ACL 2022
22ND – 27TH MAY | 60TH MEETING | DUBLIN

Workshop on Speech and Language Technologies for Dravidian Languages

MOHAMED BIN ZAYED
UNIVERSITY OF
ARTIFICIAL INTELLIGENCE

MuCoT: Multilingual Contrastive Training For Question-Answering In Low-resource Languages

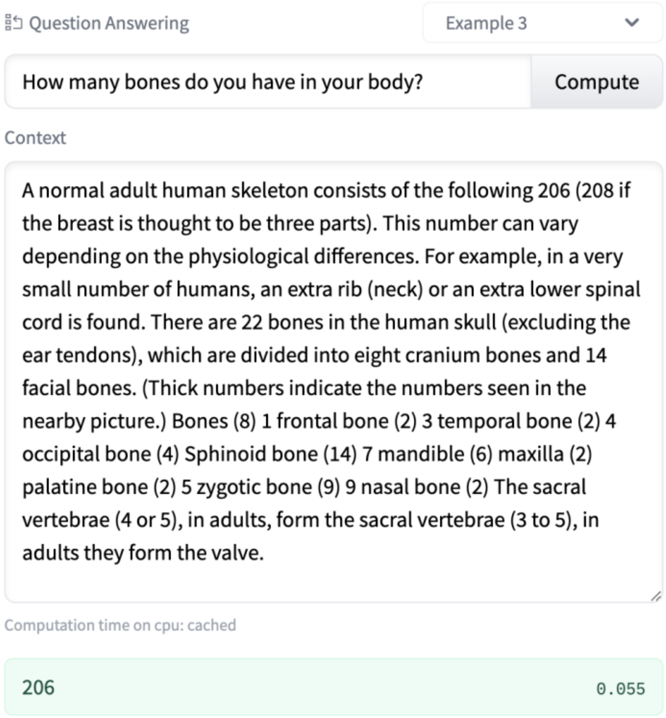
Gokul Karthik Kumar, Abhishek Singh Gehlot, Sahal Shaji Mullappilly, Karthik Nandakumar

Mohamed Bin Zayed University of Artificial Intelligence (MBZUAI)

Abu Dhabi, UAE

gokul.kumar@mbzuai.ac.ae

Multilingual Question Answering



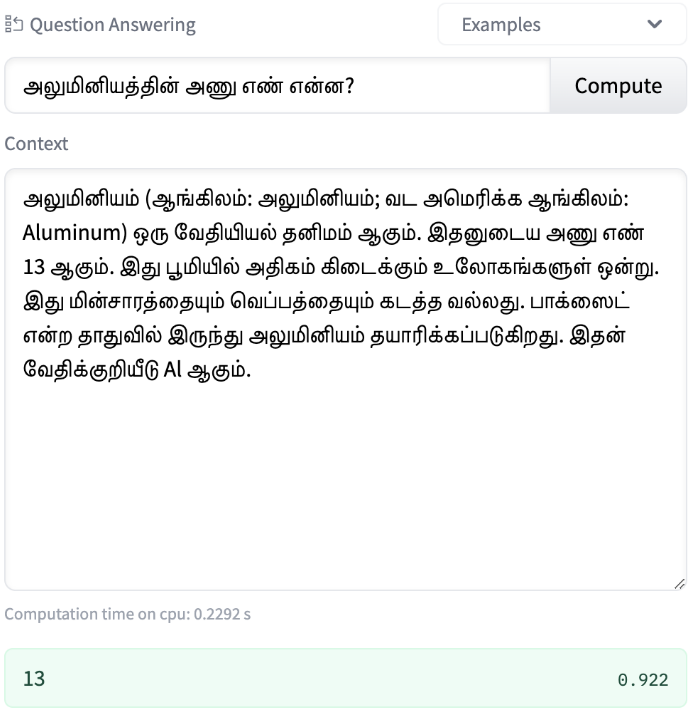
How many bones do you have in your body?

Context

A normal adult human skeleton consists of the following 206 (208 if the breast is thought to be three parts). This number can vary depending on the physiological differences. For example, in a very small number of humans, an extra rib (neck) or an extra lower spinal cord is found. There are 22 bones in the human skull (excluding the ear tendons), which are divided into eight cranium bones and 14 facial bones. (Thick numbers indicate the numbers seen in the nearby picture.) Bones (8) 1 frontal bone (2) 3 temporal bone (2) 4 occipital bone (4) Sphinoid bone (14) 7 mandible (6) maxilla (2) palatine bone (2) 5 zygotic bone (9) 9 nasal bone (2) The sacral vertebrae (4 or 5), in adults, form the sacral vertebrae (3 to 5), in adults they form the valve.

Computation time on cpu: cached

206 0.055



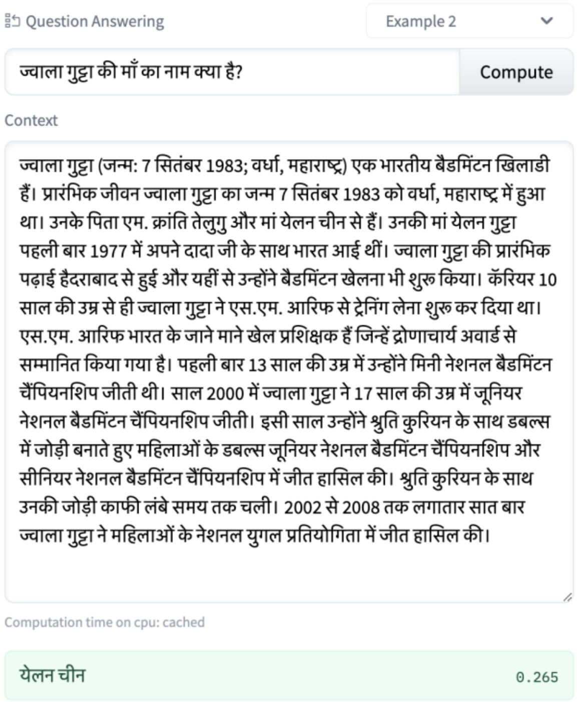
அலுமினியத்தின் அணு எண் என்ன?

Context

அலுமினியம் (ஆங்கிலம்: அலுமினியம்; வட அமெரிக்க ஆங்கிலம்: Aluminum) ஒரு வேதியியல் தனிமம் ஆகும். இதனுடைய அணு எண் 13 ஆகும். இது பூமியில் அதிகம் கிடைக்கும் உலோகங்களுள் ஒன்று. இது மின்சாரத்தையும் வெப்பத்தையும் கடத்த வல்லது. பாக்ஸைட் என்ற தாதுவில் இருந்து அலுமினியம் தயாரிக்கப்படுகிறது. இதன் வேதிக்குறியீடு Al ஆகும்.

Computation time on cpu: 0.2292 s

13 0.922



ज्वाला गुट्टा की माँ का नाम क्या है?

Context

ज्वाला गुट्टा (जन्म: 7 सितंबर 1983; वर्धा, महाराष्ट्र) एक भारतीय बैडमिंटन खिलाडी हैं। प्रारंभिक जीवन ज्वाला गुट्टा का जन्म 7 सितंबर 1983 को वर्धा, महाराष्ट्र में हुआ था। उनके पिता एम. क्रांति तेलुगु और मां येलन चीन से हैं। उनकी मां येलन गुट्टा पहली बार 1977 में अपने दादा जी के साथ भारत आई थीं। ज्वाला गुट्टा की प्रारंभिक पढ़ाई हैदराबाद से हुई और यहीं से उन्होंने बैडमिंटन खेलना भी शुरू किया। कॅरियर 10 साल की उम्र से ही ज्वाला गुट्टा ने एस.एम. आरिफ से ट्रेनिंग लेना शुरू कर दिया था। एस.एम. आरिफ भारत के जाने माने खेल प्रशिक्षक हैं जिन्हें द्रोणाचार्य अवार्ड से सम्मानित किया गया है। पहली बार 13 साल की उम्र में उन्होंने मिनी नेशनल बैडमिंटन चैंपियनशिप जीती थी। साल 2000 में ज्वाला गुट्टा ने 17 साल की उम्र में जूनियर नेशनल बैडमिंटन चैंपियनशिप जीती। इसी साल उन्होंने श्रुति कुरियन के साथ डबल्स में जोड़ी बनाते हुए महिलाओं के डबल्स जूनियर नेशनल बैडमिंटन चैंपियनशिप और सीनियर नेशनल बैडमिंटन चैंपियनशिप में जीत हासिल की। श्रुति कुरियन के साथ उनकी जोड़ी काफी लंबे समय तक चली। 2002 से 2008 तक लगातार सात बार ज्वाला गुट्टा ने महिलाओं के नेशनल युगल प्रतियोगिता में जीत हासिल की।

Computation time on cpu: cached

येलन चीन 0.265

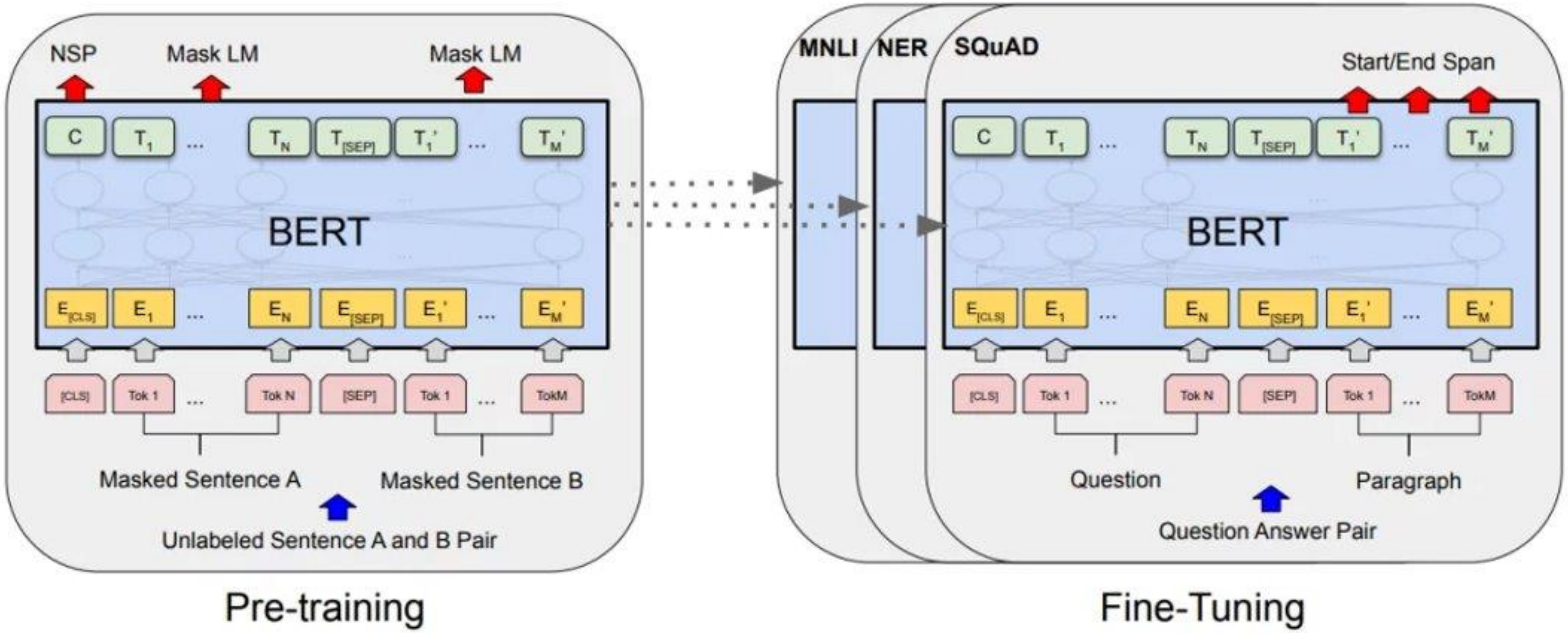
The Problem

- Accuracy of English-language Question Answering (QA) systems has improved significantly in recent years with the advent of Transformer-based models (e.g., BERT)
- Multi-lingual BERT-based models (mBERT) are often used to transfer knowledge from high-resource languages to low-resource languages
- Directly training an mBERT-based QA system for low-resource languages is challenging due to the paucity of training data

Solution Overview

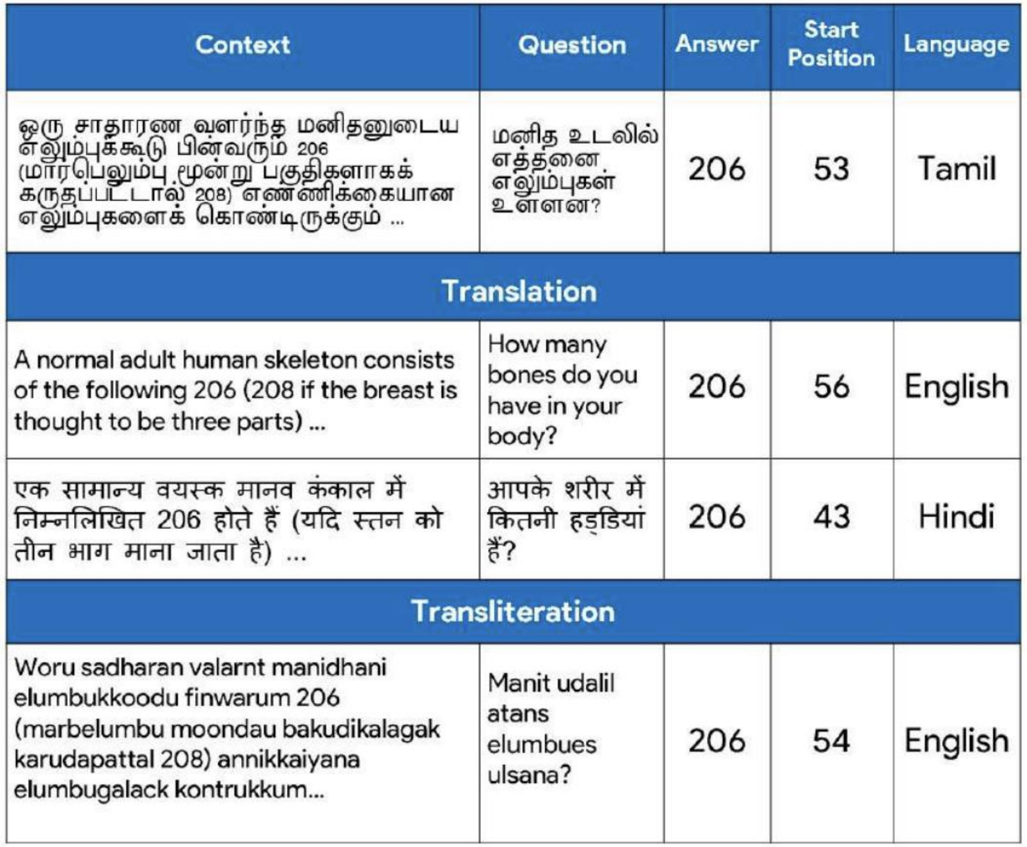
- We augment the QA samples of the target language using **translation** and **transliteration** into other languages and use the augmented data to fine-tune an mBERT-based QA model, which is already pre-trained in English
- Experiments on the Google ChAII dataset show that fine-tuning the mBERT model with **translations from the same language family boosts the question-answering performance**, whereas the performance degrades in the case of crosslanguage families
- We further show that introducing a **contrastive loss** between the translated question-context feature pairs during the fine-tuning process, prevents such degradation

1. Pretraining mBERT with SQuAD



Source: Devlin et al., 2018

2. Translation & Transliteration

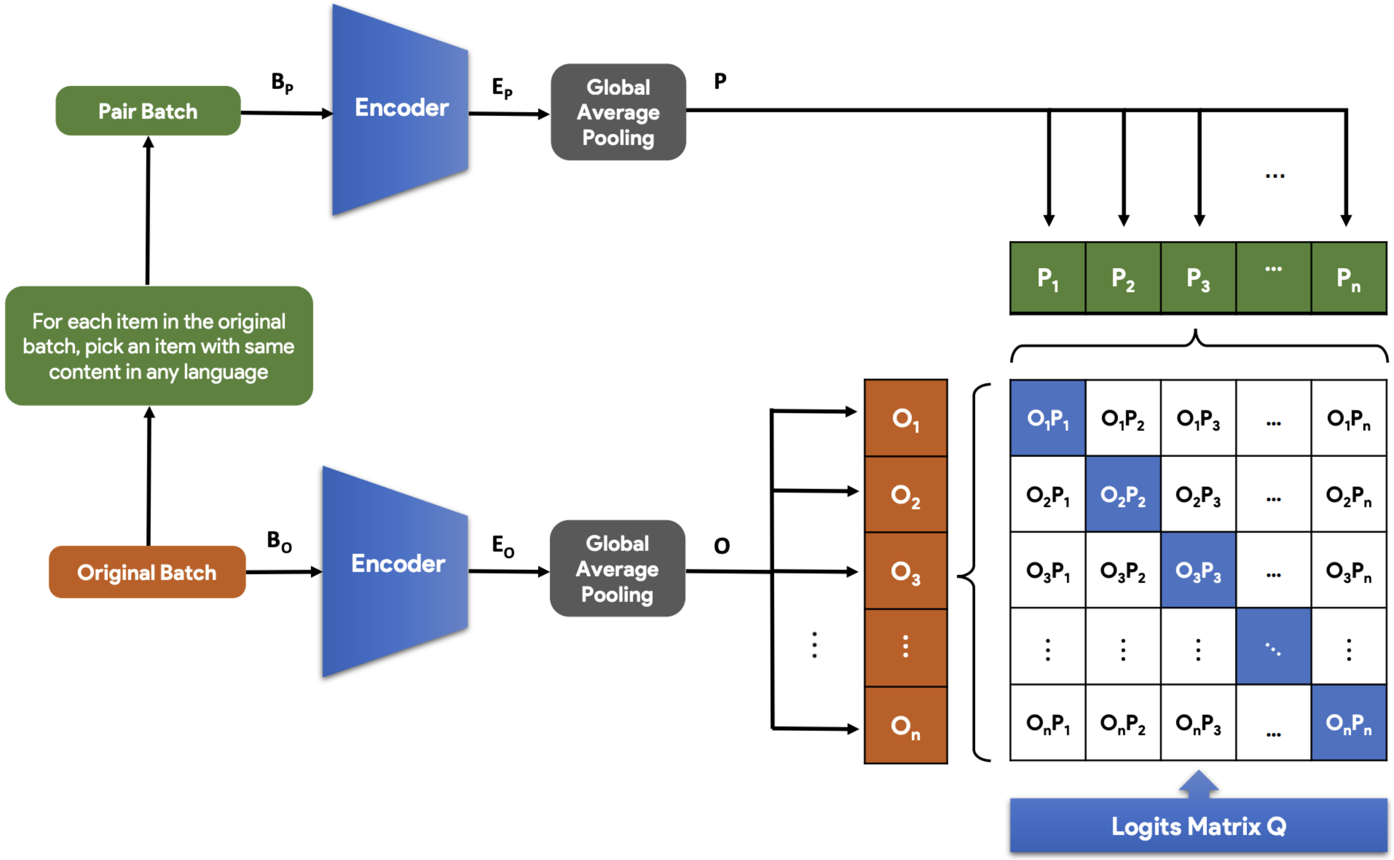


Context	Question	Answer	Start Position	Language
ஒரு சாதாரண வளர்ந்த மனிதனுடைய எலும்புக்கூடு பின்வரும் 206 (மார்பெலும்பு மூன்று பகுதிகளாகக் கருதப்பட்டால் 208) எண்ணிக்கையான எலும்புகளைக் கொண்டிருக்கும் ...	மனித உடலில் எத்தனை எலும்புகள் உள்ளன?	206	53	Tamil
Translation				
A normal adult human skeleton consists of the following 206 (208 if the breast is thought to be three parts) ...	How many bones do you have in your body?	206	56	English
एक सामान्य वयस्क मानव कंकाल में निम्नलिखित 206 होते हैं (यदि स्तन को तीन भाग माना जाता है) ...	आपके शरीर में कितनी हड्डियां हैं?	206	43	Hindi
Transliteration				
Woru sadharan valarnt manidhani elumbukkoodu finwarum 206 (marbelumbu moondau bakudikalagak karudapattal 208) annikkaiyana elumbugalack kontrukkum...	Manit udalil atans elumbues ulsana?	206	54	English

Figure 1: Example of a QA record from the ChAII QA dataset along with the translation and transliteration done on that record.

- Kudugunta et al. (2019) showed that languages under the same family have similar representations in multilingual models
- We put together translations and transliterations from related languages within the same language family to achieve better performance
- Same language family QA performance improved
- Cross language family QA performance degraded

3. Contrastive Loss



O_1P_1	O_1P_2	O_1P_3	...	O_1P_n
O_2P_1	O_2P_2	O_2P_3	...	O_2P_n
O_3P_1	O_3P_2	O_3P_3	...	O_3P_n
⋮	⋮	⋮	⋱	⋮
O_nP_1	O_nP_2	O_nP_3	...	O_nP_n

$$O = gap(enc(B_o)),$$

$$P = gap(enc(B_p)),$$

$$Q = OP^T,$$

$$L_{contrastive} = \frac{L_{ce}^{row}(Q) + L_{ce}^{column}(Q)}{2}$$

Figure 4: Logits matrix computation for the input to contrastive loss, similar to CLIP (Radford et al., 2021)

Proposed Pipeline for QA in Tamil & Hindi

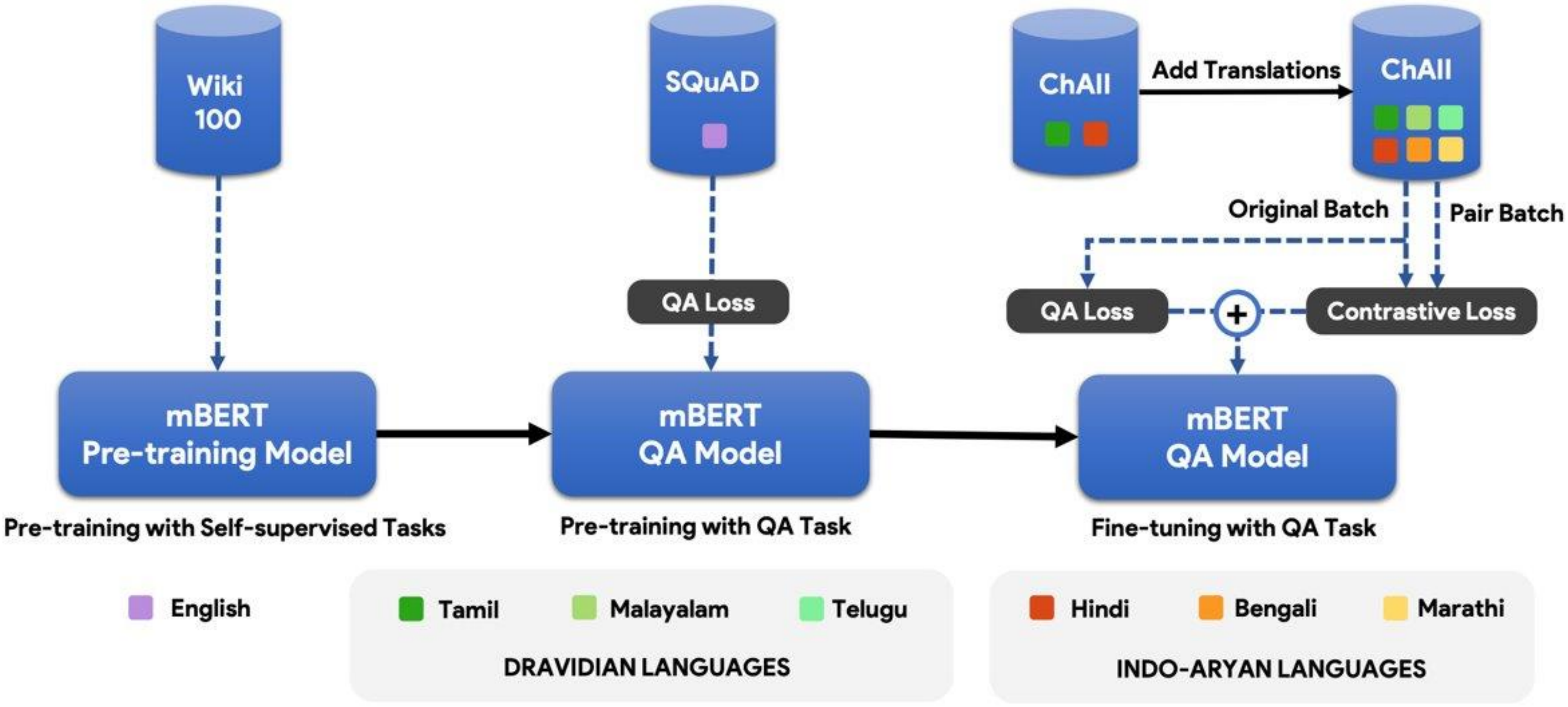
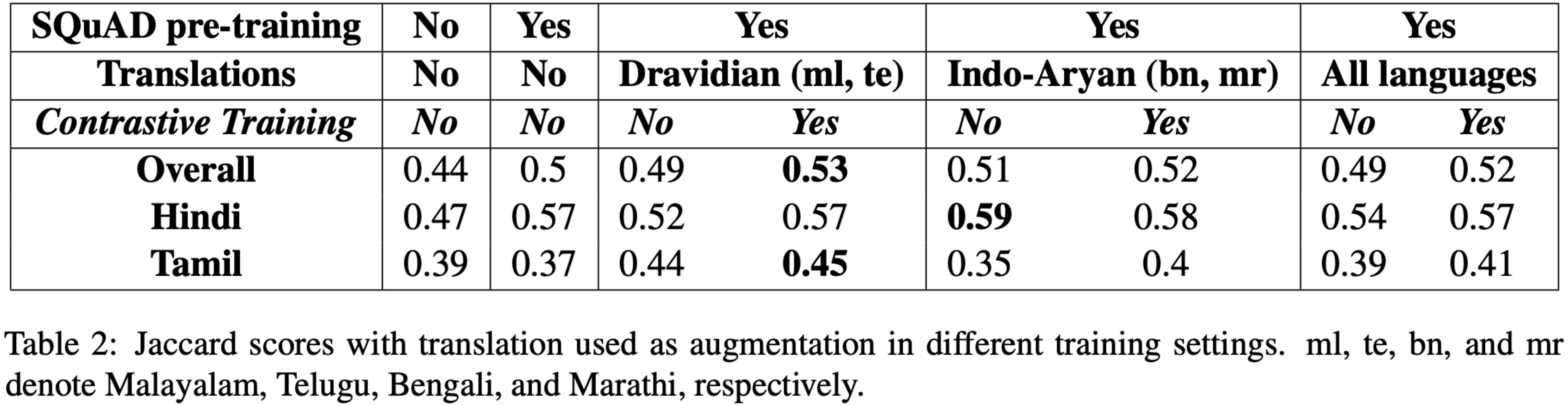


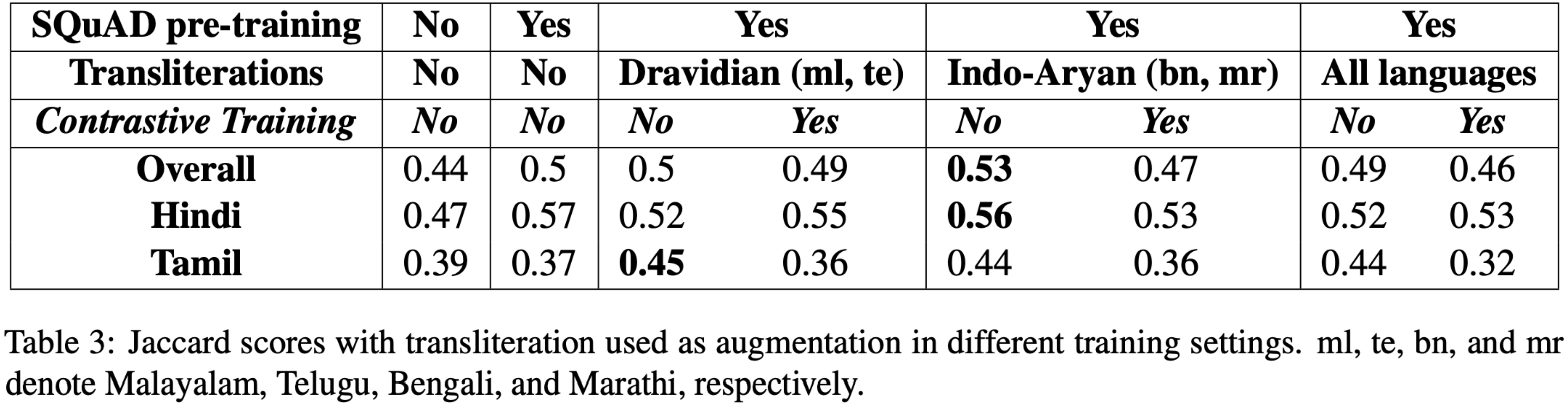
Figure 3: Proposed training pipeline of MuCoT for question answering in low resource languages Tamil and Hindi.

Results for ChAII dataset



SQuAD pre-training	No	Yes	Yes		Yes		Yes	
Translations	No	No	Dravidian (ml, te)		Indo-Aryan (bn, mr)		All languages	
Contrastive Training	*No*	*No*	*No*	*Yes*	*No*	*Yes*	*No*	*Yes*
Overall	0.44	0.5	0.49	**0.53**	0.51	0.52	0.49	0.52
Hindi	0.47	0.57	0.52	0.57	**0.59**	0.58	0.54	0.57
Tamil	0.39	0.37	0.44	**0.45**	0.35	0.4	0.39	0.41

Table 2: Jaccard scores with translation used as augmentation in different training settings. ml, te, bn, and mr denote Malayalam, Telugu, Bengali, and Marathi, respectively.



SQuAD pre-training	No	Yes	Yes		Yes		Yes	
Transliterations	No	No	Dravidian (ml, te)		Indo-Aryan (bn, mr)		All languages	
Contrastive Training	*No*	*No*	*No*	*Yes*	*No*	*Yes*	*No*	*Yes*
Overall	0.44	0.5	0.5	0.49	**0.53**	0.47	0.49	0.46
Hindi	0.47	0.57	0.52	0.55	**0.56**	0.53	0.52	0.53
Tamil	0.39	0.37	**0.45**	0.36	0.44	0.36	0.44	0.32

Table 3: Jaccard scores with transliteration used as augmentation in different training settings. ml, te, bn, and mr denote Malayalam, Telugu, Bengali, and Marathi, respectively.

Conclusion

- With Internet usage expanding every day, there is an increasing need to develop better NLP models for a variety of downstream tasks in vernacular languages
- As most of these languages do not have labeled resources that are sufficient to train standalone modern deep learning models, we need to rely on multilingual models and enhance them
- Our work is a step in this direction and is an attempt to understand and evaluate the impact of cross-lingual knowledge transfer through pre-training and finetuning
- Same phenomenon of cross-lingual transfer in other multilingual models (XLM-RoBERTa, MURIL etc.), language families and multilingual tasks