XInfoTabS: Evaluating Multilingual Tabular Natural Language Inference

https://xinfotabs.github.io/

Bhavnick Singh Minhas[1*], Anant Shankhdhar[1*], Vivek Gupta[2*], Divyanshu Aggarwal[3], Shuo Zhang[4]

[1]Indian Institute of Technology Guwahati; [2]University of Utah; [3]Delhi Technological University; [4]Bloomberg LP.

[2]on academic job market
[2]Bloomberg Ph.D. Fellow

TABULAR INFERENCE

- The **tabular natural language inference** problem is similar to standard NLI
- But here, the **premises are tabular data**
- Task: to decide whether given hypothesis is **true** (**entailment**), **false** (**contradiction**) or **undetermined** (**neutral**) given a premise table

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.

H1: The modern form of boxing started in the late 1900's. → **Contradiction**

Check out INFOTABS (Gupta et al., 2020)
https://infotabs.github.io

MOTIVATION

To date, no work has been done in the field of multilingual tabular inference.

All existing works are done entirely in English language.

MOTIVATION

To date, no work has been done in the field of multilingual tabular inference.
All existing works are done entirely in English language.

Questions

- How can we create a dataset that can be leveraged to train and evaluate multilingual models for the task?

MOTIVATION

To date, no work has been done in the field of multilingual tabular inference. All existing works are done entirely in English language.

Questions

- How can we create a dataset that can be leveraged to train and evaluate multilingual models for the task?

- How well can multilingual models (for example, XLM-RoBERTa and mBERT) reason about multilingual tabular inference?

For Tabular Natural Language Inference, it's **not enough** to **only focus in English.**

Progress in Tabular Inference must be **made across the board,** and **this includes all languages**.

OUR CONTRIBUTIONS

- XINFOTABS, is a multilingual dataset for semi-structured tabular inference which contains instances in ten diverse languages.

- To create XINFOTABS, we leverage cutting-edge machine translation models which provide high-quality translations of semi-structured tabular data.

- We access reasoning ability of state-of-the-art multilingual models trained with varying strategies over XINFOTABS.

EXAMPLE FROM XINFOTABS DATASET

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.

English Table

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.

French Table (en → fr)

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

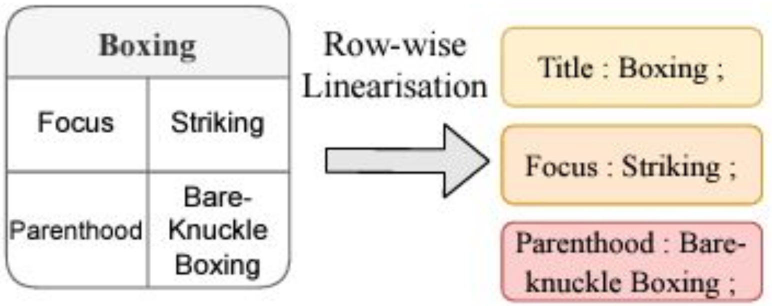
CHALLENGES

- Tabular data that is semi-structured contains succinct, non-sentential implicit information. As a result, translation is difficult.

- Translation quality is not universal. Quality varies with multilingual models (e..g mBART, M2M, MarianMT), 11 languages and data format (i.e. table, hypothesis)

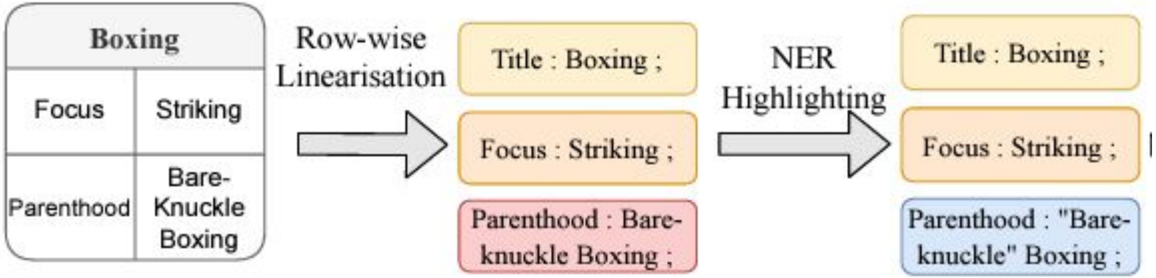
- How to measure the translations quality using automatic metric and human rating especially for tabular semi-structured data.

TRANSLATING TABLES



Since the table is a list of key value pairs, we first **linearize every row** so that both the **key and value can be translated jointly**.

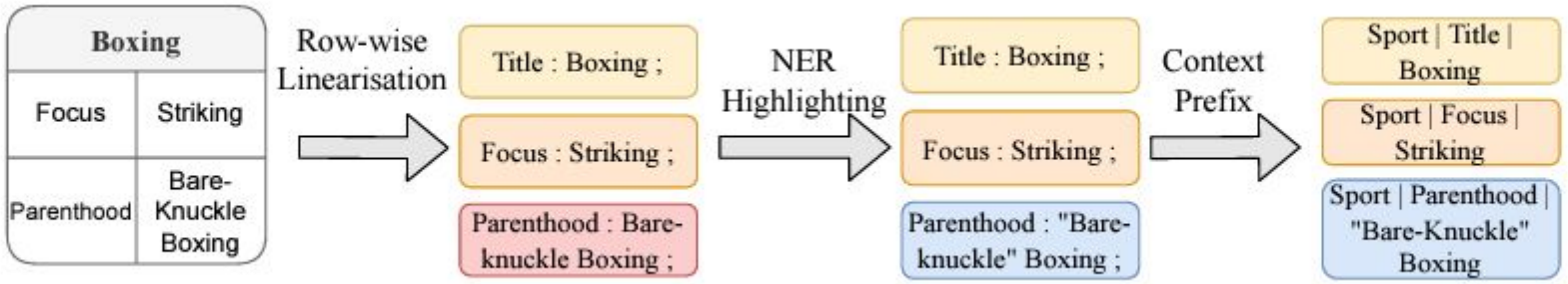
TRANSLATING TABLES



Instead of transliterating, open source machine translation models **translate named entities**.

Therefore, we **highlight** (“ ”) the **named entities** and **numbers** in the linearized rows for transliteration.

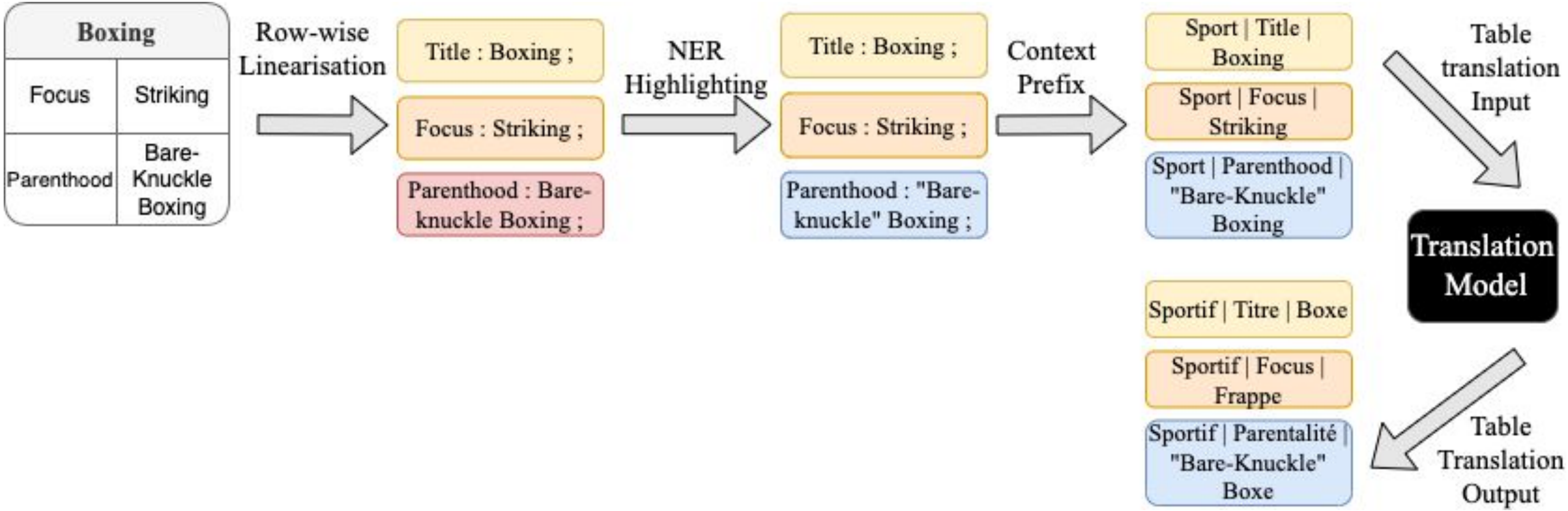
TRANSLATING TABLES



Add **additional context** in term of **Category Information**.

The category, key and value are **separated by a delimiter** (|).

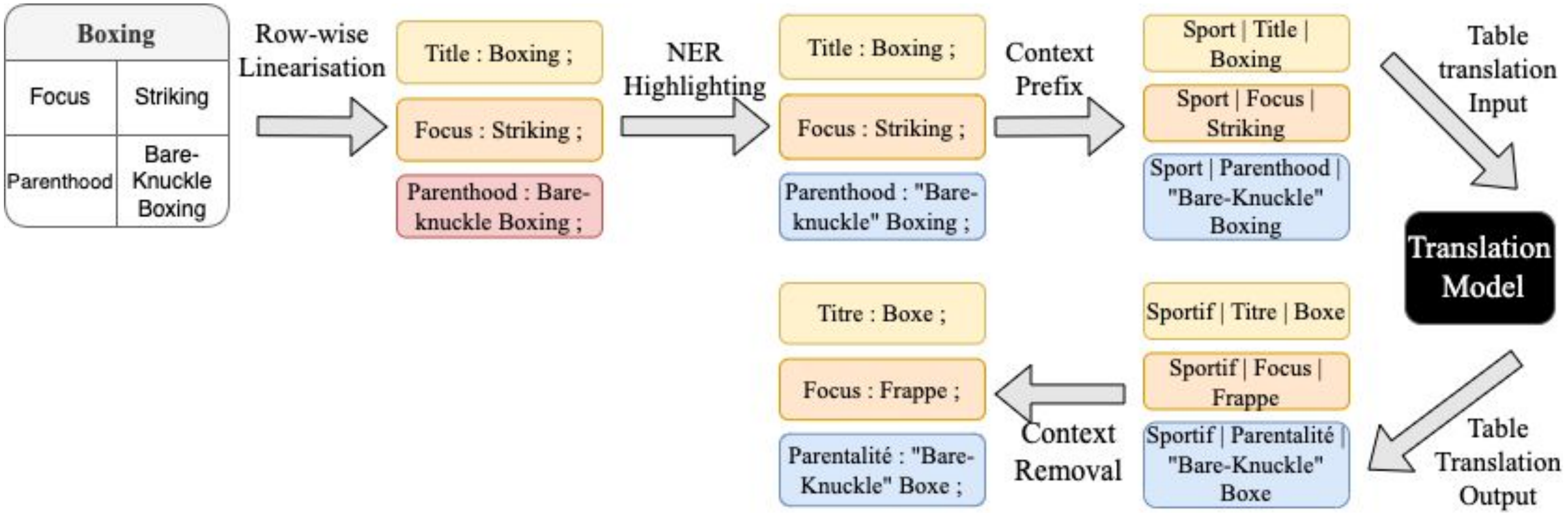
TRANSLATING TABLES



Translate each row using a suitable translation models.

For **each language**, we utilize a **different model** (*Optimal).

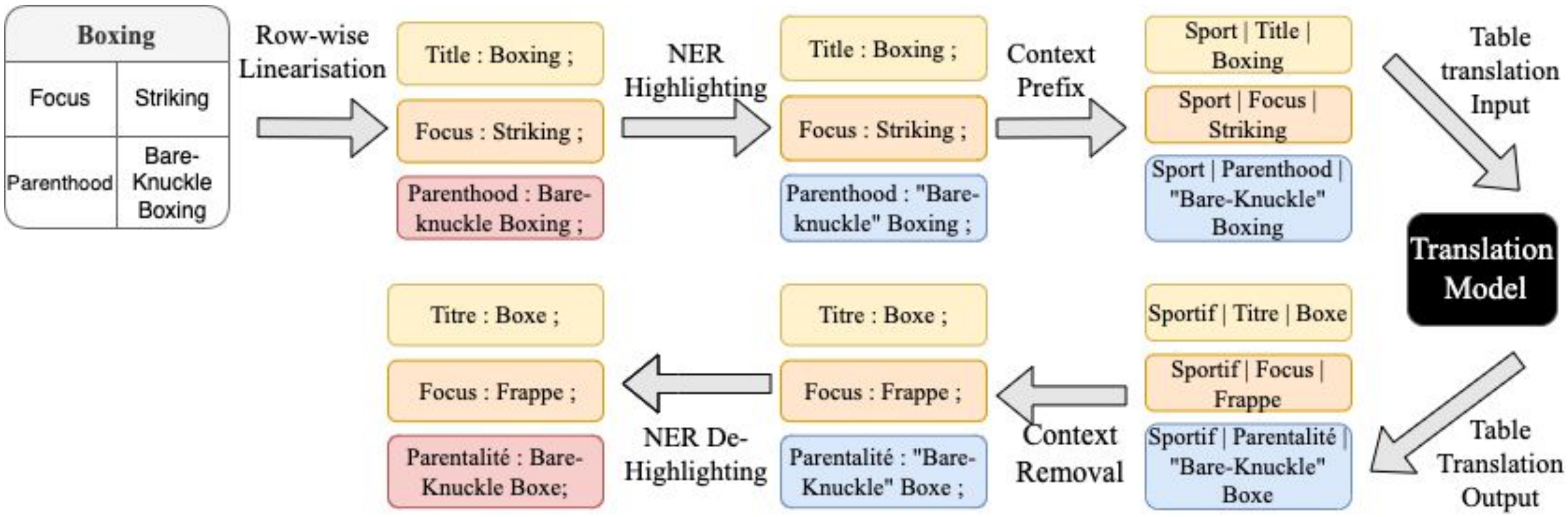
TRANSLATING TABLES



After translating we remove the added context i.e. category information.

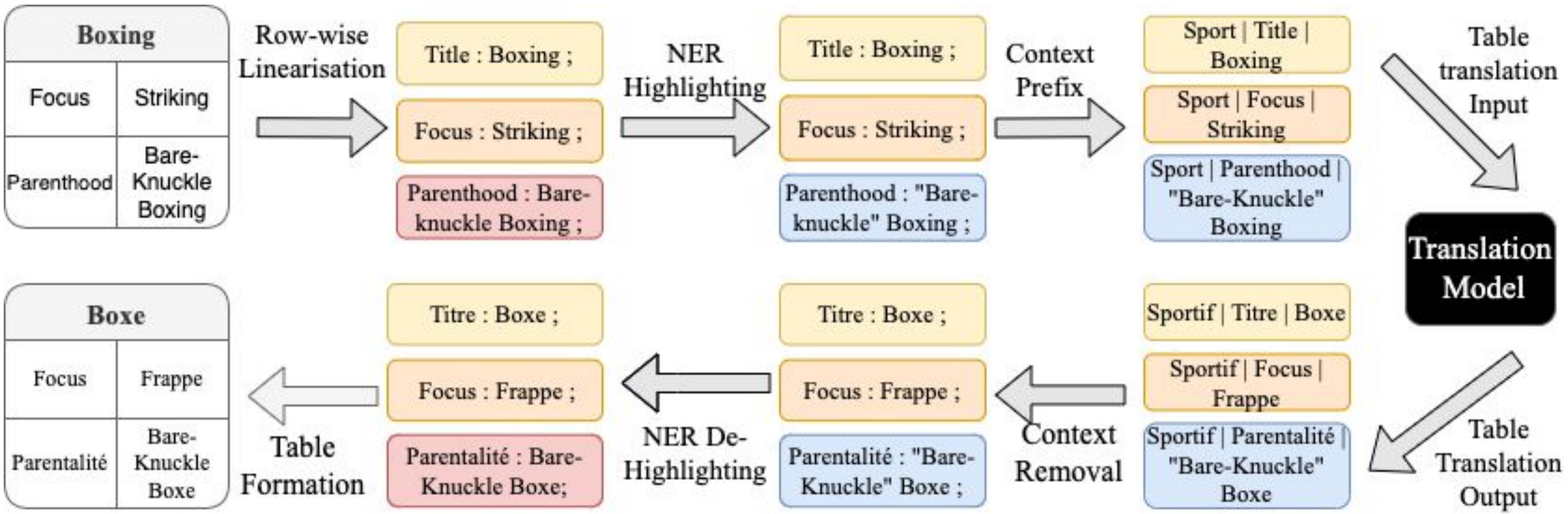
We also convert the delimiter (|) to colon (:). Also, add semi-colon (;) in row end.

TRANSLATING TABLES



Next, we **remove the highlights** (“ ”) around the **named entities** for all rows.

Translating Tables



Finally, we e**xtract the translated keys and values** from the linearised translated rows, and return them to **tabular format.**

Measuring The Quality Of Translations

Paraphrase score

Capture similarity between original and back-translated texts.

Embedding created using **all-mpnet-v2 model** trained with **Sentence BERT.**

Measuring The Quality Of Translations

Paraphrase score

Capture similarity between original and back-translated texts.

Embedding created using **all-mpnet-v2 model** trained with **Sentence BERT.**

Multilingual ParaScore

Use **multilingual-mpnet -base-v2 model** to creates embeddings for both original and translated text.

Finally, calculate the **cosine similarity** between the two embeddings.

MEASURING THE QUALITY OF TRANSLATIONS

Paraphrase score

Capture similarity between original and back-translated texts.

Embedding created using **all-mpnet-v2 model** trained with **Sentence BERT.**

Multilingual ParaScore

Use **multilingual-mpnet -base-v2 model** to creates embeddings for both original and translated text.

Finally, calculate the **cosine similarity** between the two embeddings.

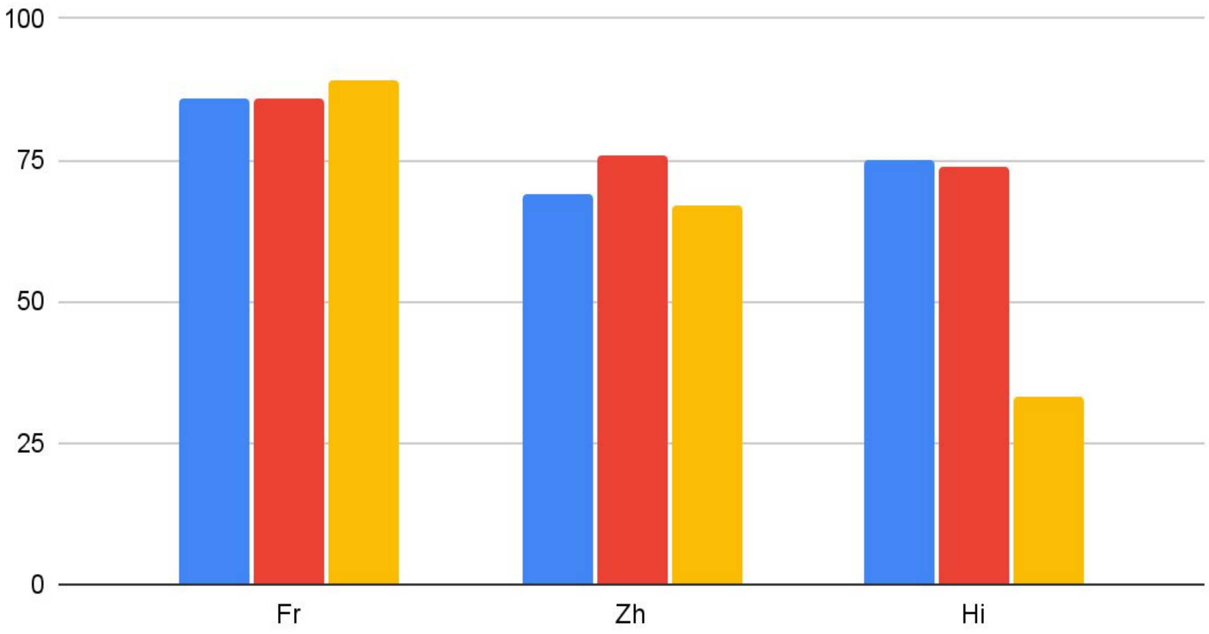
BERTScore

An automatic score which uses BERT embedding similarity to estimate translation quality.

MEASURING THE QUALITY OF TRANSLATIONS

Paraphrase score	Multilingual ParaScore	BERTScore	Human Score
Capture similarity between original and back-translated texts. Embedding created using **all-mpnet-v2 model** trained with **Sentence BERT.**	Use **multilingual-mpnet -base-v2 model** to creates embeddings for both original and translated text. Finally, calculate the **cosine similarity** between the two embeddings.	An automatic score which uses BERT embedding similarity to estimate translation quality.	Five annotators to label 500 examples per model and language. Follow, the Koehn and Monz, 2006 annotation guidelines.

Selecting The Model For Translation



HES for High (Fr), Mid (Zh), Low (Hi) and Resource Languages

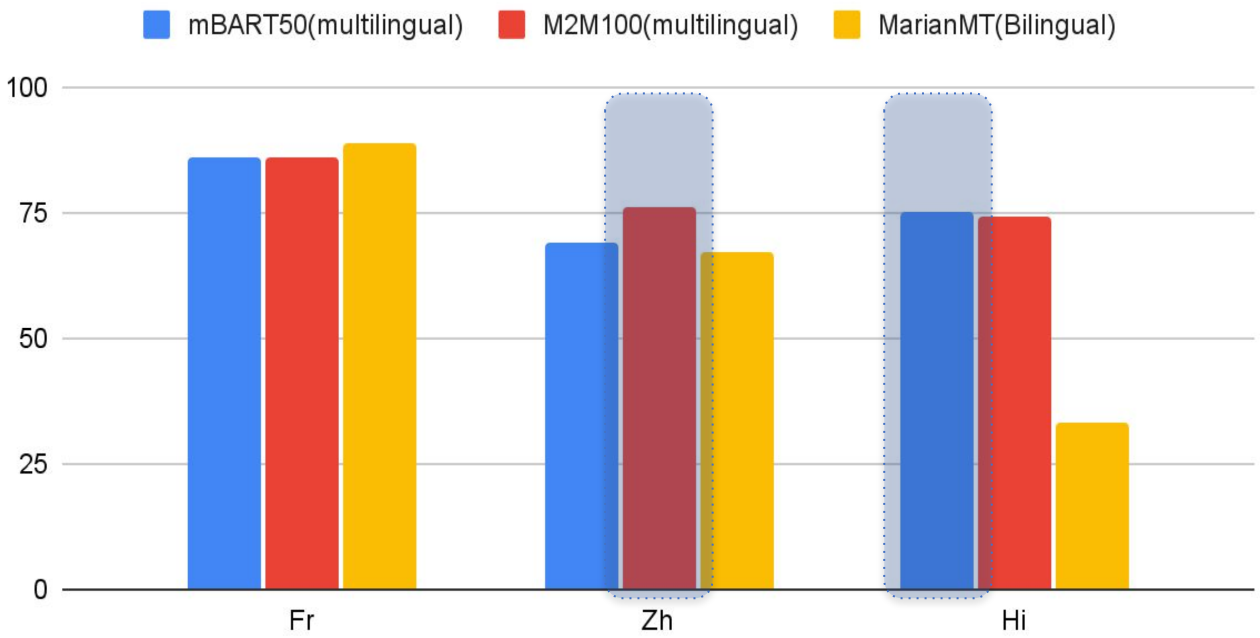
*Languages arranged in order of open source translation resource size

Translation Model Preference

High Resource

- Bi-lingual MT models (MarianMT)

SELECTING THE MODEL FOR TRANSLATION



HES for High (Fr), Mid (Zh), Low (Hi) and Resource Languages

*Languages arranged in order of open source translation resource size

Translation Model Preference

High Resource

- Bi-lingual MT models (MarianMT)

Mid & Low Resource

- Multi-lingual models (mBART or M2M)

Several Test-Splits Mitigate Artifact Issues

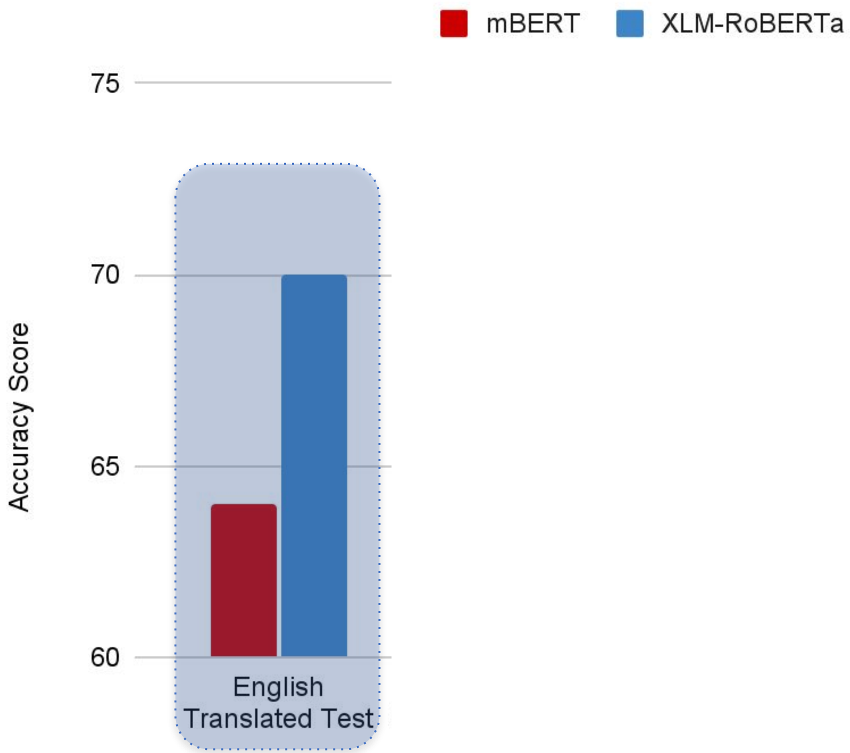
Claim: A single fixed test set is not enough

Need multiple test sets with controlled differences from each other.

- **α1** contains table from same domain (similar to dev & train set)
- **α2** has examples from same domain but entail-contradict label (e.g. ‘over’ to ‘under’) flipped by minimal change i.e. **adversarial**.
- **α3** is **zero-shot** cross domain tables (exclusive from train set domains)

Check out InfoTabS: https://infotabs.github.io

RESULTS AND ANALYSIS

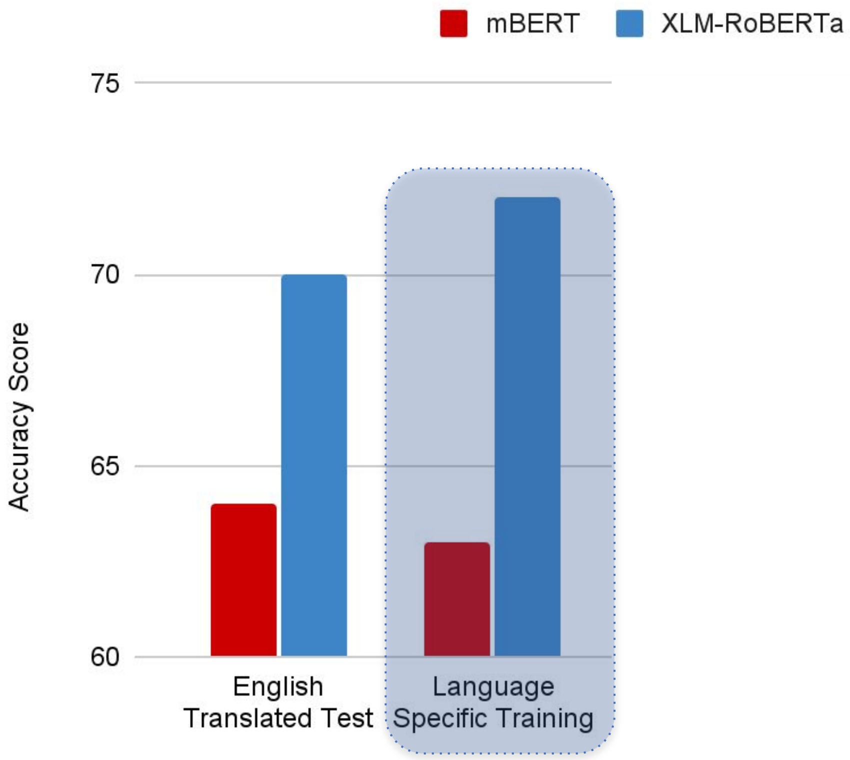


Results on α_1 test set

Translated Test

- Tables and Hypothesis are translated to English.
- Uses original INFOTABS data for training.
- English translated data is used for inference.

Results And Analysis



Results on α_1 test set

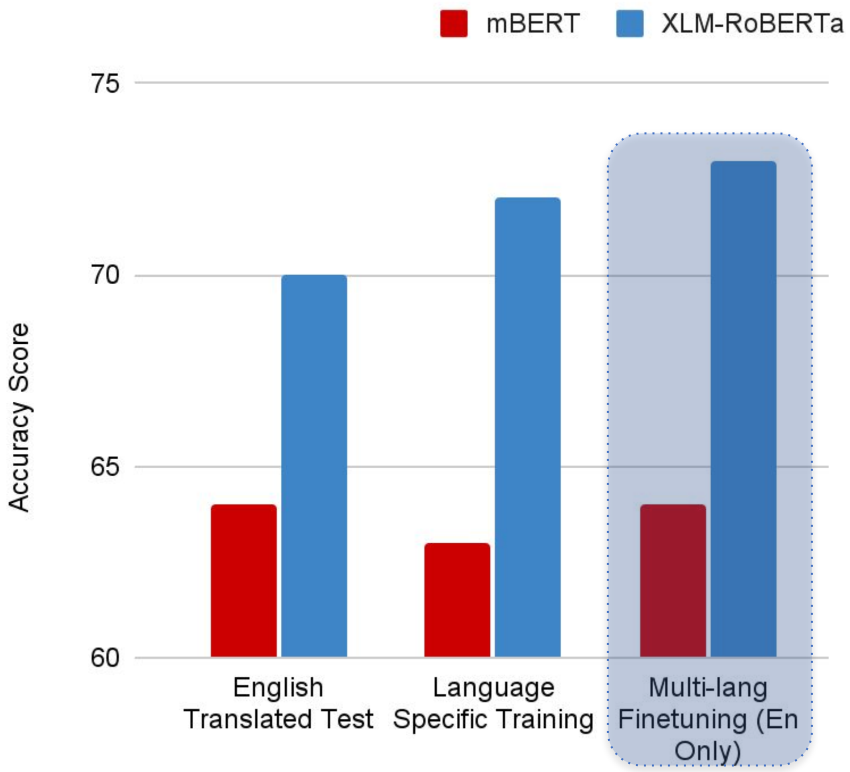
Language Specific Training

- Training and evaluation done on each language separately i.e. multiple bilingual models

- Each model is evaluated on same language set it is specifically trained on.

**we also did a cross lingual evaluation of these models*

RESULTS AND ANALYSIS

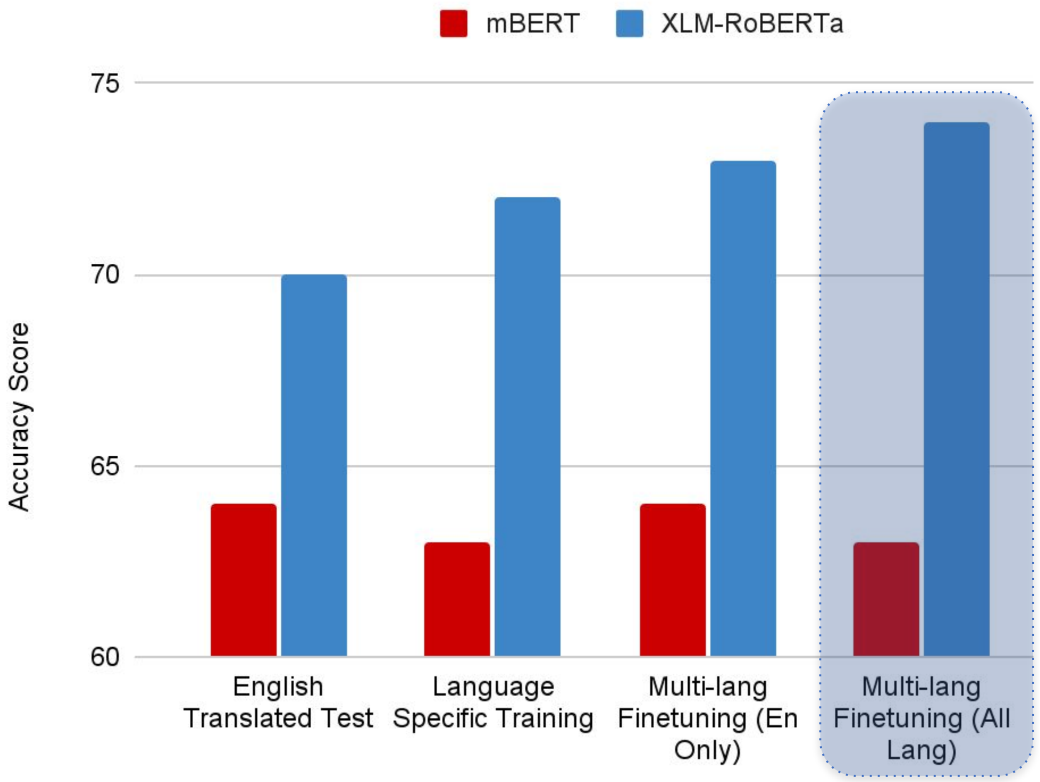


Results on $\underline{\alpha_1}$ test set

Multi Language Fine Tuning (En Only)

- Multiple models first trained for English InfoTabS data.
- Followed by Language Specific Fine tuning for each language. i.e. multiple bilingual models
- Each model is evaluated on same language set as it is specifically trained on.

RESULTS AND ANALYSIS

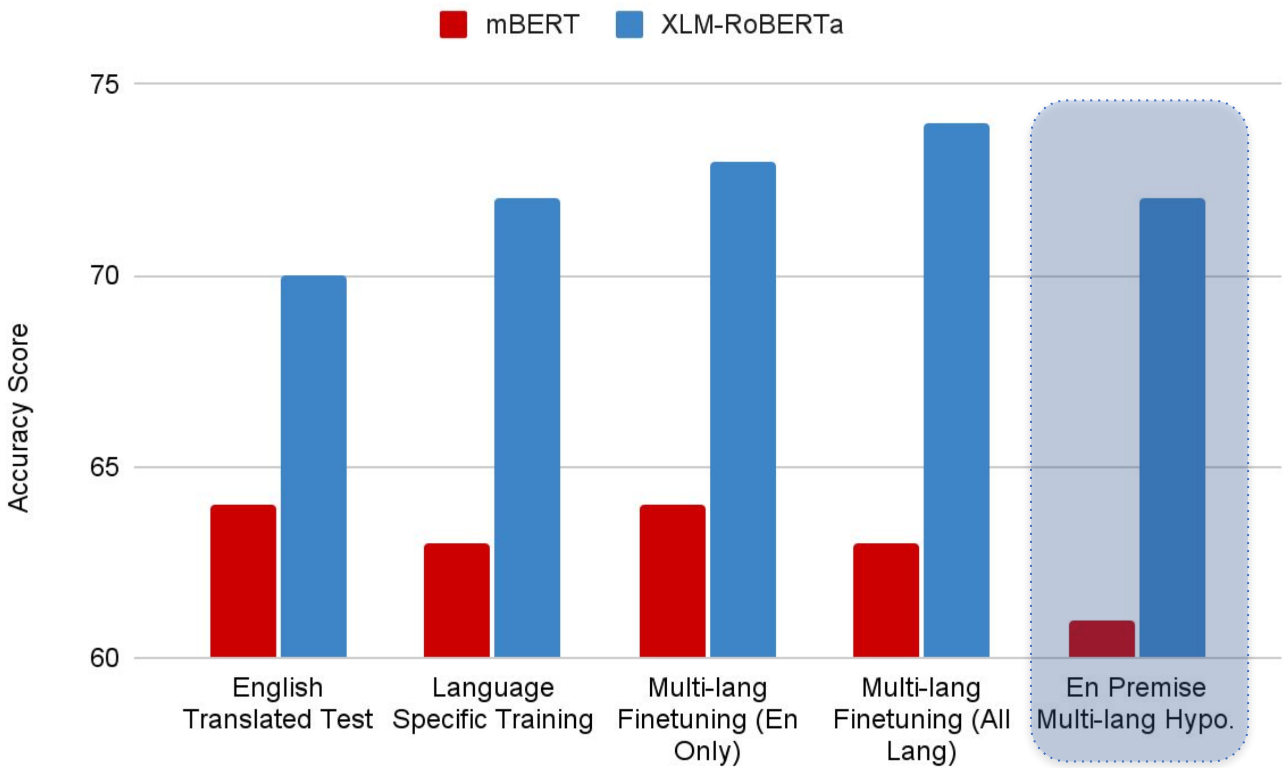


Results on α_1 test set

Multi Language Fine tuning (All Languages)

- Unified model first trained for English InfoTabS data.
- Followed by Language Specific Fine tuning for All languages.
  i.e. unified multilingual model
- Unified model is evaluated on all the language set.

RESULTS AND ANALYSIS



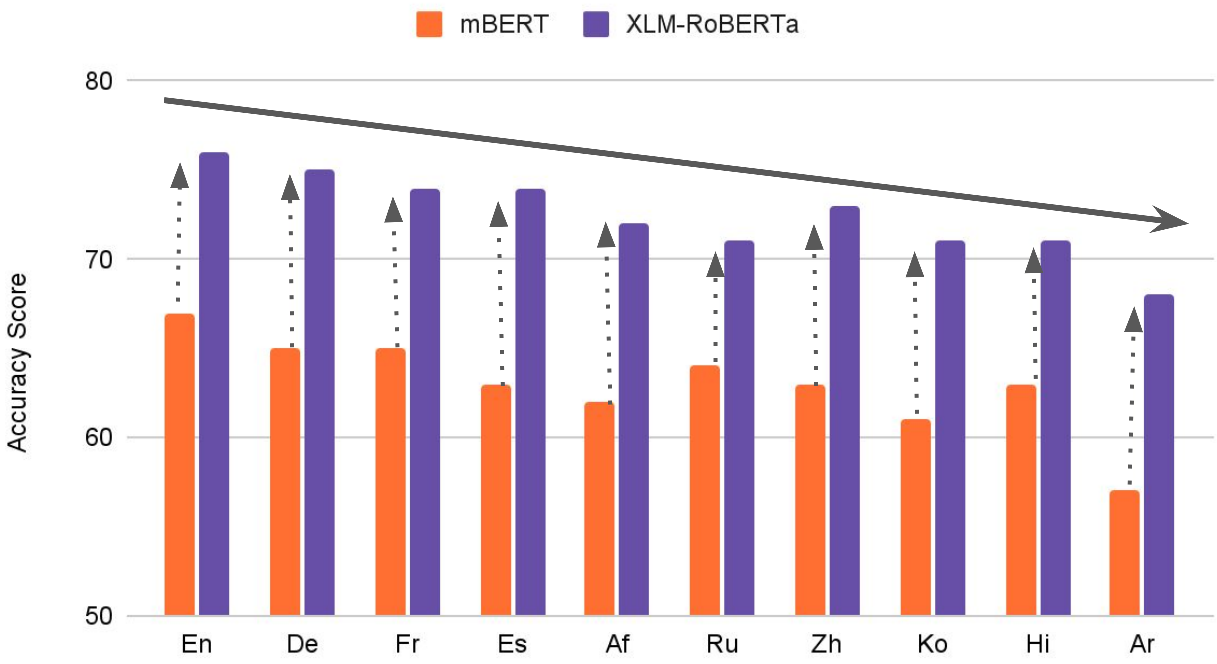
Results on α_1 test set

Bi-lingual Inference English Premise, Multilingual Hypothesis

Use English Premise with language specific hypothesis.
I.e. bilingual models

Here too, each model is evaluated on the language is trained on.

LANGUAGE SPECIFIC PERFORMANCE COMPARISON



Results on $\underline{\alpha}_1$ for Language Specific Baseline Task

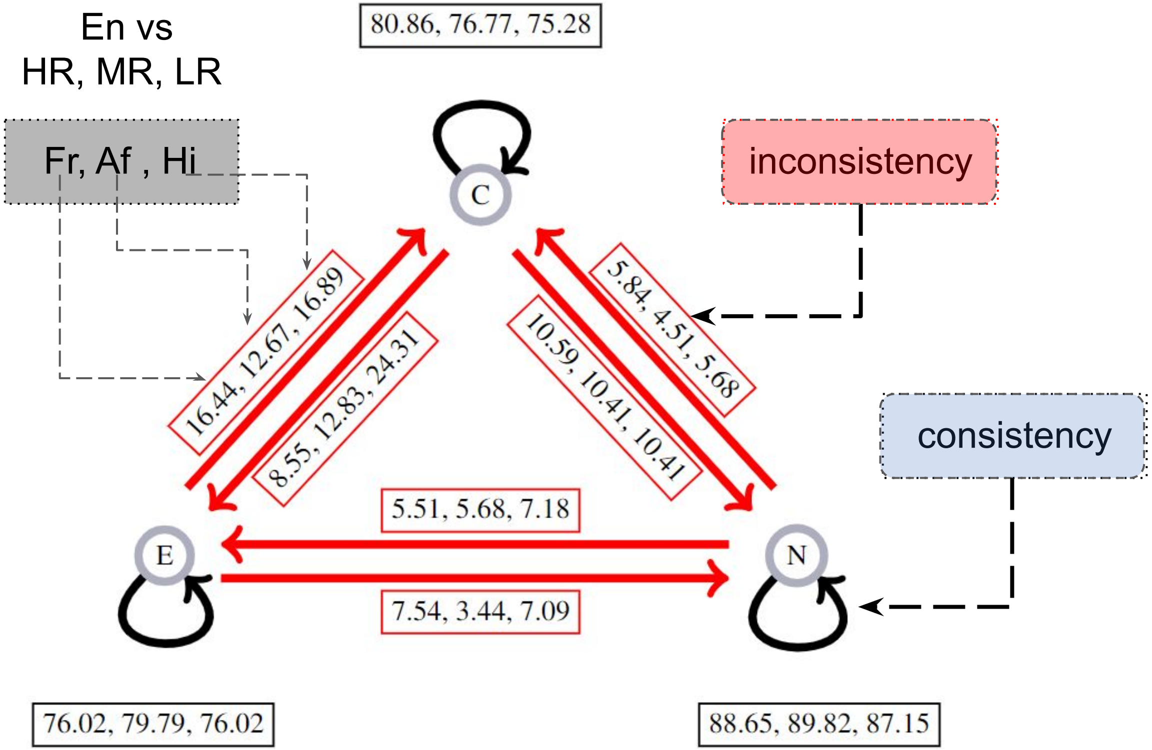
XLM-RoBERTa > mBERT

- more parameters
- learning objective
- longer training
- more languages

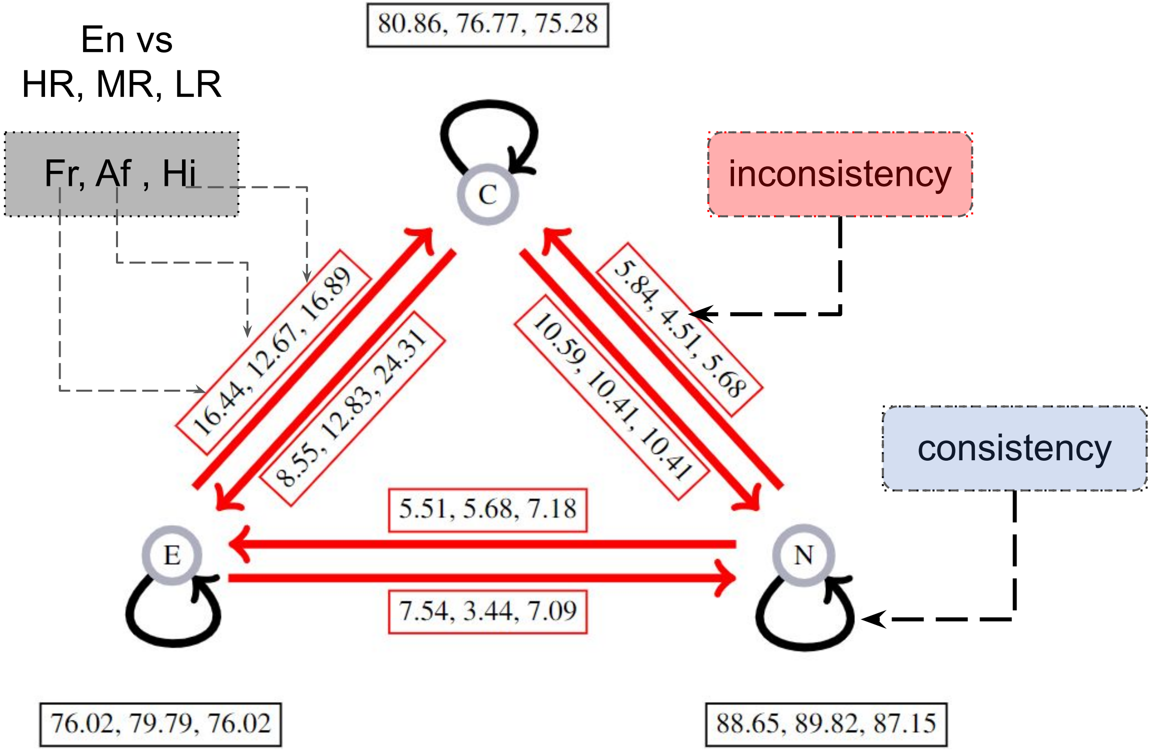
High Resource > Low Resource Performance

- mBERT more consistent

Cross Language Model Consistency



CROSS LANGUAGE MODEL CONSISTENCY

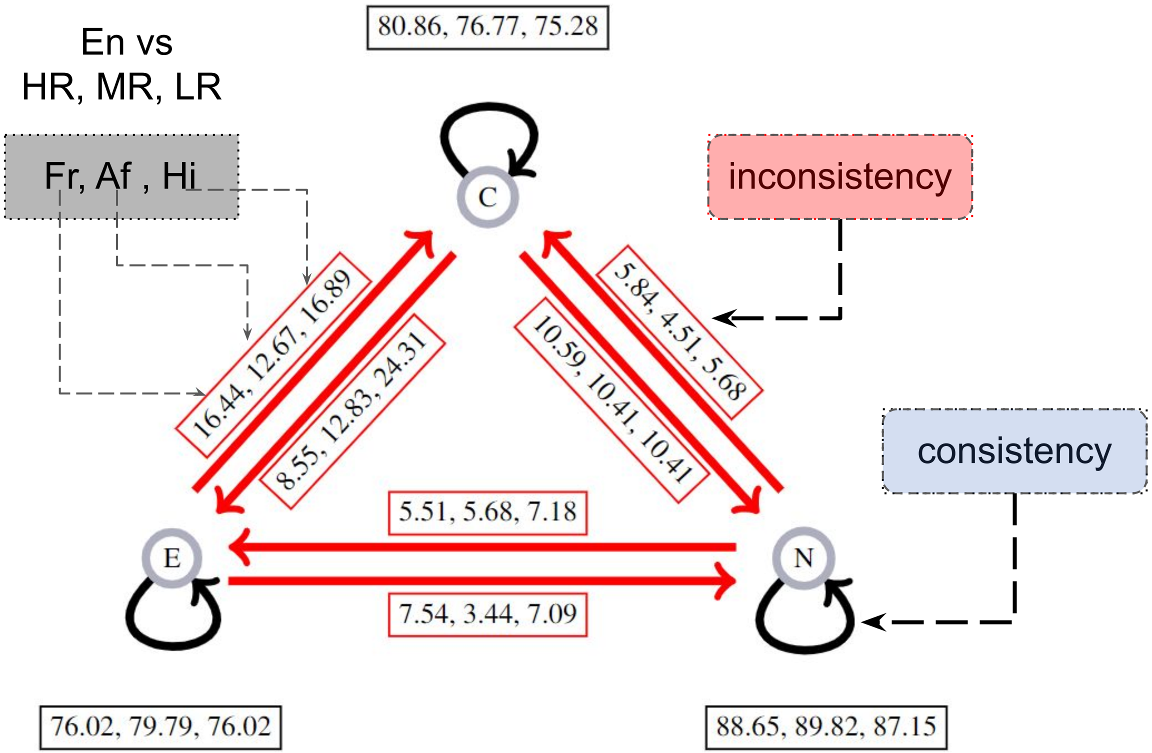


Inconsistent Summary

Label Specific

- E → C >> E → N
- C → N > C → E
- N → E == N → C

Cross Language Model Consistency



Inconsistent Summary

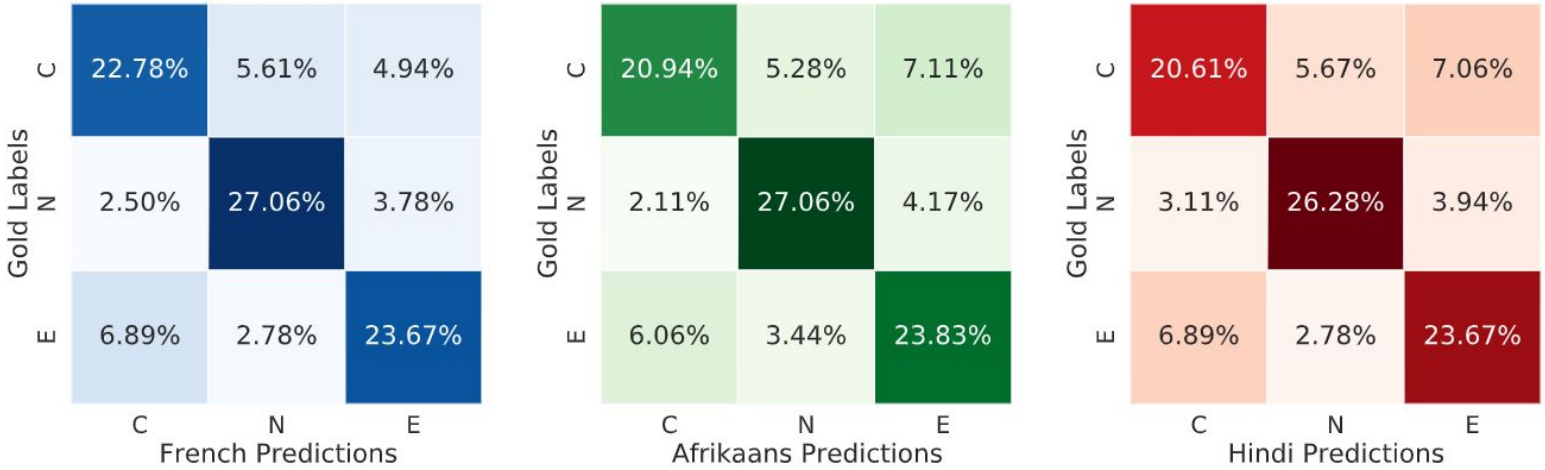
Label Specific

- E → C >> E → N
- C → N > C → N
- N → E == N → C

Language Specific

- E → C (Hi ~ Fr > Af)
- C → E (Hi >> Af > Fr)
- E → N (Hi ~Fr > Af)
- N → E (Af > Fr ~Hi)

CONFUSION MATRIX



For **low resource** model **wrongly** predict **Entailment for Contradiction**

In addition, for **Hi,** the model **predicts Neutral for Entailment instances**

TAKEAWAY

- XINFOTABS, is a multilingual dataset for semi-structured tabular inference which contains instances in ten diverse languages.

- To create XINFOTABS, we leverage cutting-edge machine translation models which provide high-quality translations of semi-structured tabular data.

- We access reasoning ability of state-of-the-art multilingual models trained with varying strategies over XINFOTABS.