CORI: CJKV Benchmark with Romanization Integration - A step towards Cross-lingual Transfer Beyond Textual Scripts

Hoang Nguyen[1], Chenwei Zhang[2], Ye Liu[3], Natalie Parde[1], Eugene Rohrbaugh[4], Philip S. Yu[1]

[1] University of Illinois at Chicago
[2] Amazon [3] Salesforce Research
[3] Harrisburg University of Science and Technology

LREC-COLING 2024

Overview

- Background
- Preliminary Study
- Dataset Construction
- Framework
- Evaluation
- Conclusion

BACKGROUND

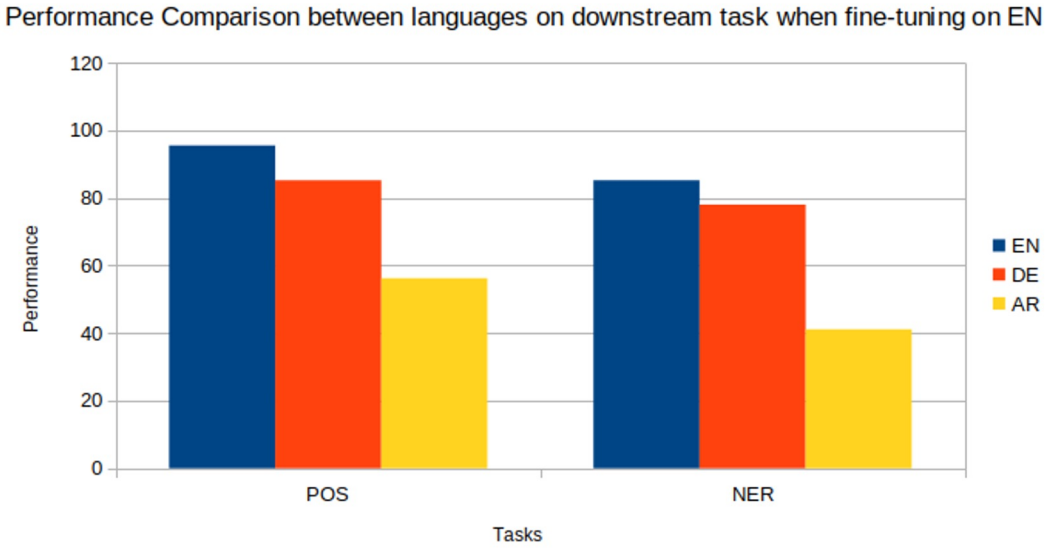
Background

EN: He was a scholar in Theology

DE: Er war ein Gelehrter der Theologie

AR: وكان عالما في اللاهوت

- Performance on low-resource language degrades on low-resource language (AR)
- Performance on degrades language with different script



Q1: Should EN always be chosen as source language?

Q2: How to overcome performance gap between languages with different textual writing scripts

Results are obtained from Yang et al., 2022. Enhancing Cross-lingual Transfer via Manifold Mixup. ICLR 2022

PRELIMINARY STUDY

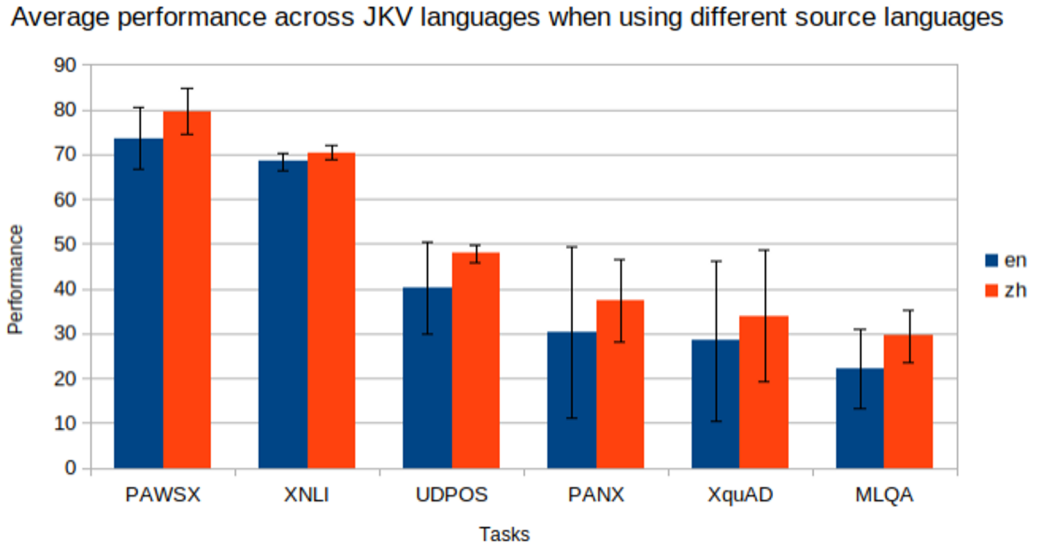
Background

Observation:

When ZH is leveraged as source language:

- Higher average performance among target JKV languages
- Lower STD among JKV

⇒ **ZH is a better source language** than EN when considering target JKV languages



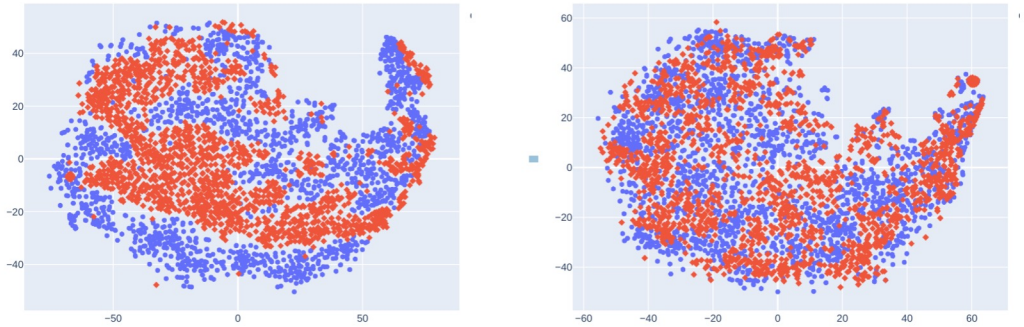
Source language matters on downstream task target language performance

Background

Observation:

- Representations of parallel inputs between JA and ZH are more aligned than between EN and ZH
- JA-ZH has higher language contact

⇒ **ZH is a better source language** than EN when considering target JKV languages



(a) EN source (blue), JA target (red)

(b) ZH source (blue), JA target (red)

Figure 1: Representation visualization of parallel sentences between source language and target lan guage (EN-JA (left) and ZH-JA (right)) of the fine-tuned XLM-R model on parallel PAWSX test set. The significant overlapping representation of ZH-JA demonstrates the lower representation discrepancy and higher language contact between source and target language.

➡ **Closely-related benchmark dataset is needed CJKV**

➡ **CORI Dataset Construction**

DATASET CONSTRUCTION

Dataset Construction

Challenge 1: Language Availability

Table 4: Details of CORI benchmark dataset. (✗) and (✓) denote unavailable and available data existent in XTREME benchmark respectively . MT, SEG, ROM correspond to Machine Translation, Pre-segmentation, Romanization processing respectively as described in Section 5. Y, N denote if the corresponding preprocessing is conducted for the dataset or not.

		MT	SEG	ROM	ZH		JA	KO	VI
					Train	Dev	Test	Test	Test
Sent-level	PAWSX	Y	Y	Y	49.4k (✗)	2k (✓)	2k (✗)	2k (✓)	2k (✗)
	XNLI	Y	Y	Y	392.7k (✓)	2.49k (✓)	5.01k (✗)	5.01k (✗)	5.01k (✓)
Token-level	UDPOS	N	N	Y	10k (✓)	1.6k (✓)	2.5k (✓)	4.7k(✓)	0.8k (✓)
	PANX	N	Y	Y	20k (✓)	10k (✓)	10k (✓)	10k (✓)	10k (✓)
Question Answering	XQuAD	Y	Y	Y	80.1k (✗)	8.87k (✗)	1.19k (✗)	1.19k (✗)	1.19k (✓)
	MLQA	Y	Y	Y	80.1k (✗)	8.87k (✗)	11.24 k (✗)	11.24k (✗)	11.24k (✓)

➡ **Multilingual Datasets have been created unequally across CJKV languages**

Dataset Construction

Challenge 2: Pre-segmentation

➡ **Inconsistent pre-segmentation across CJKV languages existent in XTREME**

Dataset Construction

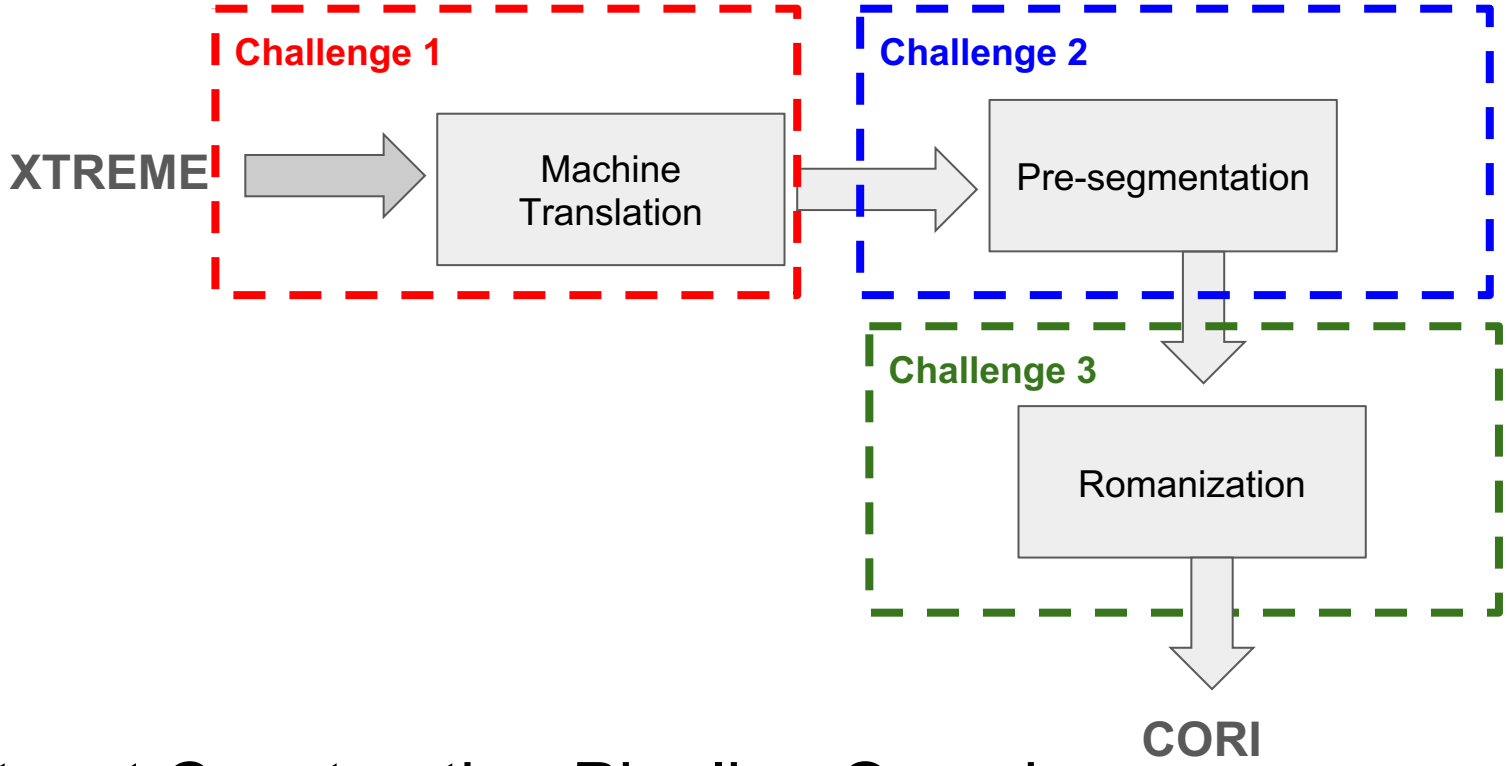
Challenge 3: Romanization

Table 1: Orthographic and Romanized representations (abbreviated as Ortho and Roman) of a sample sentence across CJKV languages where **colored segments** denote the corresponding semantic **words** defined in Section 5. (.) denotes the specific name of Romanization system of the respective language. The first sentence for each language denotes the currently preprocessed XTREME benchmark dataset.

Language	Input Type	Sample input sentence
EN	Ortho	He was a scholar in Metaphysical Literature , Theology and Classical sciences .
ZH (source)	Ortho	他是形而上学文学、神学和古典科学方面的学者。
	Ortho (seg)	他 // 是 // 形而上学 // 文学 // 、 // 神学 // 和 // 古典 // 科学 // 方面 // 的 // 学者 // 。
	Roman (Pinyin)	tā // shì // xíngérshàngxué // wénxué // 、 // shénxué // hé // gǔdiǎn // kēxué // fāngmiàn // de // xuézhě // 。
VI (target)	Ortho	Ông là một học giả về Văn học Siêu hình , Thần học và Khoa học Cổ điển .
	Ortho (seg)	Ông // là // một // học giả // về // Văn học // Siêu hình // , // Thần học // và // Khoa học // Cổ điển // .
	Roman	Ông // là // một // học giả // về // Văn học // Siêu hình // , // Thần học // và // Khoa học // Cổ điển // .
JA (target)	Ortho	彼は形而上学文学、神学、古典科学の学者でした。
	Ortho (seg)	彼 // は // 形而上学 // 文学 //、 神学 、 // 古典 // 科学 //の// 学者 //でし // た // 。
	Roman (Romaji)	kare // ha // keiji ue gaku // bungaku // , // shingaku // , // koten // kagaku // no // gakusha // deshi // ta // .
KO (target)	Ortho	그는형이상학문학, 신학및고전과학의학자이었습니다.
	Ortho (seg)	그 // 는 // 형이상학 // 문학 //, // 신학 // 및 // 고전 // 과학 // 의학자 // 이 // 었 // 습니다 // .
	Roman (Romaja)	eu // neun // hyeongisanghak // munhak // , // sinhak // mit // gojeon // gwahak // uihakja // i // eot // seupnida // .

Romanization captures linguistic contact beyond textual scripts, providing beneficial signals for cross-lingual transfer

Dataset Construction



Dataset Construction Pipeline Overview

Dataset Construction

Table 4: Details of CORI benchmark dataset. (✗) and (✓) denote unavailable and available data existent in XTREME benchmark respectively . MT, SEG, ROM correspond to Machine Translation, Pre-segmentation, Romanization processing respectively as described in Section 5. Y, N denote if the corresponding preprocessing is conducted for the dataset or not.

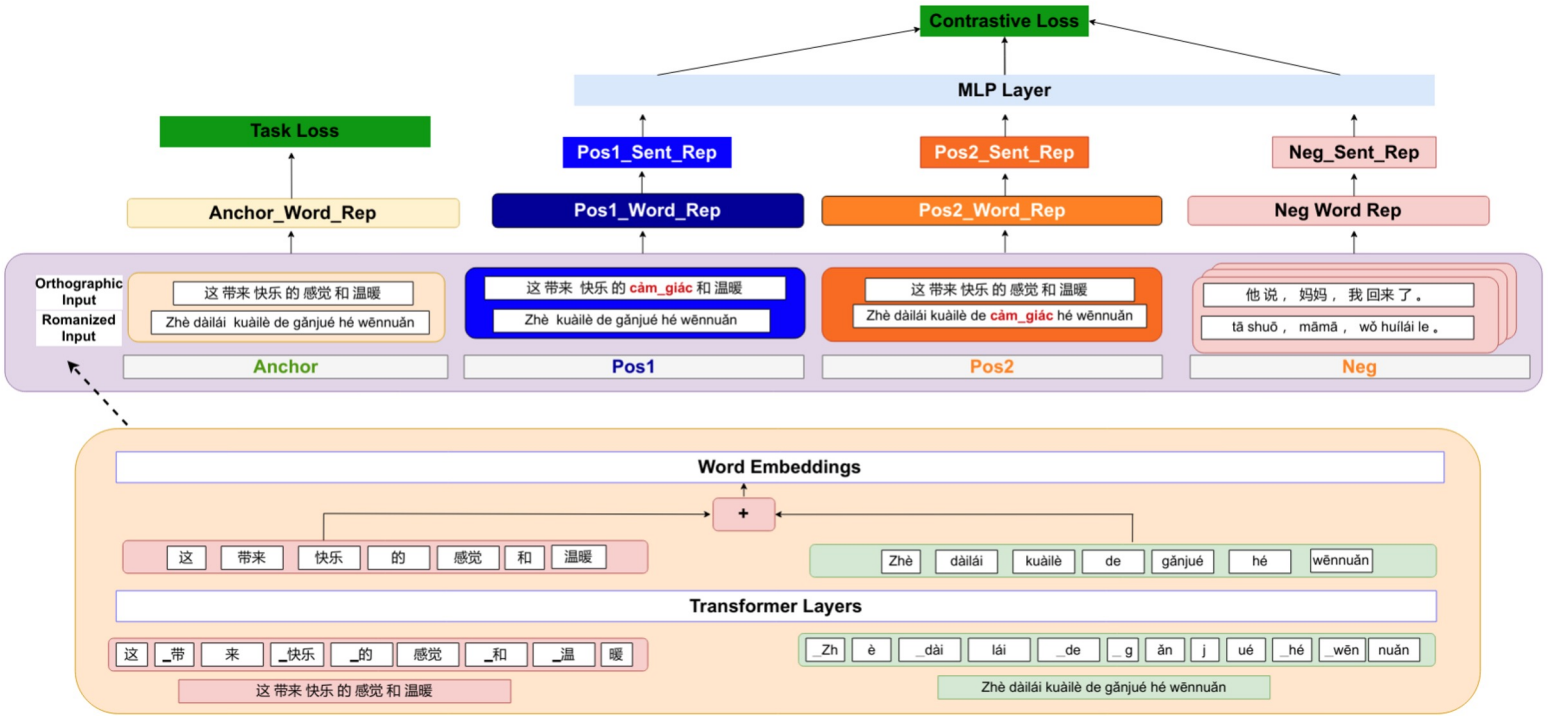
		MT	SEG	ROM	ZH		JA	KO	VI
					Train	Dev	Test	Test	Test
Sent-level	PAWSX	Y	Y	Y	49.4k (✗)	2k (✓)	2k (✗)	2k (✓)	2k (✗)
	XNLI	Y	Y	Y	392.7k (✓)	2.49k (✓)	5.01k (✗)	5.01k (✗)	5.01k (✓)
Token-level	UDPOS	N	N	Y	10k (✓)	1.6k (✓)	2.5k (✓)	4.7k(✓)	0.8k (✓)
	PANX	N	Y	Y	20k (✓)	10k (✓)	10k (✓)	10k (✓)	10k (✓)
Question Answering	XQuAD	Y	Y	Y	80.1k (✗)	8.87k (✗)	1.19k (✗)	1.19k (✗)	1.19k (✓)
	MLQA	Y	Y	Y	80.1k (✗)	8.87k (✗)	11.24 k (✗)	11.24k (✗)	11.24k (✓)

➡ **CORI addresses the presented challenges from multilingual XTREME benchmark**

1. **Language Availability**
2. **Pre-segmentation**
3. **Romanization**

FRAMEWORK

Framework: Overview

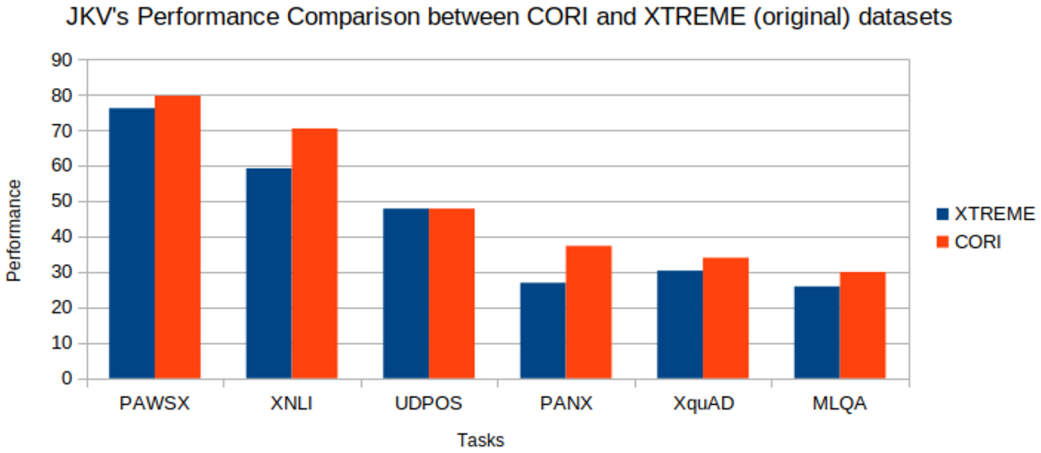


EVALUATION

Evaluation

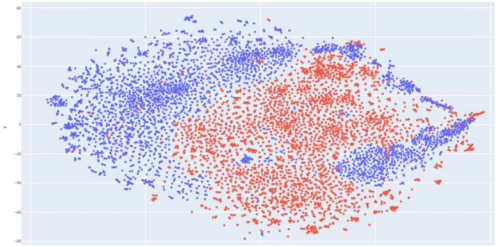
Observation:

- Higher performance on CORI than XTREME benchmark dataset
- No difference in performance on UDPOS task performance since no MT or SEG is applied to improve quality of the dataset.

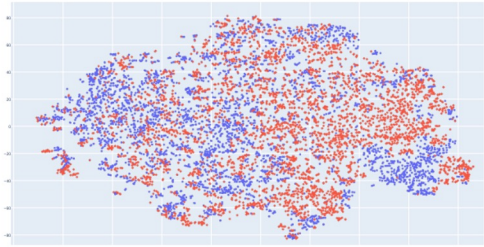


➡ **Proposed preprocessing steps for CORI are effective**

Evaluation



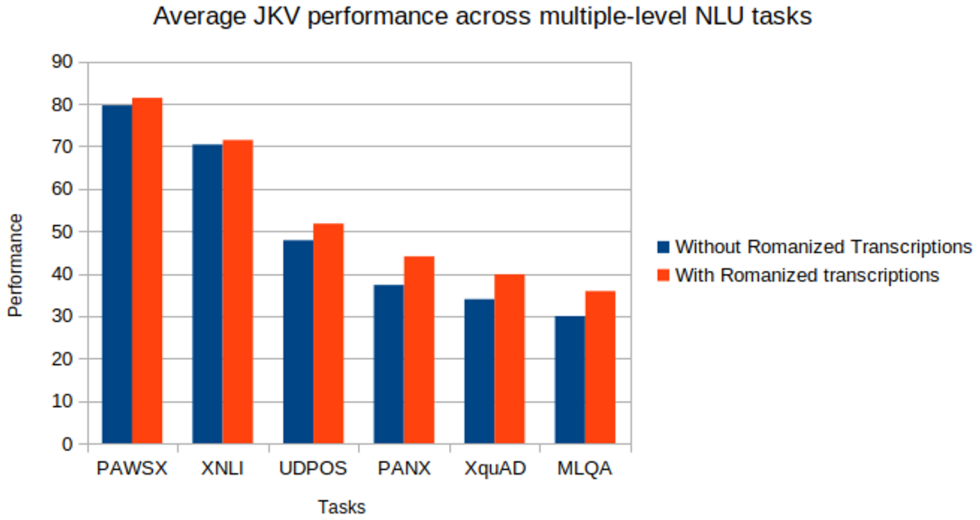
(a) Without Romanized transcriptions



(b) With Romanized transcriptions

Observation:

- Romanized transcription enhances the textual representation across CJKV languages
- Leading to improvements of downstream multi-level tasks across target JKV languages

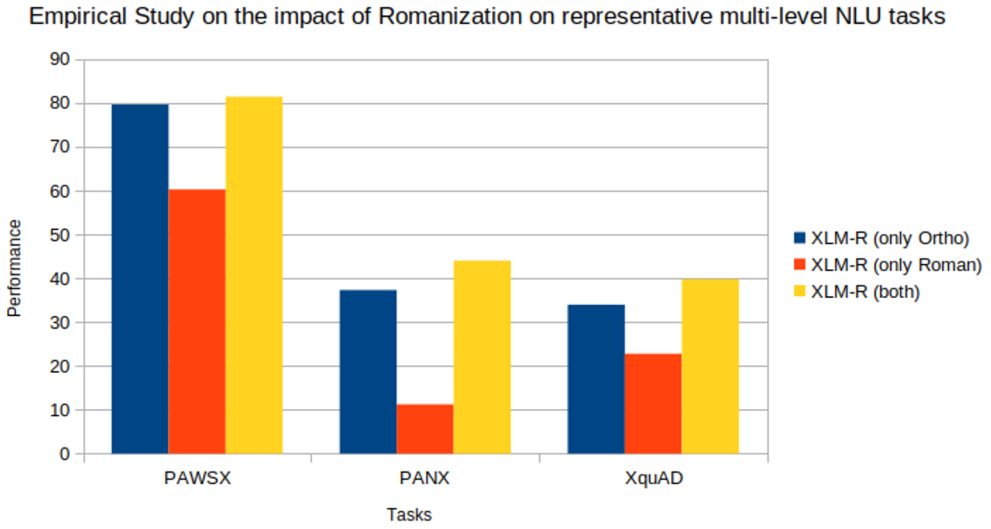


Romanization provides additional helpful signals for cross-lingual transfer

Evaluation

Observation:

- Romanization provides helpful information for downstream tasks for target JKV languages
- Relying purely on Romanization is not sufficient for cross-lingual transfer on text-based LMs



Romanization is an essential addendum, not replacement, for the orthographic representation

CONCLUSION

CONCLUSION

- Choice of source language is essential to downstream task performance for target languages
- **CORI: CJKV-specific dataset** addresses the limitations of current multilingual benchmark datasets
- **Romanization**, one type of phonemic signals, is valuable for cross-lingual transfer beyond the limitations of textual scripts

Thank you for your attendance