XINFOTABS: Evaluating Multilingual Tabular Natural Language Inference

Bhavnick Minhas[1*], **Anant Shankhdhar**[1*], **Vivek Gupta**[2*†], **Divyanshu Aggrawal**[3], **Shuo Zhang**[4]
[1]Indian Institute of Technology, Guwahati; [2]School of Computing, University of Utah
[3]Delhi Technological University; [4]Bloomberg
{bhavnick, anant.shankhdhar}@iitg.ac.in; vgupta@cs.utah.edu;
divyanshuggrwl@gmail.com; szhang611@bloomberg.net

Abstract

The ability to reason about tabular or semi-structured knowledge is a fundamental problem for today's Natural Language Processing (NLP) systems. While significant progress has been achieved in the direction of tabular reasoning, these advances are limited to English due to the absence of multilingual benchmark datasets for semi-structured data. In this paper, we use machine translation methods to construct a multilingual tabular natural language inference (TNLI) dataset, namely XINFOTABS, which expands the English TNLI dataset of INFOTABS to ten diverse languages. We also present several baselines for multilingual tabular reasoning, e.g., machine translation-based methods and cross-lingual TNLI. We discover that the XINFOTABS evaluation suite is both practical and challenging. As a result, this dataset will contribute to increased linguistic inclusion in tabular reasoning research and applications.

1 Introduction

Natural Language Inference (NLI) on semi-structured knowledge like tables is a crucial challenge for existing (NLP) models. Recently, two datasets, TabFact (Chen et al., 2019) on Wikipedia relational tables and INFOTABS (Gupta et al., 2020) on Wikipedia Infoboxes, have been proposed to investigate this problem. Among the solutions, contextual models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), when adapted for tabular data, surprisingly achieve remarkable performance.

The recent development of multi-lingual extensions of contextualizing models such as mBERT (Devlin et al., 2019) from BERT and XLM-RoBERTa (Conneau et al., 2020) from RoBERTa, has led to substantial interest in the problem of multi-lingual NLI and the creation of multi-lingual XNLI (Conneau et al., 2018) and TaxiXNLI (K et al., 2021) dataset from English MNLI (Williams et al., 2018) dataset. However, there is still no equivalent multi-lingual NLI dataset for semi-structured tabular data. To fill this gap, we propose XINFOTABS, a multi-lingual extension of INFOTABS dataset. The XINFOTABS dataset consists of ten languages, namely English ('en'), German ('de'), French ('fr'), Spanish ('es'), Afrikaans ('af'), Russian ('ru'), Chinese ('zh'), Korean ('ko'), Hindi ('hi') and Arabic ('ar'), which belong to seven distinct language families and six unique writing scripts. Furthermore, these languages are the majority spoken in all seven continents covering 2.76 billion native speakers in comparison to 360 million English language (INFOTABS) speakers[1].

The intuitive method of constructing XINFOTABS, i.e., human-driven manual translation, is too expensive in terms of money and time. Alternatively, various state-of-the-art machine translation models, such as mBART50 (Tang et al., 2020), MarianMT (Junczys-Dowmunt et al., 2018), M2M100 (Fan et al., 2020a), have greatly enhanced translation quality across a broad variety of languages. Furthermore, NLI requires simply that the translation models retain the semantics of the premises and hypotheses, which machine translation can deliver (K et al., 2021). Therefore, we use automatic machine translation models to construct XINFOTABS from INFOTABS.

Tabular data is far more challenging to translate than semantically complete and grammatical sentences with existing state-of-the-art translation systems. To mitigate this challenge, we propose an efficient, high-quality translation pipeline that utilizes Name Entity Recognition (NER) and table context in the form of category information to convert table cells into structured sentences before

*Equal Contribution †Corresponding Author

[1] Refer to Appendix Table 5 for more information.

Boxing (en)	
Focus	Punching, striking
Olympic sport	688 BC (Ancient Greece), 1904 (modern)
Parenthood	Bare-knuckle boxing
Country of origin	Prehistoric
Also known as	Western Boxing, Pugilism See note.

Boxe (fr)	
Focus	Punching, frappe
Sport olympique	688 av. J.-C. (Grèce ancienne), 1904 (moderne)
Parentalité	Bare-knuckle boxe
Pays d'origine	Préhistorique
Aussi connu sous le nom	Western Boxing, Pugilism Voir note.

Language	Hypothesis	Label
English	The modern form of boxing started in the late 1900's.	CONTRADICTION
German	Boxen hat seinen Ursprung als olympischer Sport, der vor Jahrtausenden begann.	CONTRADICTION
French	La boxe occidentale implique des punches et des frappes	ENTAILMENT
Spanish	El boxeo ha sido un evento olímpico moderno durante más de 100 años.	ENTAILMENT
Afrikaans	Bare-knuckle boks is 'n prehistoriese vorm van boks.	NEUTRAL

Table 1: An example of the XInfoTabS dataset containing English (top-left) and French (top-right) tables in parallel with the hypothesis associated with the table in five languages (below).

translation. We assess the translations via several automatic and human verification methods to ensure quality. Our translations were found to be accurate for the majority of languages, with German and Arabic having the most and least exact translations, respectively. Table 1 shows an example from the XINFOTABS dataset.

We conduct tabular NLI experiments using XINFOTABS in monolingual and multilingual settings. By doing so, we aim to assess the capacity and cross-lingual transferability of state-of-the-art multilingual models such as mBERT (Devlin et al., 2019), and XLM-Roberta (Conneau et al., 2020). Our investigations reveal that these multilingual models, when assessed for additional languages, perform comparably to English. Second, the translation-based technique outperforms all other approaches on the adversarial evaluation sets for multilingual tabular NLI in terms of performance. Thirdly, the method of intermediate-task finetuning, also known as pre-finetuning, significantly improves performance by finetuning on additional languages prior to the target language. Finally, these models perform admirably on cross-lingual tabular NLI (tables and hypotheses given in different languages), although the additional effort is required to improve them. Our contributions are as follows:

- We introduce XINFOTABS, a multi-lingual extension of INFOTABS, a semi-structured tabular inference English dataset over ten diverse languages.
- We propose an efficient pipeline for high-quality translations of semi-structured tabular data using state-of-the-art translation models.
- We conduct intensive inference experiments on XINFOTABS and evaluate the performance of state-of-the-art multilingual models with various strategies.

The dataset and associated scripts, is available at https://xinfotabs.github.io/.

2 Why the INFOTABS dataset?

There are only two public datasets, both in English, available for semi-structured tabular reasoning, namely TabFact (Chen et al., 2019) and INFOTABS (Gupta et al., 2020). We choose INFOTABS because it includes multiple adversarial test sets for model evaluation. Additionally, the INFOTABS dataset also includes the NEUTRAL label, which is absent in TabFact. The INFOTABS dataset contains 2,540 tables serving as premise and 23,738 hypothesis sentences along with associated inference labels. The table-sentence pairs are divided into development, and three evaluation sets α_1, α_2, and α_3, each containing 200 unique tables along with nine hypothesis sentences equally distributed among three inference labels (ENTAILMENT, CONTRADICTION, and NEUTRAL). α_1 is a conventional evaluation set that is lexically similar to the training data. α_2 has lexically adversarial hypotheses. And α_3 contains domain topics that are not present in the training set. The remaining 1,740 tables with corresponding 16,538 hypotheses serve as a training set. Table 2 describes the inference performance of RoBERTa$_L$ model on INFOTABS dataset. As we can see, the Human Scores are superior to that of RoBERTa$_L$ model trained with TabFact representation. Since the XINFOTABS is

translated directly from the INFOTABS, we expect a similar human baseline for XINFOTABS.

Model	dev	α_1	α_2	α_3
Human	**79.78**	**84.04**	**83.88**	**79.33**
Hypo Only	60.51	60.48	48.26	48.89
RoBERTa$_{\text{LARGE}}$	77.61	75.06	69.02	64.61

Table 2: Accuracy scores of the *Table as Struct* strategy on XINFOTABS subsets with RoBERTa$_{\text{LARGE}}$ model, hypothesis only baseline and majority human agreement results. The first three rows are reproduced from Gupta et al. (2020).

3 Table Representation

Machine translation of tabular data is a challenging task. Tabular data is semi-structured, non-sentential (ungrammatical), and succinct. The tight form of tabular cells provides inadequate context for today's machine translation models, which are primarily designed to handle sentences. Thus, table translation requires additional context and conversion. Furthermore, frequently occurring named entities in tables must be transliterated rather than translated. Figure 1 shows the table translation pipeline. We describe our approach to context addition and handling of named entities in detail in the following subsections §3.1.

3.1 Table Translation Context

There are several ways to represent tables, each with its own set of pros and cons, as detailed below:

Without Context. The most straightforward way to represent a table would be to treat every key (header) and value (cell) as separate entities and then translate them independently. This approach results in poor translations as the models have no context regarding the keys. The key *"Length"* in English in context of *Movies* would correspond to *"durée"*, meaning *duration* in French but in *Object* context, would correspond to *"longueur"*, meaning *size or span*. Thus, context is essential for accurate table translation.

Full Table. Before transferring data from the header and table cells to translation models, one may concentrate and seam each table row using a delimiter such as a colon (":") to separate key from value and a semi-colon (";") to separate rows (Wenhu Chen and Wang, 2020). This method provides full context and completely translates all table cells. However, in practice, this strategy has two major problems:

a. *Length Constraint*: All transformer-based models have a maximum input string length of 512 tokens.[2] Larger tables with tens of rows may not be translated using this approach.[3] In practice, strings longer than 256 tokens have been shown to have inferior translation quality.[4]

b. *Structural Issue*: When a linearized table is directly translated, the delimiter tokens (":" and ";") get randomly shifted. [5] The delimiter counts are also altered. Hence, the translation appears to merge characters from adjacent rows, resulting in inseparable translations. Ideally, the key and value delimiter token locations should be invariant in a successful translation.

Category Context. Given the shortcomings of the previous two methods, we devise a new strategy: we add a *general context* that describes table rows at a high level to each linearized row cell. We leverage the *table category* here, as it offers enough context to grasp the key's meaning. For the key *"Focus"* in Table 1, the category information *Sports* offers enough context to understand its significance in relation to boxing. The context added representation for this key-value pair will be *"Sports | Focus | Punching , Striking"*. We use "|" delimiter for separating the context, key, and value. Furthermore, multiple values are seperated by ",". Unlike full table translation, row structure is preserved since each row is translated independently and no row surpasses the maximum token limit. We observe an average increase of 5.5% in translation performance (cf. §4).

3.2 Handling Named Entities

Commercial translation methods, like Google Translate, correctly transliterate specified entities (such as proper nouns and dates). However, modern open-source models like mBART50 and M2M100 translate name entity labels, lowering overall translation quality. For example, *Alice Sheets* is translated to *Alice draps* in French. We propose a simple preprocessing technique to address the transliterate/translate ambiguity. First, we use the Named Entity Recognition (NER) model[6](Jiang et al., 2016) to identify entity information that must be transliterated, such as proper nouns and dates. Then, we add a unique identifier in the form

[2] Recently, models bigger than 512 tokens have been developed, e.g. (Asaadi et al., 2019; Beltagy et al., 2020), but no publicly accessible long-sequence (> 512 tokens) multilingual machine translation model exists at the moment. [3] Average # of rows in InfoTabS is: 8.8 for Train, Development, α_1 and α_2, and 13.1 for α_3. [4] Neeraja et al. (2021) raises a similar issue for NLI. [5] Using "|" instead of ":" helps key-value separation. [6] spaCy NER tagger

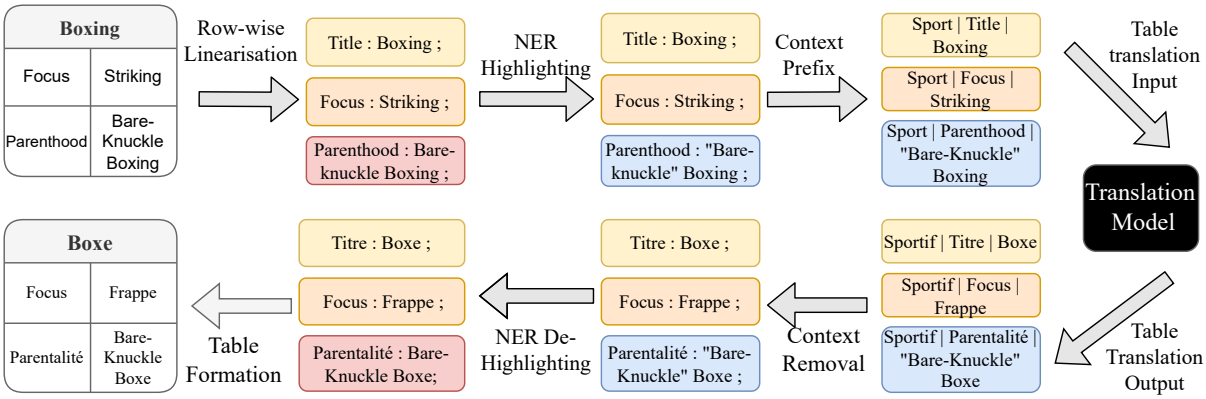


Figure 1: Table translation pipeline (§3) with premise table "Boxing" (from INFOTABS) translated into French.

of double quotations (" "), e.g., *"Alice Sheets"*, and apply the translation model. Finally, we delete the quotation mark (" ") from the translated sentence after it has been translated. This helps the models identify these entities easily due to their pre-training.

4 Translation and Verification

As mentioned previously, we now grasp how to represent a table. Consequently, these reformatted tables can now be fed into reliable translation models. To accomplish this, we assess many prominent multilingual (e.g., mBART50 (Tang et al., 2020) and M2M100 (Fan et al., 2020b)) and bilingual (e.g., MarianMT (Junczys-Dowmunt et al., 2018)) translation models as described below:

Multilingual Models. This category of models used includes widely used machine translation models trained on a large number of languages such as mBART50 (Tang et al., 2020) which can perform translation between any two languages from the list of 50 languages and M2M100 (Fan et al., 2020b) which has 100 training languages. Apart from these models, we used Google Translate[7] to compare against our dataset translation quality.

Bilingual Models. Earlier studies have revealed that bilingual models outperform multilingual models in machine translation of high-resource languages. Thus, for our experiments, we also considered language-specific bilingual translation models in MarianMT (Junczys-Dowmunt et al., 2018) repository. Because the MarianMT models were not available for a few languages (e.g., Korean (ko)) of XINFOTABS, we could not conduct experiments for some languages.

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

We also use an efficient data sampling technique to determine the ideal translation model for each language, as detailed in the next section. The results for the translations are shown in Table 3.

4.1 Translation Model Selection

Translating the complete INFOTABS dataset to find the optimal model is practically infeasible. Thus, we select a representative subset of the dataset that approximates the full dataset rather well. Finally, we use optimal models to translate the complete INFOTABS dataset. The method used for making the subset is discussed in the ***Table Subset Sampling Strategy*** and ***Hypothesis Subset Sampling Strategy*** sections given below:-

Table Subset Sampling Strategy: In a table, keys can serve as an excellent depiction of the type of data included therein. For example, if the key *"children"* is used, the associated value is almost always a valid *Noun Phrase* or a collection of them. Additionally, the type of keys for a given category remains constant across tables, but the values are always different.[8] This fact is used to sample a subset of diverse tables based on keys and categories. Specifically, we sample tables for each category based on the frequency of occurrence of keys in the dataset to guarantee diversity. The sum of the frequencies of all the keys in a table is computed for each table. Finally, the top 10% of tables with the largest frequency sum in each category are chosen to be included in the subset. In the end, we construct a subset with 11.14% tables yet containing 90.2% of the all unique keys.

Hypothesis Subset Sampling Strategy: To get a diverse subset of hypotheses, we employ Top2Vec (Angelov, 2020) embedding for each

[8] There are 2,163 unique keys in INFOTABS.

hypothesis, then use k-means clustering (Jin and Han, 2010) to choose 10% of each cluster. Sampling from each cluster ensures we cover all topics discussed in the hypothesis, resulting in a subset of 2,569 hypothesis texts.

Model Selection Strategy: To choose the translation model that will be used to generate the language datasets, we first translate the premise and hypothesis subsets for all languages using each of the existing models, as described before. Following translation, we compute the various scores detailed in Section 4.2. Finally, the model with the highest average of premise and hypothesis translation ***Human Evaluation Score*** for the specified language is chosen to translate the complete INFOTABS datasets.

4.2 Translation Quality Verification

With the emergence of Transformer-based pre-trained models, significant progress has been made in automated quality assessment using semantic similarity and human sense correlation (Cer et al., 2017) for machine translation evaluation. To verify our created dataset XINFOTABS, we use three automated metrics in addition to human ratings.

Paraphrase Score (PS). PS indicates the amount of information retained from the translated text. To capture this, we estimate the cosine similarity between the original INFOTABS text and the back-translated English XINFOTABS text sentence encodings. We utilize the all-mpnet-v2(Song et al., 2020) model trained using SBERT (Reimers and Gurevych, 2019) method for sentence encoding.

Multilingual Paraphrase Score (mPS). Different from PS, mPS directly uses the multilingual XINFOTABS text instead of the English back-translated text to compare with INFOTABS text. We produce sentence encodings for multilingual semantic similarity using the multilingual-mpnet-base-v2 model (Reimers and Gurevych, 2020) trained using the SBERT method.

BERTScore (BS). BERTScore is an automatic score that shows high human correlation and has been a widely used quality estimation metric for machine translation tasks (Zhang et al., 2019).

Human Evaluation Score (HES) We hired five annotators to label sampled subsets of 500 examples per model and language. Human verification is accomplished by supplying sentence pairs and requesting that annotators classify them as identical or dissimilar based on the meaning expressed by the sentences. For more details, refer to the Appendix §A.

Analysis. We arrive at an average language score of 85 for tables and 91 for hypotheses for the final selected models in all languages. The results are summarised in Table 3. These results are also utilized to determine the optimal models for translating the entire dataset. MarianMT is used to create the entire dataset in German, French, and Spanish, mBART50 is used to create the Tables dataset in Afrikaans, Korean, Hindi, and Arabic, and M2M100 is used to create the entire dataset in Russian and Chinese, as well as the hypothesis dataset in Afrikaans, Korean, Hindi, and Arabic.

5 Experiment and Analysis

In this section, we study the task of Multilingual Tabular NLI, utilizing our XINFOTABS dataset as the benchmark for a variety of multilingual models with multiple training-testing strategies. By doing so, we aim to assess the capacity and cross-lingual transferability of state-of-the-art multilingual models. For the inference task, we linearize the table using the *"Table as Struct"-TabFact* described in INFOTABS.

Multilingual Models: We use pre-trained multilingual models for all our inference label prediction experiments. We use a multilingual mBERT-base (cased) (Devlin et al., 2019) model pre-trained on masked language modeling. This model will be referred to as $\text{mBERT}_{\text{BASE}}$. The other model we evaluated is the XLM-RoBERTa Large (XNLI) model (Conneau et al., 2020), which is trained on masked language modeling and then finetuned for the NLI task using the XNLI dataset. This model is referred to as XLM-R Large (XNLI). For details on hyperparameters, refer to Appendix §B.

Tables 4, 6, and 7 show the performance of the discussed multilingual models for α_1, α_2, and α_3 test splits respectively. Tables 6 and 7 are shown in Appendix §C, due to limited space. On all three evaluation sets, regardless of task type, the $\text{XLM-RoBERTa}_{\text{Large}}$ model outperforms mBERT. This might be because XLM-RoBERTa has more parameters, and is better pre-trained and pre-tuned for the NLI task using the XNLI dataset.

Model	Metric	de	fr	es	af	ru	zh	ko	hi	ar	MdlAvg
MarianMT	PS	95 | 96	93 | 95	93 | 96	83 | 88	81 | 87	75 | 85	N.A.	56 | 55	60 | 79	80 | 85
	mPS	92 | 95	87 | 96	90 | 96	83 | 84	78 | 84	79 | 83	N.A.	65 | 64	66 | 74	80 | 85
	BS	93 | 94	91 | 94	92 | 94	84 | 89	81 | 87	73 | 85	N.A.	63 | 68	64 | 83	80 | 87
	HES	95 | 87	92 | 86	92 | 94	70 | 56	84 | 54	75 | 59	N.A.	40 | 23	58 | 56	76 | 64
	LnAvg	**94** | **93**	**91** | **93**	**92** | **95**	80 | 79	81 | 78	76 | 78	N.A.	56 | 53	62 | 73	79 | 80
mBART50	PS	94 | 96	93 | 95	86 | 87	88 | 92	89 | 87	81 | 87	83 | 82	85 | 82	70 | 77	85 | 87
	mPS	92 | 96	90 | 96	72 | 92	85 | 91	81 | 88	79 | 84	86 | 83	79 | 81	80 | 80	83 | 88
	BS	91 | 94	91 | 93	71 | 88	88 | 93	85 | 89	77 | 86	79 | 85	82 | 86	76 | 83	82 | 89
	HES	93 | 84	91 | 81	82 | 80	89 | 69	87 | 69	76 | 61	76 | 54	79 | 70	71 | 53	83 | 69
	LnAvg	93 | 93	91 | 91	78 | 87	**88** | 86	86 | 83	78 | 80	**81** | 76	**81** | 80	**74** | 73	83 | 83
M2M100	PS	89 | 96	92 | 94	88 | 95	91 | 94	89 | 90	83 | 82	83 | 92	83 | 88	72 | 77	86 | 90
	mPS	88 | 96	88 | 96	88 | 96	84 | 92	83 | 88	80 | 86	84 | 90	81 | 87	78 | 92	84 | 91
	BS	87 | 94	89 | 93	86 | 93	89 | 94	87 | 90	81 | 88	80 | 90	81 | 89	73 | 88	84 | 91
	HES	88 | 85	86 | 86	84 | 86	86 | 83	87 | 74	79 | 72	70 | 82	75 | 73	60 | 51	79 | 77
	LnAvg	88 | 93	89 | 92	87 | 93	88 | **91**	**87** | **86**	**81** | **82**	79 | **89**	80 | **84**	71 | **77**	83 | 87
GoogleTr	PS	91 | 94	94 | 93	92 | 93	96 | 95	79 | 86	80 | 83	87 | 89	90 | 85	60 | 81	85 | 89
	mPS	89 | 94	88 | 94	88 | 94	82 | 87	82 | 86	80 | 86	83 | 87	77 | 80	71 | 81	82 | 88
	BS	87 | 91	89 | 90	88 | 91	88 | 93	77 | 85	78 | 82	82 | 85	87 | 85	63 | 82	82 | 87
	HES	91 | 79	93 | 81	89 | 83	96 | 81	84 | 66	79 | 56	79 | 70	92 | 74	65 | 70	85 | 73
	LnAvg	90 | 90	91 | 90	89 | 90	91 | 89	81 | 81	79 | 77	83 | 83	87 | 81	65 | 79	84 | 84

Table 3: Table translation experiment results with Paraphrase Score (PS), Multilingual Paraphrase Score (mPS), BERTScore (BS), Human Evaluation Score (HES), Language Average (LnAvg) and Model Average (MdlAvg). We use the "X | Y" format, where X and Y represent the Table and hypothesis translation score respectively. **Purple** and **Orange** signifies the language average score of the model selected for table and hypothesis translation respectively.

5.1 Using English Translated Test Sets

We aim to investigate the following question: *How would models trained on original English* INFOTABS *perform on English translated multilingual* XINFOTABS*?*. We trained multilingual models using the original English INFOTABS training set, and used the English translated XINFOTABS development set, and three test sets during the evaluation. According to Table 4, German has the best language-wise performance for α_1. From Table 6, German, French, and Afrikaans have the highest average scores for α_2. French and Russian have the best scores on α_3 as shown in Table 7. Arabic has the lowest average of any language across all three test sets. Here, the model trained on English INFOTABS is being used for all the languages. Since the model is the same for all languages, the variation in performance only depends on English translation across XINFOTABS languages. On α_2 and α_3 sets, this task on average performs competitively against all other baseline tasks.

5.2 Language-Specific Model Training

In this subsection, we try to answer the question: *Is it beneficial to train a language-specific model on* XINFOTABS*?* In doing so, we finetune ten distinct models, one for each language on XINFOTABS. Comparing models on this task helps comprehend the model's intrinsic multilingual capabilities for tabular reasoning. Among the language-specific models, English has the best language average in all three test sets, while Arabic has the lowest.

Additionally, there is a substantial variation in the quality of translation and model multilingualism competence. The high-resource languages often perform better since the pre-trained models have been trained on a larger amount of data from these languages. Surprisingly, §5.2 setting has lower average mBERT scores for all three splits than §5.1 setting. The benefit of training the model in English seems to surpass any loss incurred during translating test sets into English. However, this is not the case with XLM-R(XNLI). The average scores increase substantially for α_1 split in §5.2 setting compared to §5.1 setting, decrease slightly for α_2, and remain constant for α_3. The α_1 set improves due to its similar split to the train set, whereas the α_2 set slightly worsens since it includes human-annotated perturbed hypotheses with labels flipped. Lastly, the α_3 set comprises tables from zero-shot domains i.e. unseen domain tables, so it remains constant. Our exploration of models' cross-lingual transferability is provided in Appendix§ D.

5.3 Fine-tuning on Multiple Languages

Earlier findings indicate that fine-tuning multilingual models for the same task across

Train/Test Strategy	Model	en	de	fr	es	af	ru	zh	ko	hi	ar	Model. Avg.
English Translated Test (§5.1)	mBERT$_{BASE}$	-	66	64	65	66	63	63	64	64	59	64
	XLM-R$_{LARGE}$ (XNLI)	-	73	73	72	72	72	71	69	70	62	70
	Lang. Avg.	-	**70**	69	69	69	67	67	67	67	61	68
Language Specific Training (§5.2)	mBERT$_{BASE}$	67	65	65	63	62	64	63	61	63	57	63
	XLM-R$_{LARGE}$ (XNLI)	76	75	74	74	72	71	73	71	71	68	72
	Lang. Avg.	**72**	70	69	68	67	67	68	66	67	63	68
Multiple Language Finetuning Using Only English (§5.3A)	mBERT$_{BASE}$	-	64	66	64	64	64	65	63	62	62	64
	XLM-R$_{LARGE}$ (XNLI)	-	75	74	75	74	74	73	73	72	69	73
	Lang. Avg.	-	69	**70**	69	69	69	69	68	67	66	**69**
Multiple Language Finetuning Unified Model (§5.3B)	mBERT$_{BASE}$	65	64	64	64	64	63	64	62	62	59	63
	XLM-R$_{LARGE}$ (XNLI)	76	75	74	75	73	74	74	73	72	70	**74**
	Lang. Avg.	**71**	69	69	70	69	68	69	67	67	65	**69**
English Premise Multilingual Hypothesis (§5.4)	mBERT$_{BASE}$	-	63	63	64	62	61	61	59	61	60	61
	XLM-R$_{LARGE}$ (XNLI)	-	73	73	73	72	72	73	72	71	68	72
	Lang. Avg.	-	**68**	**68**	**68**	67	67	67	66	66	64	67

Table 4: Accuracy for baseline tasks on the α_1 set. **Purple** signifies the best task average accuracy, **Orange** signifies the best language average accuracy, **Cerulean** signifies the best model accuracy. XLM-R$_{LARGE}$ represent XLM-RoBERTa$_{LARGE}$ model.

languages improves performance in the target language (Phang et al., 2020; Wang et al., 2019; Pruksachatkun et al., 2020). Thus, *do models benefit from sequential fine-tuning over several* XINFOTABS *languages?* To answer it, we investigate this strategy of pre-finetuning in two ways, (a) by using English as the predominant language for pre-finetuning, and (b) by utilizing all XINFOTABS languages to train a unified model, .

A. Using English Language. We fine-tune our models on the English INFOTABS and then on XINFOTABS in each language individually. Thus, we train nine models in total, one for each multilingual language (except English). English was chosen as the pre-finetuning language due to its strong performance in the §5.2 paradigm and prior research demonstrating English's superior cross-lingual transfer capacity (Phang et al., 2020). Across all three splits, the average score improves from the §5.2 setting, demonstrating that pre-finetuning the English dataset benefits other multilingual languages. The most significant gains are shown in lower resource languages, notably Arabic, which improved by 3% for α_1, 2% for α_2, and 1% for α_3 in comparison to the §5.2 approach.

B. Unified Model Approach. We explore whether fine-tuning on other languages is beneficial, where we fine-tune a single unified model across all XINFOTABS languages' training sets and use it for making predictions on XINFOTABS test sets. We observe that the finetuning language order affects the final model performance if done sequentially. We find that training from a high to a low resource language leads to the highest average accuracy improvement. This is due to the catastrophic forgetting trait (Goodfellow et al., 2015), which encourages training on more straightforward examples first, i.e., those with better performance. Hence, we trained in the following language order: en → fr → de → es → af → ru → zh → hi→ ko → ar.

We observe that the XLM RoBERTa Large model performs the best across all baseline tasks in the α_1 set. On average, this performance is comparable to English pre-finetuning. While the accuracy of high resource languages remains constant or marginally declines compared to the §5.2 setting, there is a substantial improvement in accuracy for low resource languages, particularly Arabic, which increases by 2%. It performs similarly to English pre-finetuning. To conclude, more fine-tuning is not always beneficial for all models, but it benefits larger models like the XLM-R Large. Models improve performance for low-resource languages compared to the §5.2 setting (i.e., no pre-finetuning), but not nearly as much as that of English-based pre-finetuning.

5.4 English Premise Multilingual Hypothesis

The premise of English's multilingual hypothesis is practical, as it is frequently observed in the real world. The majority of the world's facts and information are written in English. For instance, Wikipedia has more tables in English than in any other language, and even if a page is available, it is likely that it missing an infobox. However, because people are innately bilingual, inquiries or verification queries concerning these facts could be in a language other than English. As a result,

the task of developing cross-lingual tabular NLI is critical in the real world.

To study this problem, we look at the following question: *How effective are models with premise and hypothesis stated in distinct languages?* To answer this, we train the models using the original INFOTABS premise tables in the English language and multilingual hypotheses in XINFOTABS, i.e., nine languages. We note that XLM-R Large (XNLI) has the highest accuracy for the α_1 set. On average, the high-resource languages German, French, and Spanish perform favorably across models, whereas Arabic underperforms. Both models have shallow scores in German for the α_2 set, which defy earlier observations. This might be because the adversarial modifications in the α_2 hypothesis might not be reflected in the German translation. XLM-R Large has the highest accuracy on this set, with French and Spanish being the most accurate languages. The models for the α_3 validation set demonstrate that language average accuracy is nearly proportional to the size of translation resources. However, the scores are marginally lower on average for the α_2 set.

Surprisingly, models perform worse on average than with §5.2 setting on the α_1 and α_2 sets while performing similarly on the α_3 set. Except for α_2 on German, the average language accuracy changes are directly proportional to the language resource, implying that the constraint could be translation quality; left for future study. Refer Appendix §E for robustness and consistency analysis.

6 Discussion and Analysis

Extraction vs. Translation. One straightforward idea for constructing the multilingual tabular NLI dataset is to extract multilingual tables from Wikipedia in the considered languages. However, this strategy fails in practice for several reasons. For starters, not all articles are multilingual. For example, only 750 of the 2540 tables were from articles available in Hindi. The existence of the same title articles across several languages does not indicate that the tables are identical. Only 500 of the 750 tables with articles in Hindi had infoboxes, and most of these tables were considerably different from the English tables. The tables had different numbers of keys and different value information.

Human Verification vs. Human Translation. We selected machine translation with human verification over hiring expert translators for several reasons: (a) Hiring bilingual, skilled translators in multiple languages is expensive and challenging, (b) Human verification is a more straightforward classification task based on semantic similarity; it is also less erroneous compared to translation, (c) By selecting an appropriate verification sample size, we may further minimize the time and effort required for human inspection, (d) A competent translation system has no effect on the classification labels used in inference. As a result, the loss of the semantic connection between the table and the hypothesis is not a significant issue (K et al., 2021), and (e) Minor translation errors have no effect on the downstream NLI task label as long as the semantic meaning of the translation is retained (Conneau et al., 2018; K et al., 2021; Cohn-Gordon and Goodman, 2019; Carl, 2000).

Usage and Future Direction. The dataset can be used to test benchmarks, multilingual models, and methods for tabular NLI. In addition to language invariance, robustness, and multilingual fact verification, it may well be utilized for reasoning tasks like multilingual question answering (Demszky et al., 2018). The baselines can also be beneficial to understand models' cross-lingual transferability.

Our current table structure does not generate natural language sentences and hence does not optimize the capabilities of a machine translation model. The representation of tables can be enhanced further by adding Better Paragraph Representation (BPR) from Neeraja et al. (2021). Additionally, NER handling may be enhanced by inserting a predetermined template name into the sentence post-translation, i.e. extracting a named entity from the original sentence, replacing it with a fixed template entity, and then replacing the named entity with the template post-translation. Multiple experiments, however, would be necessary to identify suitable template entities for replacement, and hence this is left as future work. Another approach is the extraction of keys and values from multilingual Wikipedia pages is also a challenging task and left as future work. Finally, human intervention can enhance the translation quality by either direct human translation or fine-grained post-translation verification and correction.

7 Related Work

Tabular Reasoning. Recent studies investigate various NLP tasks on semi-structured tabular data, including tabular NLI and fact verification (Chen et al., 2019; Gupta et al., 2020; Zhang and Balog, 2019), tabular probing (Gupta et al., 2021), various question answering and semantic parsing tasks (Pasupat and Liang, 2015; Krishnamurthy et al., 2017; Abbas et al., 2016; Sun et al., 2016; Chen et al., 2020b; Lin et al., 2020; Zayats et al., 2021; Oguz et al., 2020; Chen et al., 2021, *inter alia*), and table-to-text generation (e.g., Parikh et al., 2020; Nan et al., 2021; Yoran et al., 2021; Chen et al., 2020a). Several strategies for representing Wikipedia relational tables were recently proposed, such as TAPAS (Herzig et al., 2020), TaBERT (Yin et al., 2020), TabStruc (Zhang et al., 2020), TABBIE (Iida et al., 2021), TabGCN (Pramanick and Bhattacharya, 2021) and RCI (Glass et al., 2021). Yu et al. (2018, 2021); Eisenschlos et al. (2020) and Neeraja et al. (2021) study pre-training for improving tabular inference.

Multilingual Datasets and Models. Given the need for greater inclusivity towards linguistic diversity in NLP applications, various multilingual versions of datasets have been created for text classification (Conneau et al., 2018; Yang et al., 2019; Ponti et al., 2020), question answering (Lewis et al., 2020; Clark et al., 2020; Artetxe et al., 2020) and structure prediction (Rahimi et al., 2019; Nivre et al., 2016). Following the introduction of datasets, multilingual leaderboards like XTREME leaderboard (Hu et al., 2020), the XGLUE leaderboard (Liang et al., 2020) and the XTREME-R leaderboard (Ruder et al., 2021) have been created to test models' cross-lingual transfer and language understanding.

Multilingual models can be broadly classified into two variants: (a) Natural Language Understanding (NLU) models like mBERT (Devlin et al., 2019), XLM (Conneau and Lample, 2019), XLM-R (Conneau et al., 2020), XLM-E (Chi et al., 2021), RemBERT (Chung et al., 2021), and (b) Natural Language Generation (NLG) models like mT5 (Xue et al., 2021), mBART (Liu et al., 2020), M2M100 (Fan et al., 2021). NLU models have been used in multilingual language understanding tasks like sentiment analysis, semantic similarity and natural language inference while NLG models are used in generation tasks like question-answering and machine translation.

Machine Translation. Modern machine translation models involve having an encoder-decoder generator model trained on either bilingual (Tran et al., 2021) or a multilingual parallel corpus with monolingual pre-training e.g. mBART (Liu et al., 2020) and M2M100 (Fan et al., 2021). These models have been shown to work very well even for low-resource languages due to cross-language transfer properties. Recently auxiliary pertaining for machine translation models have garnered attention, with a focus on autonomous quality estimation metrics (Specia et al., 2018; Fonseca et al., 2019; Specia et al., 2020). As such, automatic scores like the BERTScore (Zhang et al., 2019), Bleurt (Sellam et al., 2020) and COMET Score (Rei et al., 2020) have high human evaluation correlation, are increasingly used to assess NLG tasks.

8 Conclusion

We built the first multilingual tabular NLI dataset, namely XINFOTABS, by expanding the INFOTABS dataset with ten different languages. This is accomplished by our novel machine translation approach for tables, which yields remarkable results in practice. We thoroughly evaluated our translation quality to demonstrate that the dataset meets the acceptable standard. We further examined the performance of multiple multilingual models on three validation sets of varying difficulty, with methods ranging from the basic translation-based technique to more complicated language-specific and intermediate task finetuning. Our results demonstrate that, despite the models' success, this dataset remains a difficult challenge for multilingual inference. Lastly, we gave a thorough error analysis of the models to comprehend their cross-linguistic transferability, robustness to language change, and coherence with reasoning.

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. 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

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.

Dimo Angelov. 2020. Top2vec: Distributed representations of topics.

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.

Shima Asaadi, Saif Mohammad, and Svetlana Kiritchenko. 2019. Big BiRD: A large, fine-grained, bigram relatedness dataset for examining semantic composition. 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 505–516, Minneapolis, Minnesota. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *CoRR*, abs/2004.05150.

Michael Carl. 2000. On the meaning preservation capacities in machine translation.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

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. 2019. 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 Yang Wang. 2020b. HybridQA: A dataset of multi-hop question answering over tabular and textual data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1026–1036, 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*.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Reuben Cohn-Gordon and Noah D. Goodman. 2019. Lost in machine translation: A method to reduce meaning loss. *CoRR*, abs/1902.09514.

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, 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.

Dorottya Demszky, Kelvin Guu, and Percy Liang. 2018. Transforming question answering datasets into natural language inference datasets. *arXiv preprint arXiv:1809.02922*.

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*.

Julian Eisenschlos, Syrine Krichene, and Thomas Müller. 2020. Understanding tables with intermediate pre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 281–296, Online. 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, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020a. Beyond english-centric multilingual machine translation. *arXiv preprint*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020b. Beyond english-centric multilingual machine translation.

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.

Erick Fonseca, Lisa Yankovskaya, André F. T. Martins, Mark Fishel, and Christian Federmann. 2019. Findings of the WMT 2019 shared tasks on quality estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 1–10, Florence, Italy. Association for Computational Linguistics.

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.

Ian J. Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. 2015. An empirical investigation of catastrophic forgetting in gradient-based neural networks.

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.

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.

Ridong Jiang, Rafael E. Banchs, and Haizhou Li. 2016. Evaluating and combining name entity recognition systems. In *Proceedings of the Sixth Named Entity Workshop*, pages 21–27, Berlin, Germany. Association for Computational Linguistics.

Xin Jin and Jiawei Han. 2010. *K-Means Clustering*, pages 563–564. Springer US, Boston, MA.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. 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. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 86–95, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Philipp Koehn and Christof Monz. 2006. Manual and automatic evaluation of machine translation between European languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121, New York City. Association for Computational Linguistics.

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.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Xiaoyu Li and Francesco Orabona. 2019. On the convergence of stochastic gradient descent with adaptive stepsizes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 983–992. PMLR.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.

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. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

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*.

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.

Jason Phang, Iacer Calixto, Phu Mon Htut, Yada Pruksachatkun, Haokun Liu, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. English intermediate-task training improves zero-shot cross-lingual transfer too. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 557–575, Suzhou, China. Association for Computational Linguistics.

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.

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. Intermediate-task transfer learning with pretrained language models: When and why does it work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5231–5247, Online. Association for Computational Linguistics.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. 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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

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.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Erick Fonseca, Vishrav Chaudhary, Francisco Guzmán, and André F. T. Martins. 2020. Findings of the WMT 2020 shared task on quality estimation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 743–764, Online. Association for Computational Linguistics.

Lucia Specia, Frédéric Blain, Varvara Logacheva, Ramón F. Astudillo, and André F. T. Martins. 2018. Findings of the WMT 2018 shared task on quality estimation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 689–709, Belgium, Brussels. Association for Computational Linguistics.

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.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning.

Chau Tran, Shruti Bhosale, James Cross, Philipp Koehn, Sergey Edunov, and Angela Fan. 2021. Facebook AI's WMT21 news translation task submission. In *Proceedings of the Sixth Conference on Machine Translation*, pages 205–215, Online. Association for Computational Linguistics.

Alex Wang, Jan Hula, Patrick Xia, Raghavendra Pappagari, R. Thomas McCoy, Roma Patel, Najoung Kim, Ian Tenney, Yinghui Huang, Katherin Yu, Shuning Jin, Berlin Chen, Benjamin Van Durme, Edouard Grave, Ellie Pavlick, and Samuel R. Bowman. 2019. Can you tell me how to get past sesame street? sentence-level pretraining beyond language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4465–4476, Florence, Italy. Association for Computational Linguistics.

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

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

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*.

Tao Yu, Chien-Sheng Wu, Xi Victoria Lin, Bailin Wang, Yi Chern Tan, Xinyi Yang, Dragomir Radev, Richard Socher, and Caiming Xiong. 2021. Grappa: Grammar-augmented pre-training for table semantic parsing. *International Conference of Learning Representation*.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir Radev. 2018. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3911–3921, Brussels, Belgium. Association for Computational Linguistics.

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.

Shuo Zhang and Krisztian Balog. 2019. Auto-completion for data cells in relational tables. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, pages 761–770, New York, NY, USA. ACM.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

A Human Annotation Guidelines

Annotators Details. We employed five undergraduate students proficient in English as human evaluation annotators. They were presented with an instruction set with sample examples and annotations before the actual work. We paid the equivalent of 10 cents for every labeled example. The study's authors reviewed random annotations to confirm their quality.

Annotation Guidelines. We refer to the work by (Koehn and Monz, 2006) while setting up our annotation task and instruction guidelines. We gathered 500 table-sentence pairs representing original (en) and back-translated (en) texts per model-language into several Google spreadsheets. We had a total of 108 sheets (4 models, 9 languages, 3 Modes (table-keys, table-values, and hypothesis) and hence 54000 annotation instances. Each sheet was assigned to a single annotator, who was required to adhere to the semantic similarity task requirements, which are outlined below:

1. The Semantic Similarity task requires the annotator to classify each sentence-pair as conveying the same meaning (label 1) or conveying different meaning (label 0) than each other.
2. In case their exists a difference of syntax including spelling mistakes, punctuation error or missing special characters, the annotators was asked to ignore these as long as the sentence meaning is understandable (label 1). In case proper nouns were misspelled, the annotator must judge the spellings as phonetically similar (label 1) or not (otherwise label 0).
3. The annotators were asked to be lenient on the grammar, allowing for active-passive changes and tense change, if the sentences convey close to the same meaning i.e. (label 1).
4. In case acronyms or abbreviations were present in the sentences, the annotators were asked to mark them as same (label 1) if the sentences had proper expansion/contractions.

Code	Language	Language Family	Script Type	# of Speakers
en	English	Germanic	Latin	1.452 Billion
de	German	Germanic	Latin	134.6 Million
fr	French	Romance	Latin	274.1 Million
es	Spanish	Romance	Latin	548.3 Million
af	Afrikaans	Germanic	Latin	17.5 Million
ru	Russian	Balto-Slavik	Cryllic	258.2 Million
zh	Chinese	Sinitic	Hanzi	1.118 Billion
ko	Korean	Koreanic	Hangul	81.7 Million
hi	Hindi	Indo-Aryan	North-Indic	602.2 Million
ar	Arabic	Semitic	Arabic	274.0 Million

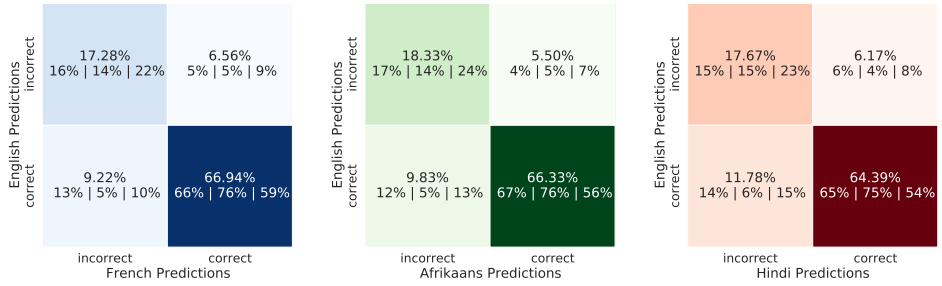
Table 5: Details regarding languages provided in the XINFOTABS, from English to Arabic in order of open-source translation resources, refer to OPUS

Train/Test Strategy	Model	en	de	fr	es	af	ru	zh	ko	hi	ar	Model. Avg
English Translated Test (§5.1)	$mBERT_{BASE}$	-	54	53	52	54	52	52	53	52	50	53
	$XLM\text{-}R_{LARGE}$ (XNLI)	-	67	66	64	65	65	63	63	63	58	64
	Lang. Avg.	-	**60**	**60**	58	**60**	59	58	58	58	54	**59**
Language Specific Training (§5.2)	$mBERT_{BASE}$	54	54	52	53	50	52	52	51	50	48	52
	$XLM\text{-}R_{LARGE}$ (XNLI)	68	66	64	66	63	64	64	64	62	57	64
	Lang. Avg.	**61**	60	58	60	57	58	58	58	56	53	58
Multiple Language Finetuning Using Only English (§5.3A)	$mBERT_{BASE}$	-	53	54	51	53	53	53	52	51	50	52
	$XLM\text{-}R_{LARGE}$ (XNLI)	-	66	67	66	66	65	65	65	64	61	**65**
	Lang. Avg.	-	59	**60**	58	59	59	59	59	58	55	**59**
Multiple Language Finetuning Unified Model (§5.3B)	$mBERT_{BASE}$	53	51	53	53	52	51	53	50	50	49	52
	$XLM\text{-}R_{LARGE}$ (XNLI)	66	64	64	63	64	64	64	63	63	60	64
	Lang. Avg.	**60**	58	59	58	58	58	58	56	57	54	58
English Premise Multilingual Hypothesis (§5.4)	$mBERT_{BASE}$	-	49	53	53	51	49	49	50	47	50	50
	$XLM\text{-}R_{LARGE}$ (XNLI)	-	63	65	65	64	65	65	63	63	61	64
	Lang. Avg.	-	56	**59**	**59**	57	57	57	57	55	55	57

Table 6: Accuracy for baseline tasks on the α_2 set. **Purple** signifies the best task average accuracy, **Orange** signifies the best language average accuracy, **Cerulean** signifies the best model accuracy. $XLM\text{-}R_{LARGE}$ represent $XLM\text{-}RoBERTa_{LARGE}$ model.

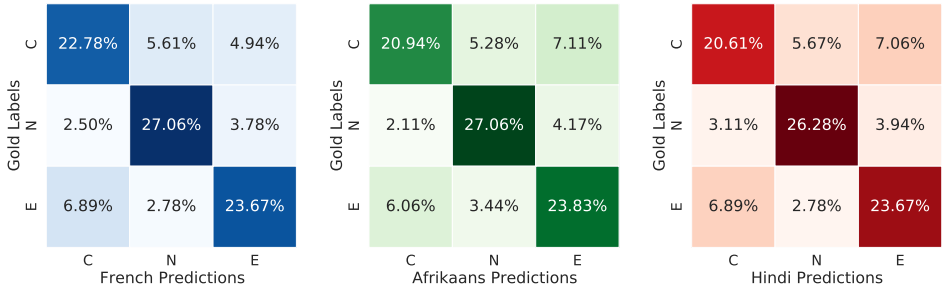
Train/Test Strategy	Model	en	de	fr	es	af	ru	zh	ko	hi	ar	Model. Avg.
English Translated Test (§5.1)	$mBERT_{BASE}$	-	52	53	52	53	53	52	52	52	50	52
	$XLM\text{-}R_{LARGE}$ (XNLI)	-	65	65	64	63	64	62	62	61	57	**63**
	Lang avg	-	58	**59**	58	58	**59**	57	57	57	53	**58**
Language Specific Training (§5.2)	$mBERT_{BASE}$	52	50	52	53	50	50	51	48	49	49	50
	$XLM\text{-}R_{LARGE}$ (XNLI)	67	65	62	64	62	62	63	60	62	57	62
	Lang avg	**60**	58	57	58	56	56	57	54	56	53	56
Multiple Language Finetuning Using Only English (§5.3A)	$mBERT_{BASE}$	-	52	50	52	52	51	51	49	49	48	50
	$XLM\text{-}R_{LARGE}$ (XNLI)	-	65	64	65	62	64	60	63	62	63	**63**
	Lang avg	-	**59**	57	58	57	57	56	56	56	54	57
Multiple Language Finetuning Unified Model (§5.3B)	$mBERT_{BASE}$	53	50	51	53	50	50	51	47	50	49	50
	$XLM\text{-}R_{LARGE}$ (XNLI)	66	64	64	64	63	64	63	62	63	60	**63**
	Lang avg	60	57	57	58	56	57	57	55	56	54	57
English Premise Multilingual Hypothesis (§5.4)	$mBERT_{BASE}$	-	51	50	51	50	50	47	45	48	48	49
	$XLM\text{-}R_{LARGE}$ (XNLI)	-	63	63	64	62	62	62	60	61	60	62
	Lang avg	-	**57**	**57**	**57**	56	56	55	54	55	54	56

Table 7: Accuracy for baseline tasks on the α_3 set. **Purple** signifies the best task average accuracy, **Orange** signifies the best language average accuracy, **Cerulean** signifies the best model accuracy. $XLM\text{-}R_{LARGE}$ represent $XLM\text{-}RoBERTa_{LARGE}$ model.



(a) (b) (c)

Figure 2: Predictions of XLM-RoBERTa for English vs (a) French, (b) Afrikaans, (c) Hindi. The percentage on top in each block represents the average across all three labels with each label percentage given below it in the order of ENTAILMENT, NEUTRAL and CONTRADICTION. (cf. Appendix §E)



(a) (b) (c)

Figure 3: Confusion Matrix: Gold Labels vs predictions of XLM-R for (a) French, (b) Afrikaans, (c) Hindi

Categories	ENTAILMENT					NEUTRAL					CONTRADICTION				
	En	Fr	Af	Hi	Avg.	En	Fr	Af	Hi	Avg.	En	Fr	Af	Hi	Avg.
Person	79	71	75	73	74	82	81	78	81	81	**59**	67	54	56	59
Musician	**88**	77	78	76	80	87	87	91	82	87	70	69	60	69	67
Movie	**70**	**63**	**57**	63	**63**	85	93	85	87	88	**81**	**76**	**78**	65	**75**
Album	76	76	81	**62**	74	**95**	90	86	**90**	**90**	76	**76**	67	62	70
City	73	58	60	67	65	71	69	**65**	**63**	**67**	67	**54**	**50**	**52**	**56**
Country	74	**61**	65	63	66	74	70	76	76	74	74	72	76	69	73
Painting	83	79	75	67	76	83	**96**	**92**	83	89	71	71	71	71	71
Animal	79	75	79	79	78	75	58	83	67	71	71	75	67	58	68
Food&Drink	**88**	83	75	**88**	**83**	83	79	71	79	78	67	63	58	54	60
Organization	83	**100**	**83**	50	79	**67**	**67**	67	67	**67**	67	67	67	**83**	71
Other	75	73	67	73	72	73	84	84	75	79	76	68	71	62	69
Avg.	79	74	72	69	74	80	79	80	77	79	71	69	65	64	67

Table 8: Category wise accuracy scores of XLM-R (large) for four languages: namely English (En), French (Fr), Afrikaans (Af) and Hindi (Hi). **Orange** denotes the least score in the column and **Purple** denotes the highest score in the column.

Reasoning type	ENTAILMENT					NEUTRAL					CONTRADICTION				
	H.En	En	Fr	Af	Ko	H.En	En	Fr	Af	Ko	H.En	En	Fr	Af	Ko
Coref	8	6	6	6	4	22	19	19	20	19	13	10	9	7	8
Entity Type	6	5	5	5	5	8	6	6	6	6	6	6	6	4	5
KCS	31	21	19	17	22	21	20	17	19	18	24	18	17	17	20
Lexical Reasoning	5	4	4	4	3	3	2	2	2	1	4	1	1	1	1
Multirow	20	14	11	11	11	16	13	12	13	11	17	15	14	10	13
Named Entity	2	0	0	0	1	2	1	1	1	2	1	1	1	1	1
Negation	0	0	0	0	0	0	0	0	0	0	6	5	5	4	5
Numerical	11	10	7	8	8	3	3	2	3	2	7	5	6	4	4
Quantification	4	2	2	2	2	13	10	10	12	10	6	2	1	2	3
Simple Lookup	3	2	1	2	2	0	0	0	0	0	1	0	1	0	0
Subjective/OOT	6	3	4	4	3	41	37	35	36	37	6	3	4	2	3
Temporal	19	16	12	13	14	11	6	6	6	5	25	18	20	15	19

Table 9: Reasoning wise number of correct predictions of XLM-R (large) for four languages: namely English (En), French (Fr), Afrikaans (Af) and Hindi (Hi) along with human scores for the english dataset

5. In presence of numbers or dates, the annotators were asked to be extremely strict and label even slightly differing dates or numbers like (XXXI v.s. 30) as completely different (label 0).
6. In case of any further ambiguity, the judgement was left to the annotators human far-sight as long as the adhere to the task definition.

We estimated the accuracy of human verification for every models and languages by averaging the annotator labels.

B Multilingual Models Hyperparameters

The XLM-R_{LARGE} (XNLI) model was taken from HuggingFace[9] models and finetuned using PyTorch Framework[10] on Google Colaboratory[11] which offer a single P100 GPU. We utilized accuracy as our metric of choice, same as INFOTABS. We used Adagrad (Li and Orabona, 2019) as our optimizer with a learning rate of $1 * 10^{-4}$. We ran our finetuning script for ten epochs with a validation interval of 1 epoch, and early stopping callback enabled with the patience of 2. Given the large model size, we had to use a batch size of 4.

The mBERT$_{BASE}$ (cased) model was trained on TPUv2 8 cores using the PyTorch Lightning[12] Framework. AdamW (Loshchilov and Hutter, 2017) was our choice of optimizer with learning rate $5 * 10^{-6}$. We ran our finetuning script for ten epochs with a validation interval of 0.5 epochs, and early stopping callback enabled with the patience of 3. Given the model's small size, we used a batch size of 64 (8 per TPU core).

C Adversarial Sets (α_2 and α_3) Performance

Tables 6 and 7 show the results for all baseline tasks on the Adversarial Validation Sets α_2 and α_3.

D Evaluating Cross-Lingual Transfer

We are also interested in knowing whether training in one language can help transfer knowledge across other languages or not. We answer the question: *What are models of cross-lingual transfer performance?*. Since we have separate models trained on languages from our dataset available, we tested them on all other languages other than the training language to study cross-lingual transfer.

The TrLangAvg scores (Training Language Average) from 10 show how models trained on XINFOTABS for one language perform on other languages for α_1, α_2 and α_3 sets respectively. XLM-R (XNLI) outperforms mBERT across all tasks. English has the best cross-lingual transferability on mBERT, whereas Spanish has the best cross-lingual transferability on XLM-R(XNLI) for the α_1 set. On mBERT, German has the best cross-lingual transferability for the α_2 dataset. On XLM-R (XNLI), German and Spanish have the best cross-lingual transferability. On mBERT, English has the best cross-lingual transferability for the α_3 dataset. On XLM-R (XNLI), English and Spanish have the best cross-lingual transferability. Furthermore, the EvLangAvg score (Evaluation Language Average) score was comparable for all languages except approximately 4% lower for Arabic ('ar') language with XLM-R(XNLI) model on all three test sets.

Overall, we observe that finetuning models on high resource languages improve their cross-lingual transfer capacity considerably more than finetuning models on low resource languages.

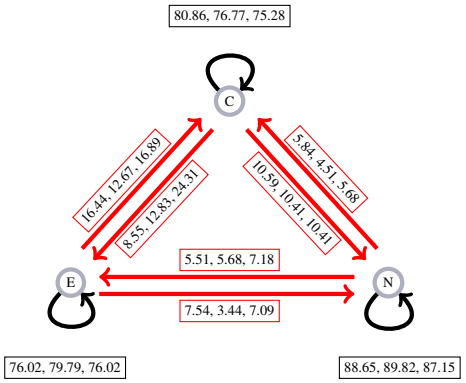


Figure 4: Consistency graph for XLM-R (large) predictions of English vs (a) French (b) Afrikaans (c) Hindi in that order respectively.

E Robustness and Consistency

In this part, we examine the findings for several languages and delve a little more into the key disparities in performance across them. We compare the results of the experiments for §5.2 setting for α_1 set of best-performing language (en) with three languages - (a) A high resource language (fr), (b) A mid resource language (af) and c) A low resource language (hi). We compute four numbers for each of the languages (l) (where l is (fr), (af), or (hi)) and (en) - the proportion of instances when (a) both are right, (b) both are erroneous (c) correct (en) but incorrect (l), and

[9] huggingface.co [10] pytorch.org [11] Google Colaboratory
[12] PyTorch Lightning

(d) correct (l) but incorrect (en). We compute this number overall as well independently for each of the inference labels, as shown in Figure 2.

We note that the majority of instances were correctly categorized in both English and all three other languages. This is followed by the number of instances in which English and all other languages categorised examples inaccurately. Additionally, we notice a greater proportion of samples that are correctly identified by English but wrongly classified by all other languages, as opposed to the contrary. Furthermore, the label NEUTRAL has the highest proportion of correctly classified examples across all languages, whereas the label CONTRADICTION has the lowest.

In Figure 3, we notice that the CONTRADICTION gets confused a lot with ENTAILMENT label across all the languages. The difference between the accuracy for the CONTRADICTION label of French vs Afrikaans and Hindi can entirely be attributed to this sort of confusion. Furthermore, ENTAILMENT gets quite confused with CONTRADICTION.

In Figure 4, we see the greatest language inconsistency with ENTAILMENT label going towards CONTRADICTION across all the languages, though this inconsistency is least in Afrikaans. The inconsistency for CONTRADICTION label being predicted as ENTAILMENT is increasing across resource size of languages from French having the least to Hindi having the highest. Otherwise, the inconsistency across languages is rather low, showing that the XLM-R_{LARGE} model is quite consistent across languages.

In Table 8, we can observe that our model on average performs worst for all ENTAILMENT belonging to Movie category, NEUTRAL and CONTRADICTION belonging to City category. In general, our model performs the worst for all hypothesis belonging to the City category possibly because of the involvement of larger table sizes on average and highly numeric and specific hypothesis statements as compared to the rest of the categories. Our models perform extremely well on all ENTAILMENT in FoodDrink category because of their smaller table size on average and hypothesis requiring no external knowledge to confirm as compared to CONTRADICTION. For ENTAILMENT our model performs remarkably well on Organization category for French, getting all the hypothesis labels correct. While for NEUTRAL, it performs well for Paintings in French language. Lastly, it performs marginally well for CONTRADICTION on Hindi for Organization as compared to the highest performing category for CONTRADICTION in English i.e. the Movie category. All language averages perform in the order of their language resource which is expected from Table 4.

Table 9 depicts a subset of the validation set which has been labeled based on different reasoning mechanisms that the model must employ to categorize the hypothesis correctly. We found the reasoning accuracy scores for 4 languages along with human evaluation score for comparison. Upon observation, we can see that regardless of language, human scores are better than the model we utilize. The variation in language is mostly minimal, but on average our model performs best for English. We notice that for some reasoning types, like Negation and Simple Look-up, humans and the model get no hypothesis right, showing the toughness of the problem. For Numerical based reasoning as well as Coref type reasoning, our model comes very close to human score evaluation. However, overall we are still far from human level performance at TNLI and much scope remains to betterment of models on this task.

Test-Split	Model	TrLang	en	de	fr	es	af	ru	zh	ar	ko	hi	TrLangAvg
α_1	mBERT$_{BASE}$	en	**67**	64	63	62	61	61	60	56	58	58	**61**
		de	63	**65**	61	62	60	59	57	56	56	57	60
		fr	64	62	**65**	62	61	59	59	55	53	57	60
		es	62	62	**63**	**63**	61	60	60	**57**	57	58	60
		af	**62**	61	61	60	**62**	59	57	55	55	55	59
		ru	63	61	61	60	59	**64**	59	56	55	55	59
		zh	55	56	58	56	59	57	**63**	55	57	58	57
		ar	57	**58**	**58**	57	**58**	**58**	57	**57**	53	57	57
		ko	58	59	58	57	57	56	58	55	**61**	57	58
		hi	59	58	59	58	57	58	58	56	54	**63**	58
	EvLangAvg		**61**	**61**	**61**	60	60	59	59	56	56	58	59
	XLM-R (XNLI)	en	**76**	73	71	73	71	**71**	71	63	70	69	71
		de	74	**75**	**74**	72	71	70	69	63	**71**	68	71
		fr	73	**74**	**74**	72	**72**	70	71	64	70	70	71
		es	**74**	73	**74**	**74**	**72**	**71**	72	65	**71**	69	**72**
		af	**72**	**72**	71	71	**72**	70	70	63	70	68	70
		ru	**73**	**73**	72	71	71	**71**	71	64	70	67	70
		zh	72	72	70	71	70	69	**73**	64	70	69	70
		ar	**71**	**71**	70	70	69	70	**71**	**68**	70	68	70
		ko	**72**	71	**72**	71	70	69	71	64	**71**	69	70
		hi	**73**	**73**	71	72	70	70	70	64	69	**71**	70
	EvLangAvg		**73**	**73**	72	72	71	70	71	64	70	69	70
α_2	mBERT$_{BASE}$	en	**54**	53	**53**	**53**	51	**52**	50	**49**	50	47	51
		de	**54**	**54**	**53**	**53**	**52**	**52**	50	**49**	50	48	**52**
		fr	52	51	52	**53**	50	50	48	**49**	**51**	47	50
		es	52	50	50	**53**	47	51	48	**49**	46	46	49
		af	49	**50**	**50**	49	**50**	**50**	47	48	48	46	49
		ru	51	50	51	51	51	**52**	49	**49**	49	49	50
		zh	49	48	49	48	49	49	**52**	47	48	48	49
		ar	**49**	48	**49**	48	47	48	47	48	47	47	48
		ko	49	49	50	48	48	47	50	47	**51**	49	49
		hi	48	47	47	48	48	49	48	46	48	**50**	48
	EvLangAvg		**51**	50	50	50	49	50	49	48	49	48	50
	XLM-R (XNLI)	en	**68**	65	64	64	**64**	63	62	**58**	63	59	63
		de	**67**	**66**	**66**	65	**64**	63	62	57	**64**	61	**64**
		fr	**67**	64	64	65	62	60	60	**58**	62	60	62
		es	**67**	**66**	65	**66**	63	**64**	62	57	**64**	61	64
		af	**66**	64	64	64	63	62	63	57	62	59	62
		ru	**66**	64	64	63	62	**64**	62	57	61	60	62
		zh	**67**	65	65	64	63	**64**	**64**	**58**	**64**	61	62
		ar	**64**	61	62	61	60	60	60	57	60	58	60
		ko	**65**	63	63	63	61	62	62	57	**64**	59	62
		hi	**67**	64	65	65	63	**64**	62	**58**	60	**62**	63
	EvLangAvg		**66**	64	64	64	63	63	62	57	62	60	63
α_3	mBERT$_{BASE}$	en	**52**	**52**	51	**53**	49	**50**	49	47	46	47	**50**
		de	50	50	**51**	50	**51**	48	48	44	46	48	49
		fr	**52**	**52**	**52**	**53**	50	**50**	49	46	44	47	**50**
		es	50	50	51	**53**	48	48	46	46	46	46	**50**
		af	50	50	50	**51**	50	49	47	47	45	48	49
		ru	**50**	48	49	**50**	49	**50**	47	45	45	46	48
		zh	49	49	50	50	49	**50**	**51**	46	**48**	49	49
		ar	**49**	**49**	**49**	**49**	48	**49**	48	**49**	47	48	48
		ko	47	46	47	47	44	45	45	43	**48**	**48**	46
		hi	**50**	49	49	49	48	46	48	46	47	**50**	48
	EvLangAvg		**50**	49	**50**	**50**	49	48	48	46	46	48	49
	XLM-R (XNLI)	en	**67**	**65**	61	**64**	62	**64**	**63**	58	**65**	**62**	**63**
		de	**65**	**65**	**63**	61	**63**	63	61	56	61	60	62
		fr	**66**	64	62	63	62	61	61	56	60	**62**	62
		es	**66**	**65**	**63**	**64**	**63**	63	62	**59**	61	**62**	**63**
		af	**65**	64	61	62	62	60	61	56	60	59	61
		ru	**65**	63	61	62	62	62	61	56	60	**62**	61
		zh	**65**	64	62	63	62	62	**63**	57	62	60	62
		ar	**63**	62	62	61	61	60	60	57	60	60	61
		ko	**64**	62	61	62	60	63	61	56	60	**62**	61
		hi	**64**	63	62	63	61	61	60	58	60	**62**	61
	EvLangAvg		**65**	64	62	63	62	62	61	57	61	61	62

Table 10: Evaluation of cross lingual transfer abilities of models on α_1, α_2, and α_3 evaluation set. TrLang refers to the language the model has been finetuned on and EvLang refers to the language the model has been evaluated on. **Purple**, **Orange** and **Cerulean** represent the highest score in the row, column and both together respectively.