon grāham melar un lebaloan kī sankhyā kī tulanā mean kam vādya bajāte haian jinake lie unhoanne kām kiyā hai

$\mathrm{H2}_{hi-trl}$: jo sṭramar ne ek giṭār vādak banane ke bād apanā upanām badal liyā

$\mathrm{H3}_{hi-trl}$: jo sṭramar tīn dashakoan se khel udyog mean sakriya the

Figure 1: Tabular premise followed by human written hypotheses (H1, H2, H3). H1 is entailed entirely from the premise, H2 is neither entailed nor contradictory, and H3 is contradictory. $\mathrm{H1}_{hi-trl}$, $\mathrm{H2}_{hi-trl}$, and $\mathrm{H3}_{hi-trl}$ are the transliterations of Hindi translations of the former, released as a part of our EI-INFOTABS dataset.

Typically, and to the authors' best knowledge, fact verification tasks, specifically TNLI, have been examined only in a monolingual setting wherein, the tabular premise and the textual hypothesis are in the same language. However, many semi-structured/tabular data sources exist only in English but require verification of hypotheses over those data sources in other languages, as discussed in §2. Therefore, we examine a modified tabular NLI task by introducing bilinguality within the premise

*Equal Contribution †Corresponding Author

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4018 - 4037
July 10-15, 2022 ©2022 Association for Computational Linguistics

and hypothesis pair. To understand this modified task, consider the example in Figure 1. The table presented in the figure has been extracted from the English Wikipedia article on *Joe Strummer*[1], a well known *musician*. Following which, are transliterated hypotheses in Hindi (hi) (and their English (en) translation) which are related to the information presented in the given table. We show transliterated hypotheses only for the ease of comprehension. We use native scripts for each language in the EI-INFOTABS dataset. For bilingual tabular NLI, a reasoning model should be able to predict the inference label entail for H1$_{hi}$, neutral for H2$_{hi}$ and contradict for H3$_{hi}$ given the English table as the primary context. In summary, our contributions are as follows:

• We introduce the task of bilingual tabular NLI (bTNLI) wherein the tabular premise is in a high resource language, while the textual hypothesis is in a low resource language. This is a practical, real world setting for fact verification on semi-structured tabular data which is further illustrated in §2.

• We create EI-INFOTABS, a dataset consisting of machine translated hypotheses in 11 Indian languages, while retaining the English tabular premises from the INFOTABS dataset. Through extensive studies shown in §3, we confirm that EI-INFOTABS is of good quality, and preserves properties important to study the bTNLI task.

• We explore several multilingual models for the bTNLI task, and establish strong baselines and share findings about their performance across multilingual models, languages, train-eval techniques, tabular reasoning categories, adversarial test splits, and both datasets (INFOTABS and EI-INFOTABS).

Overall, EI-INFOTABS dataset and our proposed train-eval strategies enable thorough examination of the challenging task of bTNLI. Furthermore, the former aslo serve as a quality benchmark for evaluating the robustness of multilingual models. The dataset and the associated scripts, are available at `https://enindicinfotabs.github.io`.

2 Motivation

Why Tabular NLI? Tabular data is termed as semi-structured as it is neither truly unstructured data like raw text, nor is it entirely structured like a database. Although semi-structured data is based on a structured scaffold, the content can be free-form text with variable length and type. Moreover, unlike a database, there is no homogeneity across various data points in a shared context. Such structural ambiguity imposes a significant cognitive load while reasoning about it. However, such data is ubiquitous in the real world (e.g. web pages, fact sheets, information tables) and we frequently make inferences from it.

Chen et al. (2020b) argue that reasoning about semi-structured data is broadly two-fold in nature. It consists of (a.) Linguistic Reasoning: a semantic deconstruction of the semi-structured data (b.) , and Symbolic Reasoning: a symbolic execution on the tabular structure.For instance, H_2 in Figure 1 requires linguistic reasoning over the phrase "*became a guitar player*" from the "*Occupation*", and the "*Instruments*" rows of the concerned table. H_1 requires symbolic reasoning in the form of conditional and arithmetic operations on the "*Labels*" and "*Instruments*" rows. Whereas, H_3 requires a combination of the two types of reasoning. Such interwoven reasoning criteria makes it challenging to model Tabular NLI task.

Why Indic Languages? Indian society is largely multilingual and consists of 122 major and 1599 other languages and dialects spanning 6 language families with over 1.3 billion native speakers[2]. Out of these, 30 languages have more than 1 million native speakers each and over 1 billion speakers cumulatively[3]. Moreover, India has the second largest online presence with over 749 million internet users and is expected to grow to over 1.5 billion users by 2040[4]. So, development of competent reasoning models for the Indic context is essential.

However, due to unfair linguistic bias on the web (Miquel-Ribé and Laniado, 2020; Joshi et al., 2020), there is a disproportionate distribution of text resources for Indian languages. Indian languages have a limited number of internet resources. Thus, they are often known as low web resource languages (LRL) (Khemchandani et al., 2021). For instance, Wikipedia entries in Hindi are just 2% of those in English, and Wikipedia entries in Assamese and Oriya are 7 times lesser than those in Hindi. This implies that a significant fraction of

[1] Joe Strummer Wikipedia

[2] Wikipedia Indian Languages [3] 2011 Indian Census
[4] www.statista.com

articles and sometime even complete categories are discussed only in the English language Wikipedia (Bao et al., 2012).

Although, efforts have been made to bridge this gap (Adar et al., 2009; Kumaran et al., 2010), there still exist several limitations: (a.) **table extraction** for an article across languages is a challenge due to absence of Wikipedia page links, their infobox tables or important keys of tables, (b.) even if tabular data exists, infobox tables in Indian languages are not updated as regularly as their English equivalents (Minhas et al., 2022) which leaves us with **outdated and untrustworthy tabular data** for inference, (c.) and lastly, **table translation** while maintaining the intent, context, and the same quality of the source English language is difficult. Often, accurate translation requires the distinction of a language specific domain expert. Due to above reasons, tabular data is mostly absent from Indic Wikipedia articles.

Thus, fact verification in a bilingual setting wherein, the premise is in English and the claim/hypothesis is in an Indic language, is of great significance. Moreover, recent advances in multilingual language models (Khanuja et al., 2021a; Kunchukuttan, 2020), datasets (Roark et al., 2020; Ramesh et al., 2022), and translation systems (Ramesh et al., 2022) for Indian languages have enabled quality examination of several Indic NLU tasks which serves as additional motivation to evaluate the task of bTNLI for Indic languages before other low resource languages.

3 EI-InfoTabS Dataset

EI-InfoTabS is an English-Indic bTNLI extension of InfoTabS (Gupta et al., 2020), an English TNLI dataset. InfoTabS consists of 23,738 pairs of tabular premises and textual hypotheses. The hypotheses are human written short assertions with an accompanying NLI label, and the tabular premises are based on 2,540 Wikipedia infoboxes from 12 diverse categories. Moreover, it consists of additional adversarial test sets apart from α_1 which is the standard test set and is lexically and topically similar to the train set - α_2 is the lexically adversarial test set which maintains topical similarity and α_3 is the topically adversarial test set. The dev and test sets (α_1, α_2, α_3) cumulatively consist of 7200 table-hypothesis pairs equally splits on all four sets.

EI-InfoTabS extends it by providing machine translated hypotheses in 11 major Indic languages namely Assamese (*as*), Bengali (*bn*), Gujarati (*gu*), Hindi (*hi*), Kannada (*kn*), Malayalam (*ml*), Marathi (*mr*), Odia (*or*), Punjabi (*pa*), Tamil (*ta*), and Telugu (*te*) for each tabular premise. In this section, we discuss the EI-InfoTabS construction and verification.

3.1 EI-InfoTabS Construction

To construct EI-InfoTabS, we machine translated the English hypotheses provided in InfoTabS to 11 major Indian languages as described earlier. We use IndicTrans (Ramesh et al., 2022), an open-sourced state-of-the-art Indic NMT model. IndicTrans is trained on the Samanantar dataset (Ramesh et al., 2022), which is the largest publicly available parallel corpus for Indic languages. Moreover, it outperforms (a) commercial NMT systems like Google-Translate [5] and Bing Microsoft Translator [6], and (b) open-source multilingual models like OPUS-MT (Tiedemann and Thottingal, 2020), mBART50 (Liu et al., 2020) and mT5 (Xue et al., 2021).

3.2 EI-InfoTabS Verification.

Given the absence of Indic reference data, it becomes challenging to measure the quality of the translations, and subsequently, of EI-InfoTabS. In this section, we describe our robust quality estimation approach to validate EI-InfoTabS.

Automatic Evaluation. We use BERTScore (Zhang* et al., 2020), an automatic scoring metric for sentence similarity, between the source and back-translated English sentences. We use IndicTrans to generate Indic to English back-translated data.

BERTScore is known to correlate better with human judgment at the sentence level (Zhang* et al., 2020) compared to conventionally used MT evaluation metrics like BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004). BERTScore calculates word level semantic similarity whereas the conventional MT metrics focus on word overlap. The results are presented in Table 1. We notice high semantic similarity scores for all the languages. However, when we analyse the examples with low scores, we note that the scores are almost always low due to the error added during the back-translation phase. The back-translation introduces

[5] https://translate.google.co.in/

[6] https://www.bing.com/translator

errors due to incorrect transliteration of *Named Entities*. Consider the following example:

• Femme aux Bras Croisés is open for public viewing.

• Back-translated: The Ox Brass Crossox is open to the public

• Hindi Translation(Transliterated): fem auksa brās kroisaiksa janatā ke lie khulā hai

The Hindi translation of the original sentence is perfect, however, the named entity "*Femme aux Bras Croisés*" when back-translated becomes "*Ox Brass Crossox* "and yields a low BERTScore of 0.86. This is broadly identified as qualitative feedback for most of the sentences with low scores across all the languages. Around 20% of the examples yield a BERTScore of 1.0 and are deemed perfect translations when reviewed by native speakers.

Human Evaluation. Broadly, we follow the guidelines recommended in (Agirre et al., 2016) to conduct human evaluation. We (a.) diversely sample source-translation pairs in each language, (b.) prepare a common Direct Assessment (Graham et al., 2013) scoring strategy, and (c.) get the sampled data evaluated on the basis of that strategy.

Diverse Sampling. We sample 50 diverse hypotheses from the dev split of EI-INFOTABS for each Indic language. Using the k-DPP algorithm (Kulesza and Taskar, 2011) over the mBERT sentence representations, we're able to achieve syntactically and semantically diverse samples spanning the different table categories.

Direct Assessment. We adopt the human evaluation strategy for low resource machine translation laid out in (Guzmán et al., 2019). We ask native Indic language speakers proficient in English to score a source-translation pair from 0-100. The score highlights the perceived translation quality of the source-translation pair. For each language, we get the samples annotated by two different annotators. In Table 1, we report the average scores for each language along with the Pearson correlation coefficient (r) as a measure for inter-rater reliability. For more details on human evaluation strategy refer to Appendix §C.

Discussion. We report our evaluation results in Table 1. Automatic evaluation and our corresponding analysis on it shows that EI-INFOTABS consists of fluent, semantically accurate translations across all Indic languages. Moreover, we note competitive Direct Assessment scores for each language, and a positive r value which indicates that the native speakers agree on the good quality of EI-INFOTABS.

Language	DA	BS^{IT}	BS^{GT}	r
Bengali ('bn')	0.87	0.95	0.99	0.64
Marathi ('mr')	0.81	0.94	0.98	0.68
Gujarati ('gu')	0.89	0.95	0.98	0.38
Oriya ('or')	0.94	0.94	0.98	0.35
Hindi ('hi')	0.89	0.96	0.99	0.40
Punjabi ('pa')	0.86	0.95	0.98	0.34
Kannada ('kn')	0.87	0.95	0.98	0.70
Tamil ('ta')	0.85	0.94	0.98	0.59
Malayalam ('ml')	0.85	0.94	0.98	0.50
Telugu ('te')	0.84	0.94	0.98	0.39
Assamese ('as')	0.83	0.94	-	0.65

Table 1: Here, we compare the Average Direct Assessment (DA) scores provided by native speakers with Average BERTScore F1 scores for IndicTrans En-Indic-En back-translated data (BS^{IT}), and Average BERTScore F1 scores for Google-Translate En-Indic-En back-translated data (BS^{GT}). Additionaly, we also present the Pearson correlation coeffecient as a measure of inter-rater reliability. Higher score implies better quality for each of the metric.

4 Experimental Pipeline

We design our experimental pipeline along the lines of the research question: *How well do existing pre-trained multlilingual language models perform on the bTNLI task?* In this section, we propose various modeling strategies and examine how they might address the challenges and nuances of the proposed inference bTNLI task.

4.1 Table Representations

It is necessary to linearize semi-structured tabular data into a textual premise in order to reduce the task of Tabular Inferencing to a standard NLI task for which existing state-of-the-art language models can be adapted directly. We use and compare the previously proposed linearization methods (a.) Better Paragraph Representations (BPR) (Neeraja et al., 2021), (b.) and Premise as Structure - TabFact (Chen et al., 2020b; Gupta et al., 2020) (cf. Appendix §A). Henceforth, by premise, we refer to the linearized representation of the tabular premise i.e. the infobox table.

Strategy	Model	bn	hi	gu	pa	mr	te	ta	ml	kn	as	or	ModAvg
Translate-Train	mBERT	62	**64**	61	62	61	60	61	59	61	60	35	59
	IndicBERT	**54**	**54**	53	48	51	51	52	47	**54**	34	53	50
	MuRIL	**67***	**67***	**67***	**66**	**65**	**65**	**66**	**66**	**66**	**65**	**64**	66
	LnAvg	61	62	60	59	59	57	60	56	60	53	51	57
Translate-Test	mBERT	53	**55**	50	54	51	50	53	51	52	45	34	50
	IndicBERT	37	35	35	34	36	36	34	38	**39**	**39**	38	36
	MuRIL	**63**	**65***	**62**	**62**	**62**	**60**	**61**	**61**	**62**	**60**	**59**	62
	LnAvg	51	52	49	50	50	49	49	50	51	48	44	49
Bilingual-Train	mBERT	63	**66**	62	64	62	63	63	62	64	62	36	61
	IndicBERT	53	53	52	53	52	50	52	**54**	53	**54**	53	53
	MuRIL	**68***	**67**	**66**	**67**	**65**	**67**	**66**	**66**	**66**	**65**	**65**	66
	LnAvg	61	62	60	61	60	60	60	61	61	60	51	60
Multilingual-Train	mBERT	63	**64**	62	63	62	62	61	62	63	62	36	60
	IndicBERT	53	**54**	53	52	52	50	51	50	50	53	51	52
	MuRIL	**67**	**68***	**67**	**67**	**66**	**66**	**67**	**67**	**67**	**66**	**66**	67
	LnAvg	61	62	61	61	60	59	60	60	60	60	51	60
EnTranslate-Test	mBERT	65	**67***	63	**66**	**63**	**62**	**64**	62	**64**	**62**	62	64
	IndicBERT	56	**57**	56	**57**	55	56	**57**	**57**	**57**	**57**	56	56
	MuRIL	**65**	**67***	**65**	65	**63**	**62**	**64**	**63**	**64**	61	**62**	64
	LnAvg	63	64	61	63	60	60	62	61	62	60	60	62
Translate-Train-X	mBERT	**55**	**55**	53	54	53	53	54	53	53	50	36	51
	IndicBERT	**41**	**41**	39	36	39	40	40	40	40	34	40	39
	MuRIL	**64**	**65***	**64**	**63**	**64**	**64**	**63**	**63**	**64**	**63**	**62**	64
	LnAvg	53	53	52	51	52	52	52	52	52	49	46	51
Bilingual-Train-X	mBERT	**56**	**56**	55	**56**	**56**	**56**	55	55	55	55	41	54
	IndicBERT	**42**	**42**	41	40	41	41	41	41	**42**	**42**	41	41
	MuRIL	**65***	**64**	**65***	**65***	**64**	**65***	**65***	**65***	**65***	**65***	**65***	65
	LnAvg	54	54	54	53	54	54	54	54	54	54	49	53

Table 2: Performance in terms of accuracy when evaluated on the α_1 test set. Higher value implies better performance. Here, LnAvg represents the average accuracy for a language across all models, while ModAvg represents the average accuracy of a model across all languages. A value in **Purple** represents the best accuracy for that model across all languages. An underlined value in **Blue** represents the best accuracy for that language across all models. A value in **Green** with an asterisk(*) represents the cases where language-wise and model-wise values coincide. As we fine-tune on a specific Indic language in the train-eval strategies Translate-Train-X and Bilingual-Train-X, we report the training average of accuracy on the remaining 10 Indic languages for them. We do not include the results of XLM-RoBERTa as the model fails to converge on these experiments on multiple runs with a distinct set of hyper-parameters as explained in Appendix §D. The results for the α_2 and α_3 sets are provided in Appendix §E.

4.2 Multilingual Models

Owing to the multilingual setting of this task, we utilise pre-trained multilingual models to encode the linearized English tabular premise along with the Indic hypothesis into contextual representations for classification. We consider two kinds of pre-trained multilingual models (a.) **Indic Specific** which includes IndicBERT and MuRIL due to their indic specific pre-training, and (b.) **Generic** which includes mBERT and XLM-Roberta due to their pre training in more than hundred languages. For more details refer to Appendix §B.

4.3 Training and Evaluation Strategies

In order to examine the inter-woven relationships among the 11 languages, and the corresponding impact on multilingual models' performance, we design a set of train-eval strategies for this task.

Translate-Train: We fine-tune and evaluate the models on EN-IN$_i$ premise-hypothesis pairs where IN$_i$ is one of the 11 Indic languages. This baseline evaluates the performance of the multilingual models on EI-INFOTABS when fine-tuned on Indic hypotheses. We also evaluated these models across all languages i.e. cross lingual zero-shot setting **Translate-Train-X**.

Translate-Test: We fine-tune the multilingual models on EN-EN premise-hypothesis pairs from the INFOTABS dataset and evaluate on EN-IN$_i$ premise-hypothesis pairs. This baseline evaluates the Zero-shot Cross-Lingual Transfer ability of the reasoning models from INFOTABS to EI-INFOTABS.

Bilingual-Train: We fine-tune the multilingual models on both EN-EN and EN-IN$_i$ premise-hypothesis pairs, and evaluate on EN-IN$_i$ premise-hypothesis pairs. This baseline evaluates whether addition of English hypotheses while fine-tuning aids the performance of the multilingual models prepared in Translate-Train. We also evaluated these models across all languages i.e. cross lingual zero-shot setting **Bilingual-Train-X**.

Multilingual-Train: We fine-tune the multilingual models on all available training data across all Indic languages and the English language. We evaluate the models on EN-IN$_i$ premise-hypothesis pairs on each 11 Indic languages. This baseline assesses if fine-tuning on several languages to produce a unified multilingual model improves performance.

EnTranslate-Test: We fine-tune the multilingual models on EN-EN premise-hypothesis pairs from INFOTABS and evaluate on EN-ENIN$_i$ premise-hypothesis pairs where ENIN$_i$ represents IN$_i$ to EN back-translated hypotheses. This approach evaluate the translate then test baseline on the EI-INFOTABS dataset.

5 Results and Analysis

In this section, we discuss and analyse the results obtained on conducting the experiments as per the various strategies laid out in §4. We present the results in Table 2 for each experiment on the α_1 test set using the BPR linearization algorithm. The values represent classification accuracy. We analyze the findings thoroughly across multilingual models, languages, train-eval techniques, tabular reasoning categories, adversarial test splits, and both datasets (INFOTABS and EI-INFOTABS).

5.1 Across Multilingual Models

We observe that MuRIL performs best across all languages and experiments except EnTranslate-Test, beating IndicBERT and mBERT. MuRIL's superior performance can be justified on the grounds of (a) the large size of the hidden layers, (b) Indic specific pre-training data, and (c) Indic specific pre-training objectives (Khanuja et al., 2021a). MuRIL's architecture consists of 237M parameters, compared to mBERT's 167M and IndicBERT's 33M, which makes it extremely competitive on any Indic NLU task. IndicBERT's relatively small size explains why it performs the worst, even though it is pre-trained on Indic specific data. mBERT comes in a close second to MuRIL, failing to perform adequately only on Odia (*or*). mBERT isn't pre-trained on Assamese (*as*) or Odia which justifies its extremely low performance on Odia. **However, we note competitive results on the Assamese language. This could be attributed to the fact that Assamese is closely related to Bengali (*bn*) linguistically. They both share the Bengali-Assamese script and are mutually intelligible** (Khemchandani et al., 2021).

#	dev	α_1	α_2	α_3
0	15.56%	16.33%	27.61%	25.67%
1-3	11.17%	10.83%	11.39%	14.22%
4-6	7.16%	7.5%	6.72%	9.61%
7-9	9.55%	10.67%	10.22%	12.67%
10-11	**56.56%**	**54.67%**	**44.06%**	**37.83%**

Table 3: Percentage of examples predicted correctly by our best performing model for the given number of Indic languages. For instance, 7.16% of examples in the dev set are predicted correctly for at least 4 and at max 6 Indic languages.

mBERT's performance gets boosted in EnTranslate-Test as mBERT is pre-trained on a significant amount of English data which makes it extremely competitive in modeling English NLU tasks. MuRIL performs similarly even though it is trained on lesser amount of English data. This could be due to Indic artifacts like sentence structure and inadequately transliterated named entities being present in the back-translated sentences which MuRIL has been trained to handle better than mBERT.

5.2 Across Languages

We observe that the models perform best on Hindi (*hi*) and Bengali. This is expected as they are high resource languages in the Indic context. **Additionally, as explained in §5.1, we note that pre-training or fine-tuning on Bengali aids the performance on Assamese due to their high degree of relatedness.** Table 3 shows the measure of agreement across the languages. We note that almost all languages agree on 55% of the predictions on the dev set and the α_1 test set. This reduces to 38% on the α_3 test set. **This indicates that for a majority of examples from the non-adversarial test sets, MuRIL performs uniformly across languages. However, its performance across languages starts varying more on the adversarial test sets (α_2 and α_3).**

5.3 Train-Eval Strategies

Translate-Train's results show that the multilingual models converge and perform adequately when fine-tuned on EI-INFOTABS. Moreover, when fine-tuned along with English data - as described

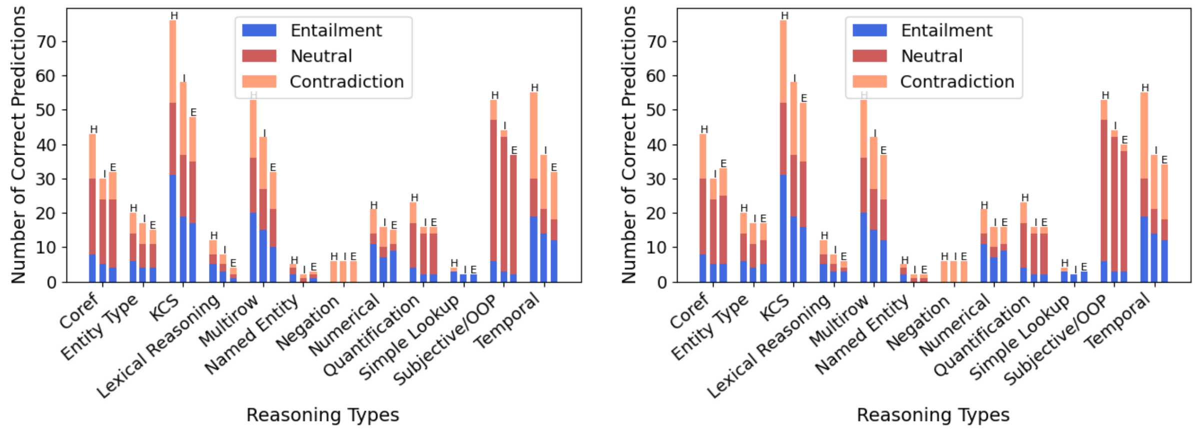

Figure 2: Here, we compare human benchmarks (H), predictions of the best performing model trained on INFOTABS from (Neeraja et al., 2021) (I) and predictions of our best performing model (E), MuRIL (Multilingual-Train), on the examples annotated with reasoning category in the dev split for Oriya (left) and Hindi (right).

in Bilingual-Train - mBERT and IndicBERT perform marginally better while MuRIL doesn't report a change in performance. **MuRIL, when fine-tuned on all languages as described in Multilingual-Train, performs best on** EI-INFOTABS **and forms the benchmark for this task**. mBERT and IndicBERT, however, perform worse on Multilingual-Train when compared to Bilingual-Train. This indicates that these models fail to generalise their reasoning ability across all languages and aren't as multilingually robust as MuRIL. The results on Translate-Test are the lowest across all train/eval strategies which indicates a poor Zero-shot Cross-Lingual Transfer from INFOTABS to EI-INFOTABS. However, the performance of MuRIL on Translate-Test is comparable with its performance on Translate-Train unlike mBERT and IndicBERT. This indicates that MuRIL can generalize well across English and Indic languages which are linguistically distinct.

Translate-Train-X and Bilingual-Train-X evaluate the average Cross-Lingual Transfer performance of the models trained in Translate-Train and Bilingual-Train. **We observe higher performance in Bilingual-Train-X over Translate-Train-X which indicates that addition of English training data aids the Cross Lingual Transfer from one Indic language to another. Moreover, the average performance of MuRIL on Bilingual-Train-X is comparable to that on Translate-Train which suggests that MuRIL robustly generalises across Indic languages.** Both, Bilingual-Train-X and Translate-Train-X perform better than Translate-Test due to high

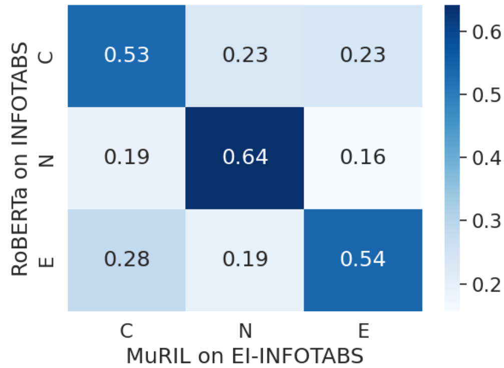

Figure 3: Consistency Matrix which measures the deviation of our best performing model, MuRIL (Multilingual-Train)'s predictions on the α_1 test set for Hindi as compared to that of RoBERTa$_{\text{LARGE}}$ on the α_1 test set of INFOTABS.

language relatedness among Indic languages when compared with English. The results on EnTranslate-Test are extremely promising for both MuRIL and mBERT. Their performance is very close to that of the best performing model, MuRIL, on Translate-Train. This indicates that back-translation doesn't lead to a significant loss in information required for the bTNLI task.

5.4 Tabular Reasoning Categories

We conduct a fine-grained analysis on how our best model, MuRIL (Multilingual-Train), performs on various reasoning categories. We present the results in Figure 2 for Hindi and Odiya. We observe that MuRIL performs similarly on EI-INFOTABS as RoBERTa$_{\text{LARGE}}$ does on INFOTABS for entity type, named entity, negation, numerical, quantification and simple lookup reasoning types. Additionally, MuRIL performs better for the co-reference resolution reasoning type. This is broadly

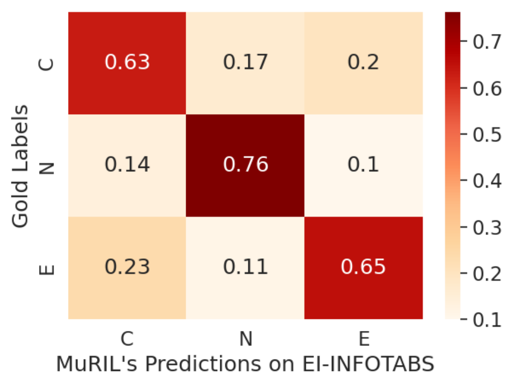

Figure 4: Confusion Matrix for the predictions of our best performing model, MuRIL (Multilingual-Train), on the Hindi α_1 set from EI-INFOTABS.

observed across all the Indic languages. Both RoBERTa$_{\text{LARGE}}$ and MuRIL perform poorly for knowledge and common sense, multi-row, co-reference, and temporal reasoning types.

5.5 Across Adverserial Test Splits

The results for the other evaluation sets α_2 and α_3 are provided in Appendix §E. Across all the experiments, we note that the fine-tuned models perform best on α_1, followed by α_2 and α_3 respectively. **Moreover, we note that on most baselines, the average performance of a fine-tuned model drops by roughly 10% when tested on α_2 or α_3.** This is similar to the observations reported on INFOTABS (Neeraja et al., 2021) and presented in Table 4. Low performance of the multilingual models on the α_2 test set of EI-INFOTABS indicates that (a.) multilingual models learn shallow lexical features to make inferences on EI-INFOTABS just like the monolingual models do on INFOTABS, (b.) and IndicTrans carefully captures the lexical adversity in the α_2 test set of INFOTABS.This commends the ability of IndicTrans to handle lexical nuances. Low performance on α_3 test set of EI-INFOTABS suggests that the multilingual models learn categorical features and perform adversely when evaluated on unseen category.

5.6 EI-INFOTABS v/s INFOTABS

Table 4 reports the human benchmarks and the baselines with the BPR linearization algorithm on each validation set in INFOTABS. We observe that the baselines on EI-INFOTABS are within an absolute margin of 10% when compared to those on INFOTABS. **This suggests that EI-INFOTABS is more challenging than INFOTABS which was expected due to the presence of (a.) bilinguality within the premise-hypothesis pair, and (b.) the low resource nature of Indic languages.**

Figure 3 reports the consistency of predictions of MuRIL on the α_1 test set of Hindi EI-INFOTABS when compared against that of RoBERTa$_{\text{LARGE}}$ on the α_1 test set of INFOTABS. We observe that MuRIL behaves noticeably different than RoBERTa$_{\text{LARGE}}$. MuRIL disagrees with RoBERTa$_{\text{LARGE}}$ on 47% of examples with the Contradiction and Entailment labels. However, for Neutral labels, it only disagrees on around 36% of the examples. Moreover, from our discussion in §5.4, we observe that MuRIL outperforms RoBERTa$_{\text{LARGE}}$ on certain reasoning categories.

Model (Rep)	Dev	α_1	α_2	α_3
BERT$_{\text{B}}$ (BPR)	63.00	63.54	52.57	48.17
RoBERTa$_{\text{B}}$ (TabFact)	68.06	66.7	56.87	55.26
RoBERTa$_{\text{L}}$ (BPR)	76.42	75.29	66.50	64.26
RoBERTa$_{\text{L}}$ (TabFact)	77.61	75.06	69.02	64.61
Human	**79.78**	**84.04**	**83.88**	**79.33**

Table 4: The human benchmarks and several baselines on evaluation set of INFOTABS as reported in Gupta et al. (2020) (TabFact) and Neeraja et al. (2021) (BPR). Here subscript X$_{\text{L}}$ and X$_{\text{B}}$ represent X model L: Large and B: Base versions respectively.

However, the models fine-tuned on EI-INFOTABS broadly mimic the performance of RoBERTa$_{\text{LARGE}}$ on INFOTABS. Figure 4 presents the confusion matrix of MuRIL's predictions on the α_1 test set of Hindi. We observe a similar distribution across all Indic languages. As noted in Gupta et al. (2020), MuRIL also tends to predict Neutral hypotheses with the highest confidence as they mostly contain out of table or subjective information terms. Moreover, both models confuse Entailment with Contradiction inference label and vice-versa. We observe that the model predictions on EI-INFOTABS is similar to RoBERTa$_{\text{LARGE}}$ predictions on INFOTABS.

6 Further Discussion

EI-INFOTABS is the first Tabular NLI dataset in the Indic context which enables preliminary studies in this field. Moreover, it introduces bilinguality for fact verification scenarios which is of huge significance in low resource contexts. It motivates the development of cross-lingual reasoning models, and helps in evaluation of robustness of multilingual models. For instance, our experiments on EI-INFOTABS clearly indicate that MuRIL is a significantly more robust multilingual model when compared to mBERT as it is able to generalize its reasoning ability across all Indic languages.

Although, we explain how machine translation doesn't affect the semantics of the hypotheses, it does come with a few challenges. We identified a few instances wherein the IndicTrans model translates named entities, instead of transliterating them. This is observed only, but not always, when a named entity has an English dictionary word in it. For instance, "*Death Proof*", name of a movie, gets translated and not transliterated in two out of nine hypotheses containing the phrase. This is mostly observed in the *Movies* category. However, this doesn't affect our reasoning models and they perform on par on this category when compared with RoBERTa$_{\text{LARGE}}$'s performance on INFOTABS. This is so because such translations when shallow parsed indicate that the translated entity still acts as the Noun Phrase in the sentence. This helps the translation, though technically imperfect, retain the intended semantic structure.

7 Related Work

Tabular Reasoning. Tabular NLI has been of keen interest recently. Datasets like TabFact (Chen et al., 2020b), INFOTABS (Gupta et al., 2020) were the first resources on TNLI and they enabled a fine-grained examination of the task. Beyond NLI, there has been a thorough examination of various other NLP tasks on semi-structured data. For instance, question answering (Abbas et al., 2016; Chen et al., 2020c; Zayats et al., 2021; Oguz et al., 2020; Chen et al., 2021, and others), semantic parsing and retrieval (Krishnamurthy et al., 2017; Sun et al., 2016; Pasupat and Liang, 2015; Lin et al., 2020, and others), tabular probing (Gupta et al., 2021), generative tasks including table-to-text (Parikh et al., 2020; Nan et al., 2021; Yoran et al., 2021; Chen et al., 2020a,d, and others). Other works have explored creating task-independent representations for Wikipedia infoboxes (Herzig et al., 2020; Yin et al., 2020; Zhang et al., 2020; Iida et al., 2021; Pramanick and Bhattacharya, 2021; Glass et al., 2021, and others), and boosting tabular reasoning by pre-training and external knowledge incorporation (Neeraja et al., 2021; Varun et al., 2022, and others).

Multilingual Models. Multilingual, and specifically Cross-Lingual transfer (Deshpande et al., 2021; Patil et al., 2022, and other), has been widely discussed in the context of low resource languages. Several datasets (Conneau et al., 2018; Yang et al., 2019; Ponti et al., 2020; Artetxe et al., 2020; Nivre et al., 2016; Lewis et al., 2021, and others), benchmarks and leaderboards (Hu et al., 2020; Liang et al., 2020; Ruder et al., 2021; Khanuja et al., 2021b, and others), and evaluation frameworks (Tarunesh et al., 2021; K et al., 2021; Srinivasan et al., 2021) have emerged which focus entirely on evaluation of multilingual NLU. Further, multilingual language models have been developed for (a.) Natural Language Understanding (Devlin et al., 2019; Conneau and Lample, 2019; Conneau et al., 2020; Chi et al., 2021; Chung et al., 2021, and others), (b.) and Natural Language Generation (Xue et al., 2021; Fan et al., 2021, and others).

Indic Resources. Indic NLP, recently, has seen a recent surge in the number of datasets (Ramesh et al., 2022; Roark et al., 2020; Haddow and Kirefu, 2020a; Abadji et al., 2022; Kolluru et al., 2021, and others), multilingual models (Dabre et al., 2021; Kakwani et al., 2020; Khanuja et al., 2021a, and others), toolkits (Arora, 2020; Bhat et al., 2015; Jain et al., 2020, and others), translation systems (Ramesh et al., 2022), and dedicated benchmarks for evaluation (Kakwani et al., 2020; Krishna et al., 2021). This has enabled the Indian NLP research community to construct competent models for a variety of challenging NLP tasks.

8 Conclusion

We motivate and introduce the bilingual tabular NLI for fact verification tasks, and release EI-INFOTABS- a first of its kind tabular NLI dataset for making inferences in 11 Indic languages over English tabular data. Our robust quality estimation experiments show that the machine translated datasets closely preserve the semantics of the source and are fluent. We show that pre-trained multilingual models find this task challenging, however, still perform close to the benchmarks on INFOTABS with Translate-test and Translate-train providing good performance. The analysis also shows the similarity of inference capabilities across languages. The dataset offers immense potential as it opens up avenues in (a) multilingual tabular NLI, (b) bilingual claim verification, (c) and evaluation of multilingual models.

9 Ethical Considerations

In terms of demographic and socioeconomic characteristics, we attempted to establish a balanced, bias-free dataset. The EI-INFOTABS dataset is derived from the INFOTABS dataset, which is devoid of bias. The only possible source of prejudice can be the translation pipeline. Our qualitative analysis indicates that translation quality is reasonably good and there aren't any observable biases like gender in the translation. The dataset is intended and useful for studying language model representations in a cross-lingual and structured data setting. The paper points out that low-resource languages can benefit from reasoning over structured data in other languages. This is a relatively new research topic and further work will help understand limitations as well as uncover new directions. Hence, we recommend the use of this dataset at this point exclusively for scholarly, non-commercial purposes.

Acknowledgement

We thank members of the Utah NLP group for their valuable insights and suggestions at various stages of the project; and reviewers their helpful comments. The authors also thank Manila Devaraj, Arka Singha, Anagh Chattopadhyay, Maitrey Mehta, Souvik Banerjee, Rupambara Padhi, Jayant Duneja, Nithila Prakash, Tanvi Kamble, Anirudh Palutla for helping with annotations for quality estimation. Additionally, we appreciate the inputs provided by Vivek Srikumar and Ellen Riloff. Vivek Gupta acknowledges support from Bloomberg's Data Science Ph.D. Fellowship.

References

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a Cleaner Document-Oriented Multilingual Crawled Corpus. *arXiv e-prints*, page arXiv:2201.06642.

Faheem Abbas, M. K. Malik, M. Rashid, and Rizwan Zafar. 2016. Wikiqa — a question answering system on wikipedia using freebase, dbpedia and infobox. *2016 Sixth International Conference on Innovative Computing Technology (INTECH)*, pages 185–193.

Eytan Adar, Michael Skinner, and Daniel S Weld. 2009. Information arbitrage across multi-lingual wikipedia. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 94–103.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511, San Diego, California. Association for Computational Linguistics.

Gaurav Arora. 2020. iNLTK: Natural language toolkit for indic languages. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 66–71, Online. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Patti Bao, Brent Hecht, Samuel Carton, Mahmood Quaderi, Michael Horn, and Darren Gergle. 2012. Omnipedia: bridging the wikipedia language gap. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1075–1084.

Irshad Ahmad Bhat, Vandan Mujadia, Aniruddha Tammewar, Riyaz Ahmad Bhat, and Manish Shrivastava. 2015. Iiit-h system submission for fire2014 shared task on transliterated search. In *Proceedings of the Forum for Information Retrieval Evaluation*, FIRE '14, pages 48–53, New York, NY, USA. ACM.

Wenhu Chen, Ming-Wei Chang, Eva Schlinger, William Yang Wang, and William W. Cohen. 2021. Open question answering over tables and text. In *International Conference on Learning Representations*.

Wenhu Chen, Jianshu Chen, Yu Su, Zhiyu Chen, and William Yang Wang. 2020a. Logical natural language generation from open-domain tables. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7929–7942, Online. Association for Computational Linguistics.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020b. Tabfact: A large-scale dataset for table-based fact verification. In *International Conference on Learning Representations*.

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Wang. 2020c. Hybridqa: A dataset of multi-hop question answering over tabular and textual data. *Findings of EMNLP 2020*.

Zhiyu Chen, Wenhu Chen, Hanwen Zha, Xiyou Zhou, Yunkai Zhang, Sairam Sundaresan, and William Yang Wang. 2020d. Logic2Text: High-fidelity natural language generation from logical forms. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2096–2111, Online. Association for Computational Linguistics.

Zewen Chi, Shaohan Huang, Li Dong, Shuming Ma, Saksham Singhal, Payal Bajaj, Xia Song, and Furu Wei. 2021. Xlm-e: Cross-lingual language model pre-training via electra. *CoRR*, abs/2106.16138.

Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2021. Rethinking embedding coupling in pre-trained language models. In *International Conference on Learning Representations*.

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. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems*, 32:7059–7069.

Alexis CONNEAU and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh M. Khapra, and Pratyush Kumar. 2021. Indicbart: A pre-trained model for natural language generation of indic languages.

Ameet Deshpande, Partha Talukdar, and Karthik Narasimhan. 2021. When is bert multilingual? isolating crucial ingredients for cross-lingual transfer. *arXiv preprint arXiv:2110.14782*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Michael Glass, Mustafa Canim, Alfio Gliozzo, Saneem Chemmengath, Vishwajeet Kumar, Rishav Chakravarti, Avi Sil, Feifei Pan, Samarth Bharadwaj, and Nicolas Rodolfo Fauceglia. 2021. Capturing row and column semantics in transformer based question answering over tables. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1212–1224, Online. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Vivek Gupta, Riyaz A Bhat, Atreya Ghosal, Manish Srivastava, Maneesh Singh, and Vivek Srikumar. 2021. Is my model using the right evidence? systematic probes for examining evidence-based tabular reasoning. *arXiv preprint arXiv:2108.00578*.

Vivek Gupta, Maitrey Mehta, Pegah Nokhiz, and Vivek Srikumar. 2020. INFOTABS: Inference on tables as semi-structured data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2309–2324, Online. Association for Computational Linguistics.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The flores evaluation datasets for low-resource machine translation: Nepali-english and sinhala-english. *arXiv preprint arXiv:1902.01382*.

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

Barry Haddow and Faheem Kirefu. 2020b. Pmindia - A collection of parallel corpora of languages of india. *CoRR*, abs/2001.09907.

Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Eisenschlos. 2020. TaPas: Weakly supervised table parsing via pre-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4320–4333, Online. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Hiroshi Iida, Dung Thai, Varun Manjunatha, and Mohit Iyyer. 2021. TABBIE: Pretrained representations of tabular data. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3446–3456, Online. Association for Computational Linguistics.

Kushal Jain, Adwait Deshpande, Kumar Shridhar, Felix Laumann, and Ayushman Dash. 2020. Indic-transformers: An analysis of transformer language models for indian languages.

Nanjiang Jiang and Marie-Catherine de Marneffe. 2019. Evaluating BERT for natural language inference: A case study on the CommitmentBank. 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 6086–6091, Hong Kong, China. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Karthikeyan K, Aalok Sathe, Somak Aditya, and Monojit Choudhury. 2021. Analyzing the effects of reasoning types on cross-lingual transfer performance.

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.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021a. Muril: Multilingual representations for indian languages.

Simran Khanuja, Melvin Johnson, and Partha Talukdar. 2021b. MergeDistill: Merging language models using pre-trained distillation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2874–2887, Online. Association for Computational Linguistics.

Yash Khemchandani, Sarvesh Mehtani, Vaidehi Patil, Abhijeet Awasthi, Partha Talukdar, and Sunita Sarawagi. 2021. Exploiting language relatedness for low web-resource language model adaptation: An Indic languages study. 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 1312–1323, Online. Association for Computational Linguistics.

Keshav Kolluru, Martin Rezk, Pat Verga, William W Cohen, and Partha Talukdar. 2021. Multilingual fact linking. *arXiv preprint arXiv:2109.14364*.

Kalpesh Krishna, Deepak Nathani, Xavier Garcia, Bidisha Samanta, and Partha Talukdar. 2021. Few-shot controllable style transfer for low-resource settings: A study in indian languages. *arXiv preprint arXiv:2110.07385*.

Jayant Krishnamurthy, Pradeep Dasigi, and Matt Gardner. 2017. Neural semantic parsing with type constraints for semi-structured tables. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1516–1526, Copenhagen, Denmark. Association for Computational Linguistics.

Alex Kulesza and Ben Taskar. 2011. k-dpps: Fixed-size determinantal point processes. In *ICML*.

A Kumaran, Naren Datha, B Ashok, K Saravanan, Anil Ande, Ashwani Sharma, Sridhar Vedantham, Vidya Natampally, Vikram Dendi, and Sandor Maurice. 2010. Wikibabel: A system for multilingual wikipedia content. In *American Machine Translation Association (AMTA) Workshop*. Citeseer.

Anoop Kunchukuttan. 2020. The IndicNLP Library. `https://github.com/anoopkunchukuttan/indic_nlp_library/blob/master/docs/indicnlp.pdf`.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and R Soricut. 2019. A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. 2021. Question and answer test-train overlap in open-domain question answering datasets. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1000–1008, Online. Association for Computational Linguistics.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, et al. 2020. Xglue:

A new benchmark dataset for cross-lingual pre-training, understanding and generation. *arXiv preprint arXiv:2004.01401*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Xi Victoria Lin, Richard Socher, and Caiming Xiong. 2020. Bridging textual and tabular data for cross-domain text-to-SQL semantic parsing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4870–4888, Online. Association for Computational Linguistics.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Bhavnick Minhas, Anant Shankhdhar, Vivek Gupta, Divyanshu Aggrawal, and Shuo Zhang. 2022. Xinfotabs: Evaluating multilingual tabular natural language inference. In *Proceedings of the Fifth Workshop on Fact Extraction and VERification (FEVER)*. Association for Computational Linguistics.

Marc Miquel-Ribé and David Laniado. 2020. The wikipedia diversity observatory: A project to identify and bridge content gaps in wikipedia. In *Proceedings of the 16th International Symposium on Open Collaboration*, pages 1–4.

Linyong Nan, Dragomir Radev, Rui Zhang, Amrit Rau, Abhinand Sivaprasad, Chiachun Hsieh, Xiangru Tang, Aadit Vyas, Neha Verma, Pranav Krishna, Yangxiaokang Liu, Nadia Irwanto, Jessica Pan, Faiaz Rahman, Ahmad Zaidi, Mutethia Mutuma, Yasin Tarabar, Ankit Gupta, Tao Yu, Yi Chern Tan, Xi Victoria Lin, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. 2021. DART: Open-domain structured data record to text generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 432–447, Online. Association for Computational Linguistics.

J. Neeraja, Vivek Gupta, and Vivek Srikumar. 2021. Incorporating external knowledge to enhance tabular reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2799–2809, Online. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Barlas Oguz, Xilun Chen, Vladimir Karpukhin, Stan Peshterliev, Dmytro Okhonko, Michael Schlichtkrull, Sonal Gupta, Yashar Mehdad, and Scott Yih. 2020. Unified open-domain question answering with structured and unstructured knowledge. *arXiv preprint arXiv:2012.14610*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. ToTTo: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1186, Online. Association for Computational Linguistics.

Panupong Pasupat and Percy Liang. 2015. Compositional semantic parsing on semi-structured tables. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1470–1480, Beijing, China. Association for Computational Linguistics.

Vaidehi Patil, Partha Talukdar, and Sunita Sarawagi. 2022. Overlap-based vocabulary generation improves cross-lingual transfer among related languages. *arXiv preprint arXiv:2203.01976*.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Aniket Pramanick and Indrajit Bhattacharya. 2021. Joint learning of representations for web-tables, entities and types using graph convolutional network. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages

1197–1206, Online. Association for Computational Linguistics.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Brian Roark, Lawrence Wolf-Sonkin, Christo Kirov, Sabrina J. Mielke, Cibu Johny, Isin Demirsahin, and Keith Hall. 2020. Processing South Asian languages written in the Latin script: the Dakshina dataset. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2413–2423, Marseille, France. European Language Resources Association.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Anirudh Srinivasan, Sunayana Sitaram, Tanuja Ganu, Sandipan Dandapat, Kalika Bali, and Monojit Choudhury. 2021. Predicting the performance of multilingual nlp models. *arXiv preprint arXiv:2110.08875*.

Huan Sun, Hao Ma, Xiaodong He, Wen-tau Yih, Yu Su, and Xifeng Yan. 2016. Table cell search for question answering. In *Proceedings of the 25th International Conference on World Wide Web*, pages 771–782.

Ishan Tarunesh, Somak Aditya, and Monojit Choudhury. 2021. Lonli: An extensible framework for testing diverse logical reasoning capabilities for nli. *arXiv preprint arXiv:2112.02333*.

Wilson L Taylor. 1953. "cloze procedure": A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Yerram Varun, Aayush Sharma, and Vivek Gupta. 2022. Trans-kblstm: An external knowledge enhanced transformer bilstm model for tabular reasoning. In *Proceedings of Deep Learning Inside Out (DeeLIO): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*. 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.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. 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 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. 2020. TaBERT: Pretraining for joint understanding of textual and tabular data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8413–8426, Online. Association for Computational Linguistics.

Ori Yoran, Alon Talmor, and Jonathan Berant. 2021. Turning tables: Generating examples from semi-structured tables for endowing language models with reasoning skills. *arXiv preprint arXiv:2107.07261*.

Vicky Zayats, Kristina Toutanova, and Mari Ostendorf. 2021. Representations for question answering from documents with tables and text. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2895–2906, Online. Association for Computational Linguistics.

Hongzhi Zhang, Yingyao Wang, Sirui Wang, Xuezhi Cao, Fuzheng Zhang, and Zhongyuan Wang. 2020. Table fact verification with structure-aware transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1624–1629, Online. Association for Computational Linguistics.

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 Details: Table Representation

1. **Premise as Paragraph**: (Chen et al., 2020b), (Gupta et al., 2020) employ universal templates to construct close to natural language sentences for isolated cells in a

Hyper Parameter	XLM-RoBERTa	IndicBERT	MuRIL-base-cased	mBERT-cased
Initial Learning Rate	[1e-4,1e-9]	5e-5	5e-5	5e-5
Batch Size	128	128	128	128
Weight Decay	0.01	0.01	0.01	0.01
Max Seq Length	128	128	128	128
Model Size	278M	33.7M	237M	177M
Warmup Steps	500	500	500	500

Table 5: Hyper Parameters used for Fine-Tuning the corresponding multilingual models.

row, and then, concatenate them to obtain a single paragraph representation. (Gupta et al., 2020) suggest constructing sentences of the form *"The* `k` *of* `t` *is* `v`*"* for a cell having key `k`, value `v` in a table with title `t`. E.g. in figure 1 for the row *Born* the premise sentence would be *"The* `born` *of* `Joe Strummer` *is* `21 August 1952 (1952-08-21) Ankara, Turkey`*"*

However, (Neeraja et al., 2021) identify that such templates can often lead to ungrammatical sentences and propose the ***B****etter* ***P****aragraph* ***R****epresentation* (BPR) approach. BPR utilises type specific templates based on the entity type of a key, and the overall category of the table itself resulting in grammatical sentences. (Neeraja et al., 2021) note a significant increase in performance while employing BPR over the universal template. We adopt BPR as one of our representation approaches. E.g. for same *Born* key in figure 1 the premise sentence with BPR representation would be *"*`Joe Strummer` *was born on* `August 21, 1952 (1952-08-21)` *at* `Ankara, Turkey`*"*

2. **Premise as Structure**: Unlike the natural language like *Premise as Paragraph* representations, here, we try to represent the row as structural text as proposed by (Chen et al., 2020b). Every isolated cell in a row is represented as *"*`k` *:* `v`*"* where `k` is the key, and `v` is the value of the cell. A row's structural representation is a semi-colon ";" separated sequence of the structural representations of all the isolated cells in that row. E.g. for the same *Born* key in figure 1 the premise sentence will be represented as *"*`Born` *:* `August 21, 1952 (1952-08-21), Ankara, Turkey`*"*

B Details: Multilingual Models

Indic Specific: This class of multilingual models are pre-trained entirely on Indic language data along with English. We use *MuRIL Base* (Khanuja et al., 2021a), and *IndicBERT* (kak) pre-trained multilingual models. *MuRIL* is a *BERT* (Devlin et al., 2019) based model trained with *Masked Language Modeling* (Taylor, 1953) and *Translation Language Modeling* (CONNEAU and Lample, 2019) objectives. It is trained on (a.) Common Crawl OSCAR corpus[7] and Wikipedia[8] monolingual data for 16 Indic languages along with the English language, (b.) PMIndia (Haddow and Kirefu, 2020b) along with other in-house parallel corpora, (c.) and the Dakshina Dataset (Roark et al., 2020) along with other parallel in-house transliterated corpora. *IndicBERT* is an *ALBERT* (Lan et al., 2019) based model trained on *IndicCorp* (kak).

Generic: This class of multilingual models are pre-trained on a wide array of languages from around the world. We use *mBERT-cased* (Devlin et al., 2019) and *XLM-RoBERTa* (con) pre-trained multilingual models.

C Human Evaluation Strategy

We requested our colleagues who are native speakers and are proficient in English to help us with this task while disclosing the intentions. We provide them with instructions adopted from the Direct Assessment (Graham et al., 2013) strategy for low resource machine translation in (Guzmán et al., 2019). We sample 50 pairs of source, translation pairs and ask the annotators to provide a continuous score between 0 to 100. 0–10 range represents a translation that is completely incorrect and inaccurate. 70–90 range represents a translation that closely preserves the meaning of the source sentence while the 90–100 range represents a perfect translation.

[7] Oscar Corpus [8] Tensorflow Datasets

D Model Hyper-Parameters

Table 5 reports the hyper-parameters used for fine-tuning the multilingual models on EI-INFOTABS. We use the Huggingface Transformers[9] library to script these experiments. We were unable to successfully converge XLM-RoBERTa in multiple runs spanning a distinctive set of hyper-parameters. Figure 5 shows the loss plots for XLM-RoBERTa and mBERT when fine-tuned on EI-INFOTABS. It is distinctively visible that XLM-RoBERTa is unable to converge on EI-INFOTABS on a significant amount of steps unlike mBERT.

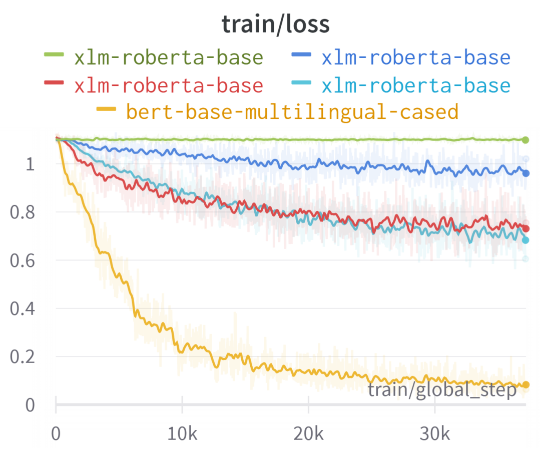

Figure 5: Train Loss for multiple runs of XLM-RoBERTa with distinct set of hyper-parameters compared with that of mBERT. Each run spans roughly 37,000 steps.

Fine-Tuning Settings. We follow the conventionally used pipeline for fine-tuning *BERT* for Sequence Classification (Jiang and de Marneffe, 2019). We concatenate the premise and the hypothesis strings using a `[SEP]` token in between them, prepend this sequence with a `[CLS]` token, tokenize this sequence using the pre-trained tokenizer for the respective model, and provide the obtained sequence as input to the pre-trained model. We attach a three-way classification head with cross-entropy loss on top of the pooled output obtained from the previous step. With an initial learning rate of 5e-05 with AdamW optimizer (Loshchilov and Hutter, 2018), we fine-tune each model on 4 1080Ti GPUs with a batch size of 32 per GPU over 10 epochs.

[9] Transformer Hugging Face

E Performance on the α_2 and α_3 Adversarial Sets

Tables 6 and 7 report the results for the adverserial test sets α_2 and α_3 respectively using the BPR linearization method.

F Zero Shot Cross-Lingual Transfer

Tables 8 and 9 report the performance of MuRIL on Translate-Train-X and Bilingual-Train-X. We note that models trained on linguistically closer pairs of languages are able to admirably transfer their performance to each other. Notably, Assamese ('as') and Bengali ('bn') being immensely closely related, support this hypothesis. Moreover, we note the same for closely related Indo-European languages Bengali, Hindi, Gujarati ('gu'), and Marathi ('mr'). Models trained on these languages distinctively transfer their performance better on each other compared to languages from the Dravidian language family - Malayalam ('ml'), Telugu ('te'), Tamil ('ta'), Kannada ('kn'). Dravidian languages are not as closely related due to differences in scripts and sentence structures which is observed in the results as well.

Strategy	Model	bn	hi	gu	pa	mr	te	ta	ml	kn	as	or	ModAvg
Translate-Train	mBERT	51	**52**	48	49	48	48	49	48	49	47	36	48
	IndicBERT	**46**	44	44	44	**46**	**46**	45	45	**46**	34	45	44
	MuRIL	**56***	**56***	**52**	**55**	**54**	**52**	**55**	**53**	**55**	**54**	**53**	54
	LnAvg	51	51	48	49	49	49	50	48	50	45	45	49
Translate-Test	mBERT	**47**	**47**	44	45	44	44	45	43	46	41	34	44
	IndicBERT	41	38	37	38	39	38	38	38	38	**42**	39	39
	MuRIL	**52**	**51**	**51**	**50**	**50**	**51**	**51**	**53***	**53***	49	50	51
	LnAvg	47	45	44	44	44	45	44	44	45	44	41	44
Bilingual-Train	mBERT	**52**	**52**	50	51	50	50	51	49	51	50	37	49
	IndicBERT	45	45	45	**48**	47	45	46	45	46	44	45	45
	MuRIL	**56***	**55**	**54**	**56***	**54**	**54**	**53**	**54**	**56***	**54**	**53**	54
	LnAvg	51	51	50	51	50	50	50	49	51	49	45	50
Multilingual-Train	mBERT	50	**51**	**51**	50	50	**51**	49	48	50	48	35	48
	IndicBERT	46	46	46	**47**	45	45	44	44	45	45	44	45
	MuRIL	**55**	**55**	**56***	**55**	**55**	**54**	**55**	**55**	**55**	**54**	**54**	55
	LnAvg	50	51	51	51	50	50	49	49	50	49	45	50
EnTranslate-Test	mBERT	**55**	**55**	**55**	53	53	54	54	54	54	54	53	54
	IndicBERT	**48**	47	47	**48**	47	47	46	46	46	47	47	47
	MuRIL	**56***	**55**	**55**	**56***	**54**	**55**	**54**	**55**	**54**	**54**	**54**	55
	LnAvg	53	52	52	52	51	52	51	52	51	52	52	52
Translate-Train-X	mBERT	**45**	**45**	44	**45**	42	42	**45**	41	**45**	40	35	43
	IndicBERT	**40**	38	37	36	37	39	38	**40**	38	34	38	38
	MuRIL	**54***	**53**	**52**	**54***	**53**	**52**	**53**	**51**	**52**	**54***	**52**	53
	LnAvg	46	45	44	45	44	44	45	44	45	42	41	44
Bilingual-Train-X	mBERT	**47**	**47**	46	46	46	46	46	44	46	**47**	43	46
	IndicBERT	39	**40**	39	39	39	39	**40**	39	39	**40**	39	39
	MuRIL	**54**	**54**	**54**	**55***	**54**	**53**	**53**	**54**	**54**	**54**	**54**	54
	LnAvg	46	47	46	47	47	46	46	46	46	47	45	46

Table 6: Performance in terms of accuracy when evaluated on the α_2 test set. Higher value implies better performance. Here, LnAvg represents the average accuracy for a language across all models, while ModAvg represents the average accuracy of a model across all languages. A value in **Purple** represents the best accuracy for that model across all languages. An underlined value in **Blue** represents the best accuracy for that language across all models. A value in **Green** with an asterisk(*) represents the cases where language-wise and model-wise values coincide. As we fine-tune on a specific Indic language in the fine-tuning strategies Translate-Train-X and Bilingual-Train-X, we report the training average of the concerned language. We do not include the results of XLM-RoBERTa as the model fails to converge on these experiments on multiple runs with a distinct set of hyper-parameters as explained in Appendix §D.

Strategy	Model	bn	hi	gu	pa	mr	te	ta	ml	kn	as	or	ModAvg
Translate-Train	mBERT	47	**48**	46	45	46	46	47	46	46	43	35	45
	IndicBERT	43	**44**	40	42	41	42	43	39	43	33	41	41
	MuRIL	**52**	**54***	**52**	**53**	**52**	**51**	**52**	**51**	**54**	**51**	**50**	52
	LnAvg	47	49	46	47	46	47	47	45	48	43	42	46
Translate-Test	mBERT	44	**46**	43	45	43	45	**46**	43	44	39	33	43
	IndicBERT	**36**	**36**	35	34	35	35	35	35	35	**36**	34	35
	MuRIL	**53***	**52**	**51**	**51**	**51**	**50**	**51**	**51**	**50**	**50**	**49**	51
	LnAvg	45	44	43	43	43	43	44	43	43	42	39	43
Bilingual-Train	mBERT	**49**	**49**	47	47	48	46	**49**	47	**49**	46	34	46
	IndicBERT	42	41	42	42	40	42	41	**44**	42	42	41	42
	MuRIL	52	**53***	**52**	**51**	**51**	**52**	**51**	**52**	**51**	**53**	**51**	52
	LnAvg	48	48	47	47	46	47	47	47	47	47	42	47
Multilingual-Train	mBERT	**47**	**47**	46	**47**	46	45	46	45	**47**	46	36	45
	IndicBERT	**42**	41	**42**	40	40	**42**	41	41	40	**42**	40	41
	MuRIL	**54***	**54***	**52**	**53**	**52**	**53**	**53**	**53**	**54***	**54***	**53**	53
	LnAvg	47	47	47	47	46	47	47	46	47	47	43	46
EnTranslate-Test	mBERT	51	**52**	50	51	50	50	**52**	49	50	50	49	50
	IndicBERT	46	**48**	46	46	46	46	46	45	46	45	44	46
	MuRIL	**53***	**52**	**51**	**51**	**51**	**50**	**50**	**50**	**51**	**50**	**48**	51
	LnAvg	50	51	49	49	49	49	49	48	49	48	47	49
Translate-Train-X	mBERT	**44**	**44**	43	**44**	43	**44**	**44**	43	**44**	41	34	42
	IndicBERT	**37**	36	36	35	36	**37**	**37**	**37**	36	33	36	36
	MuRIL	**51**	**52***	**52***	**52***	**50**	**51**	**51**	**51**	**52***	**51**	**51**	51
	LnAvg	44	44	43	44	43	44	44	43	44	42	40	43
Bilingual-TrainX	mBERT	**45**	44	44	44	44	44	**45**	44	**45**	44	37	44
	IndicBERT	**37**	36	36	36	36	**37**	38	**37**	**37**	**37**	36	37
	MuRIL	**51**	**51**	**51**	**51**	**51**	**51**	**52***	**52***	**51**	**52***	**52***	51
	LnAvg	44	44	44	44	44	44	45	44	44	44	42	44

Table 7: Performance in terms of accuracy when evaluated on the α_3 test set. Higher value implies better performance. Here, LnAvg represents the average accuracy for a language across all models, while ModAvg represents the average accuracy of a model across all languages. A value in **Purple** represents the best accuracy for that model across all languages. An underlined value in **Blue** represents the best accuracy for that language across all models. A value in **Green** with an asterisk(*) represents the cases where language-wise and model-wise values coincide. As we fine-tune on a specific Indic language in the fine-tuning strategies Translate-Train-X and Bilingual-Train-X, we report the training average of the concerned language. We do not include the results of XLM-RoBERTa as the model fails to converge on these experiments on multiple runs with a distinct set of hyper-parameters as explained in Appendix §D.

	bn	hi	gu	pa	mr	te	ta	ml	kn	as	or	TrainAvg
bn	67	66	64	62	63	63	63	60	63	64	62	63
hi	66	67	65	65	64	62	63	64	65	62	62	64
gu	63	64	66	65	62	64	63	63	63	63	64	64
pa	63	64	63	65	61	61	62	62	62	62	61	62
mr	65	66	63	64	65	62	62	63	64	62	62	63
te	65	62	63	64	62	64	62	63	64	63	63	63
ta	63	64	63	61	62	62	65	62	64	61	59	62
ml	65	62	62	63	62	63	62	65	64	63	61	63
kn	65	65	65	64	63	63	63	64	66	62	62	64
as	63	63	63	62	63	62	63	63	63	65	61	63
or	64	61	60	62	60	61	61	60	62	62	64	61
TestAvg	64	64	63	63	62	62	63	63	64	63	62	63

	bn	hi	gu	pa	mr	te	ta	ml	kn	as	or	TrainAvg
bn	55	54	54	53	53	52	53	53	53	54	52	53
hi	54	56	53	52	52	51	52	52	53	53	52	53
gu	54	52	52	50	51	50	52	50	51	50	50	51
pa	55	54	54	55	53	52	54	52	53	52	53	53
mr	54	53	53	53	53	51	52	51	52	52	52	52
te	54	52	52	52	52	51	51	52	52	52	51	52
ta	56	55	54	51	53	53	54	52	53	51	51	53
ml	53	50	53	49	51	50	50	52	52	51	50	51
kn	54	53	52	51	51	50	52	50	54	51	51	52
as	56	53	55	52	53	52	53	53	54	54	51	53
or	55	51	50	53	51	51	51	51	52	51	53	52
TestAvg	54	53	53	52	52	51	52	52	53	52	51	52

	bn	hi	gu	pa	mr	te	ta	ml	kn	as	or	TrainAvg
bn	52	51	50	51	50	49	50	49	51	52	51	50
hi	53	54	51	53	52	51	52	51	52	50	50	52
gu	53	51	52	52	50	51	50	52	51	50	50	51
pa	53	52	51	53	51	51	53	51	51	51	52	52
mr	52	50	50	51	51	48	49	49	49	50	49	50
te	51	51	49	50	50	51	50	50	51	49	49	50
ta	53	52	50	48	51	49	52	51	50	51	50	51
ml	52	49	50	51	52	50	49	50	51	50	49	50
kn	52	53	50	51	50	51	51	51	53	51	50	51
as	51	50	50	51	50	49	50	50	51	51	52	51
or	52	51	49	50	51	50	50	51	51	50	50	50
TestAvg	52	51	50	51	51	50	51	51	51	50	50	51

Table 8: Complete results (accuracy) for Translate-Train-X of MuRIL on the α_1, α_2, and α_3 test splits respectively.

	bn	hi	gu	pa	mr	te	ta	ml	kn	as	or	TrainAvg
bn	67	67	64	65	64	63	63	64	64	63	63	64
hi	65	67	63	65	63	63	63	62	63	63	61	63
gu	65	67	66	65	63	64	64	64	65	64	62	64
pa	66	66	65	66	64	62	64	64	64	64	63	64
mr	64	67	64	65	65	62	64	63	64	64	61	64
te	66	66	65	65	64	66	65	64	65	64	64	65
ta	65	67	65	64	63	63	65	63	64	63	63	64
ml	67	67	66	66	63	63	64	66	63	63	61	64
kn	67	67	65	65	64	63	64	64	66	63	63	65
as	66	66	65	64	65	63	64	64	64	65	62	64
or	65	68	65	65	64	63	65	63	65	63	65	64
TestAvg	66	67	65	65	64	63	64	63	64	63	63	64

	bn	hi	gu	pa	mr	te	ta	ml	kn	as	or	TrainAvg
bn	55	54	54	53	53	53	53	52	53	52	52	53
hi	55	54	54	54	53	52	54	53	54	52	52	53
gu	54	54	53	53	53	53	54	53	53	54	52	53
pa	55	55	54	55	53	52	55	54	56	54	53	54
mr	55	54	54	53	53	53	54	53	55	53	52	54
te	55	53	53	53	53	54	53	51	53	53	52	53
ta	53	53	53	53	52	52	53	52	54	52	51	52
ml	57	54	55	53	54	53	53	53	54	53	52	54
kn	56	54	53	54	52	51	53	53	56	54	52	53
as	56	54	54	54	52	53	54	54	55	53	52	54
or	55	54	53	53	52	53	54	54	54	53	53	53
TestAvg	55	54	54	54	53	53	54	53	54	53	52	53

	bn	hi	gu	pa	mr	te	ta	ml	kn	as	or	TrainAvg
bn	52	50	49	50	52	51	50	52	50	49	49	50
hi	51	52	49	52	50	49	50	51	51	49	48	50
gu	51	51	51	51	51	51	50	51	51	49	49	51
pa	52	52	50	51	51	51	50	52	50	49	48	51
mr	52	50	51	50	51	50	50	51	50	49	49	50
te	51	51	51	51	52	52	51	51	51	49	50	51
ta	52	52	51	52	51	51	51	51	51	51	50	51
ml	53	52	52	50	52	52	51	52	50	50	50	51
kn	50	52	51	52	50	51	51	50	50	49	48	50
as	53	52	51	52	52	52	51	52	50	53	49	52
or	53	52	51	52	51	50	52	52	51	51	50	51
TestAvg	52	52	51	51	51	51	51	51	50	50	49	51

Table 9: Complete results (accuracy) for Bilingual-Train-X of MuRIL on the α_1, α_2, and α_3 test splits respectively.