

Script and Orthographic Normalization

The Pandora's Box of Low-Resource Language Technology

Sina Ahmadi
George Mason University

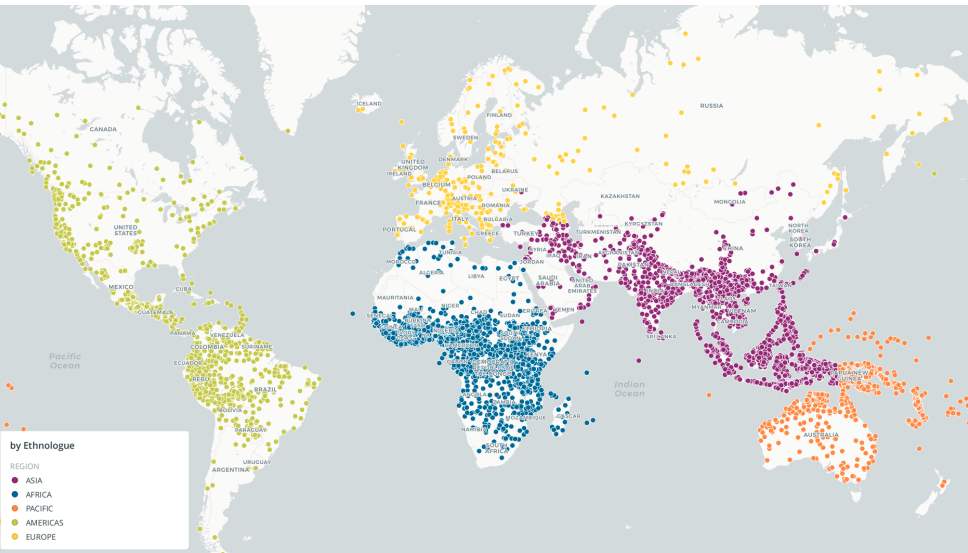
ContribuLing - INALCO
May 12, 2023



Context

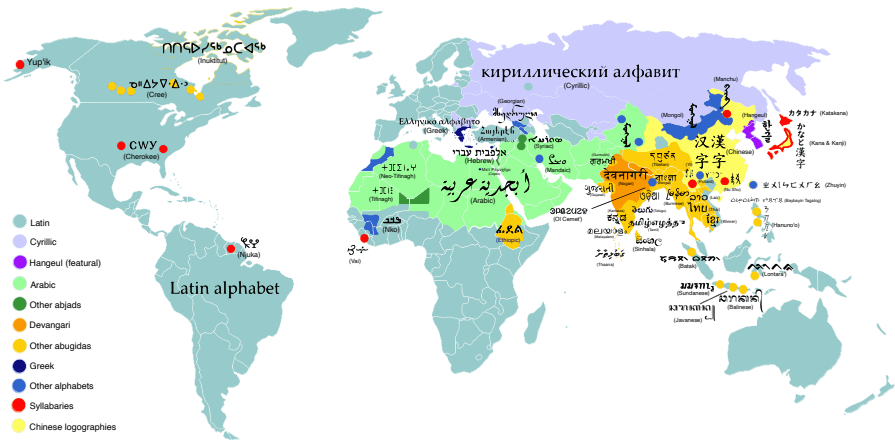
Context: Languages and Writing Systems

- More than 7,000 “languages” are spoken (Ethnologue, 2023).



Context: Languages and Writing Systems

- More than 7,000 “languages” are spoken (Ethnologue, 2023).
- Almost 300 writing systems exist (and many adopted ones)
- Less than 4,000 languages have a written form

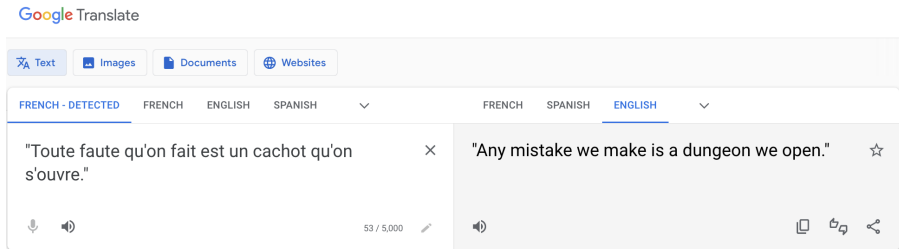


Context: Language is Text!

- Resources, tools & applications in NLP heavily rely on text, e.g.:

Context: Language is Text!

- Resources, tools & applications in NLP heavily rely on text, e.g.:
- Machine translation



The screenshot shows the Google Translate interface. At the top, the Google Translate logo is visible. Below it, there are four tabs: 'Text', 'Images', 'Documents', and 'Websites'. The 'Text' tab is selected. The interface is split into two columns. The left column shows the source text in French: "Toute faute qu'on fait est un cachot qu'on s'ouvre." Below the text are icons for voice input and output, and a character count of "53 / 5,000". The right column shows the translated text in English: "Any mistake we make is a dungeon we open." Below the text are icons for voice output, copy, share, and a star for bookmarks. The language selection dropdowns at the top of the columns show 'FRENCH - DETECTED' and 'ENGLISH' selected.

Context: Language is Text!

- Resources, tools & applications in NLP heavily rely on text, e.g.:

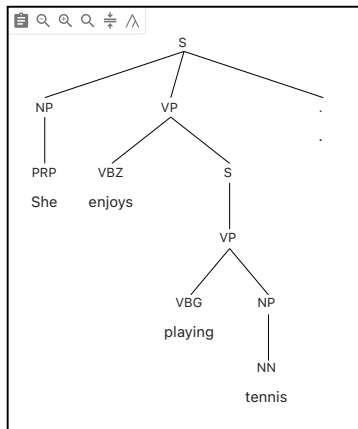
- Machine translation
- Syntactic parsing

Sentence:

Berkeley Neural Parser

She enjoys playing tennis.

Parse tree:



Context: Language is Text!

- Resources, tools & applications in NLP heavily rely on text, e.g.:

- Machine translation
- Syntactic parsing
- Large language models



Écris un poème décrivant Wikimedia France

GPT-3.5



En France, une organisation est née
Wikimédia est son nom, une bibliothèque
Virtuelle, gratuite, pour tous les curieux
Découvrir, apprendre, échanger, audacieux

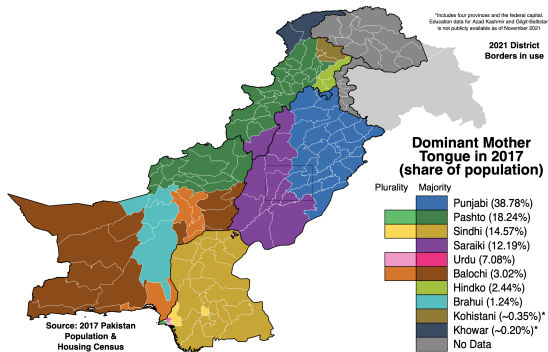
Leur mission est de partager
Les connaissances, les cultures, les idées
De rendre accessibles les savoirs
Afin que chacun puisse s'en emparer

En son sein, Wikimedia France
S'engage pour que chacun avance
Leur dévouement et leur passion
Sont à l'image de leur mission

Context: Language Communities

Most countries are X-lingual, but not all officially!

- Pakistan:
→ Urdu and English



Context: Language Communities

Most countries are X-lingual, but not all officially!

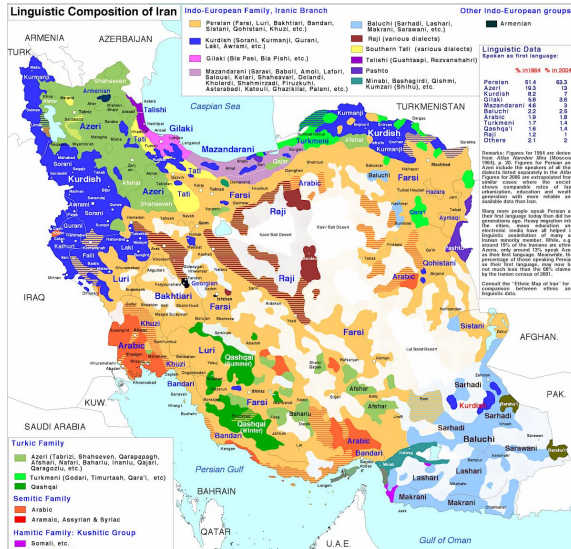
- Pakistan:
→ Urdu and English
- India:
→ Hindi, Kashmiri,
Sindhi and 20 more



Context: Language Communities

Most countries are X-lingual, but not all officially!

- Pakistan:
→ Urdu and English
- India:
→ Hindi, Kashmiri, Sindhi and 20 more
- Iraq:
→ Arabic and Kurdish
- Iran:
→ Persian!



And what if these languages are written in the dominant language's script?



Unconventional Writing

Unconventional Writing

Unconventional writing

Unconventional writing refers to the usage of the **script of another language**, presumably that of a dominant language.

- Unsystematic writing

Unconventional Writing

Unconventional writing

Unconventional writing refers to the usage of the **script of another language**, presumably that of a dominant language.

- Unsystematic writing
- Not necessarily complying with orthography

Unconventional Writing

Unconventional writing

Unconventional writing refers to the usage of the **script of another language**, presumably that of a dominant language.

- Unsystematic writing
- Not necessarily complying with orthography
- Impact of donor language on code switching

Unconventional Writing

Unconventional writing

Unconventional writing refers to the usage of the **script of another language**, presumably that of a dominant language.

- Unsystematic writing
- Not necessarily complying with orthography
- Impact of donor language on code switching
- No specific rule for mapping graphemes-phonemes

Unconventional Writing

Unconventional writing

Unconventional writing refers to the usage of the **script of another language**, presumably that of a dominant language.

- Unsystematic writing
- Not necessarily complying with orthography
- Impact of donor language on code switching
- No specific rule for mapping graphemes-phonemes

A few examples:

Unconventional Writing

Unconventional writing

Unconventional writing refers to the usage of the **script of another language**, presumably that of a dominant language.

- Unsystematic writing
- Not necessarily complying with orthography
- Impact of donor language on code switching
- No specific rule for mapping graphemes-phonemes

A few examples:

- *ana raye7 el gam3a el sa3a 3 el 3asr.* (Arabic in Latin script, aka Arabizi)

Unconventional Writing

Unconventional writing

Unconventional writing refers to the usage of the **script of another language**, presumably that of a dominant language.

- Unsystematic writing
- Not necessarily complying with orthography
- Impact of donor language on code switching
- No specific rule for mapping graphemes-phonemes

A few examples:

- *ana raye7 el gam3a el sa3a 3 el 3asr.* (Arabic in Latin script, aka Arabizi)
- *Tora ti na sou po* (Greek in Latin, aka Greeklish)

Unconventional Writing

Unconventional writing

Unconventional writing refers to the usage of the **script of another language**, presumably that of a dominant language.

- Unsystematic writing
- Not necessarily complying with orthography
- Impact of donor language on code switching
- No specific rule for mapping graphemes-phonemes

A few examples:

- *ana raye7 el gam3a el sa3a 3 el 3asr.* (Arabic in Latin script, aka Arabizi)
- *Tora ti na sou po* (Greek in Latin, aka Greeklish)
- *mer6 pr tn mess pr mn anif* (French SMS language)

Unconventional Writing: Perso-Arabic scripts



Unconventional Writing: Perso-Arabic scripts

- Traditionally used for Arabic
- Used for over a millennium
- A *Reichssprache* for centuries

Arabic



أ ب ت ج
ح خ د ذ ر ز
س ش ع غ
ف ق ل م ن
ه و ي
آ إ أ ث و
ص ض ط
ي ة

Unconventional Writing: Perso-Arabic scripts

- Used for writing in more than 20 languages/varieties



Unconventional Writing: Perso-Arabic scripts

- Used for writing in more than 20 languages/varieties
- Persian, Urdu, Kurdish, Uyghur, Kashmiri etc.



Unconventional Writing: Perso-Arabic scripts

- Used for writing in more than 20 languages/varieties
- Persian, Urdu, Kurdish, Uyghur, Kashmiri etc.
- 400M speakers in the Middle East and the Subcontinent



Unconventional Writing: Perso-Arabic scripts

- Used for writing in more than 20 languages/varieties
- Persian, Urdu, Kurdish, Uyghur, Kashmiri etc.
- 400M speakers in the Middle East and the Subcontinent
- Discriminatory language policies (Sheyholislami, 2012)
→ pernicious sociolinguistic effects on language attitudes



Unconventional Writing: Perso-Arabic scripts

Language	Unconventional script	Unconventional writing	Conventional writing
Gilaki	Persian	یتہ زون نم ہیسہ گہ گیلکن اون جی گب زنن	یتہ زوؤن؎ نؤم ہیسہ گہ گیلکؤن اؤن؎ جی گب زنن
Kashmiri	Urdu	برور چھ اکھ وراسے جانور۔	برور چھ اکھ وراسے جانور۔
Kurmanji	Arabic	قایمقام الامدی بئرثوا پارزکار دھوک دا	قایمقامن تامیدیین بھرسقا پاریزگاری دھوکی دا
Sorani	Arabic	ھەر لہ یەکەم شانۆوە دیارە فەھدیان دەوێت	ھەر لہ یەکەم شانۆوە دیارە فەھەدیان دەوێت
Sindhi	Urdu	مدیني ڈانھن ھجرت وقت فقط ھي ۽ گھرواري سان گڏ ھئي	مدیني ڈانھن ھجرت وقت فقط ھي ۽ گھرواري سان گڏ ھئي

Unconventional Writing: the Pandora's Box



Unconventional Writing: Questions

Some questions to think about:

- 1 How does unconventional writing affect NLP?

Unconventional Writing: Questions

Some questions to think about:

- ① How does unconventional writing affect NLP?
- ② How can this phenomenon be effectively remediated?

Unconventional Writing: Questions

Some questions to think about:

- 1 How does unconventional writing affect NLP?
- 2 How can this phenomenon be effectively remediated?
- 3 Do we always need to write a language?

Script Normalization

Script Normalization

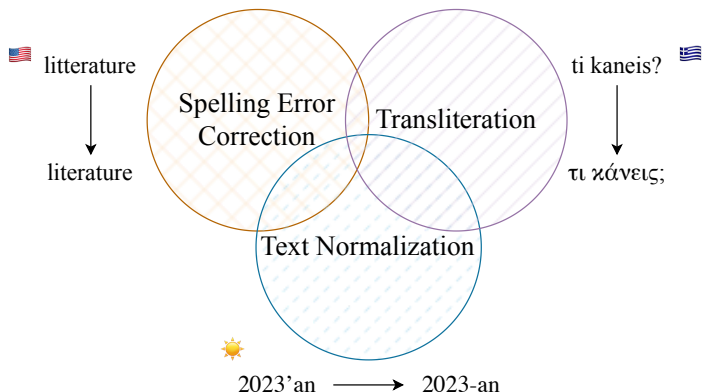
Script Normalization

Normalization of a text written in an unconventional script based on the conventional script and orthography

Script Normalization

Script Normalization

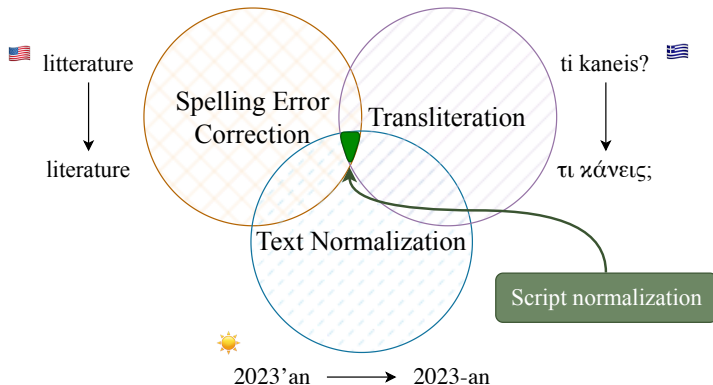
Normalization of a text written in an unconventional script based on the conventional script and orthography



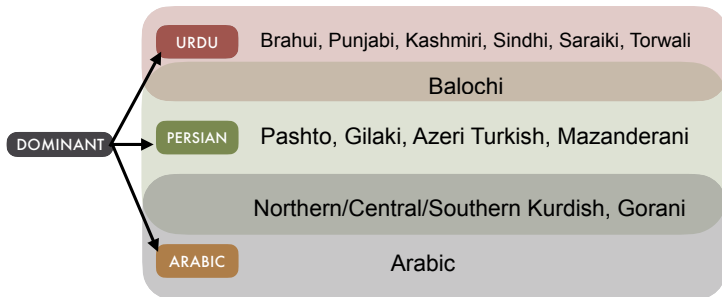
Script Normalization

Script Normalization

Normalization of a text written in an unconventional script based on the conventional script and orthography



Script Normalization: Selected Languages



Script Normalization: Approach

1 Data collection – Not easy!

Language	639-3	WP	script type	diacritics	ZWNJ	Dominant
Azeri Turkish	azb	azb	Abjad	✓	✓	Persian
Kashmiri	kas	ks	Alphabet	✓	✗	Urdu
Gilaki	glk	glk	Abjad	✓	✓	Persian
Gorani	hac	-	Alphabet	✗	✗	Persian, Arabic, Sorani
Kurmanji	kmr	-	Alphabet	✗	✗	Persian, Arabic
Sorani	ckb	ckb	Alphabet	✗	✗	Persian, Arabic
Mazanderani	mzn	mzn	Abjad	✓	✓	Persian
Sindhi	snd	sd	Abjad	✓	✗	Urdu
Persian	fas	fa	Abjad	✓	✓	-
Arabic	arb	ar	Abjad	✓	✗	-
Urdu	urd	ur	Abjad	✓	✓	-

Script Normalization: Approach

- 1 Data collection – Not easy!
- 2 Script mapping
 - Common characters
 - Visual resemblance
 - Orthographic rules

Language	Unconventional script	Source	Target
Azeri Turkish	Persian	چ	چ
Sorani	Arabic	ز	ذ / ض / ظ / ز
Kashmiri	Urdu	اُ	اُ / ا
Sindhi	Urdu	ي	ے / ي / ی

Script Normalization: Approach

- 1 Data collection – Not easy!
- 2 Script mapping
 - Common characters
 - Visual resemblance
 - Orthographic rules
- 3 Character-alignment matrix
→ sequence alignment based on dictionaries

	Sorani to Arabic
▼ ئ:	
_:	0.9829
ئ:	1.0066
ا:	1
▼ ا:	
ا:	1.9559
▼ ب:	
ب:	1.999
▼ د:	
د:	2
▼ ی:	
ي:	1.9222000000000001
▼ ت:	
ت:	1.7437
ط:	1.1711

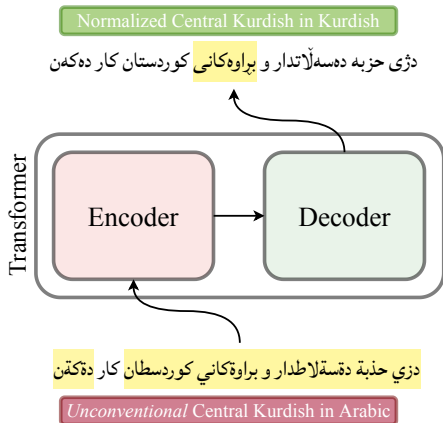
Script Normalization: Approach

- 1 Data collection – Not easy!
- 2 Script mapping
 - Common characters
 - Visual resemblance
 - Orthographic rules
- 3 Character-alignment matrix
 - sequence alignment based on dictionaries
- 4 Synthetic data generation
 - randomly generate pairs
 - inject noise

Noise %	Sentence
Clean	دووهمین پيشانگهها فوتوگرافهريڻ كورد ل بهلجيجا Second Kurdish photographers' exhibition in Belgium
20	دووهمين پيشانگهها فوتوگرافهريڻ كورد ل بهلجيجا
40	دووهمين پيشانگهها فطگرافه رن كورد ل بهلجيجا
60	دووهمين پيشانگهها فوتوگرافه رن كورد ل بهلجيجا
80	دووهمين پيشانگهها فوتوگرافهريڻ كورد ل بهلجيجا
100	دووهمين پيشانگهها فوتوگرافهريڻ كورد ل بهلجيجا

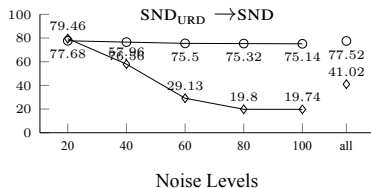
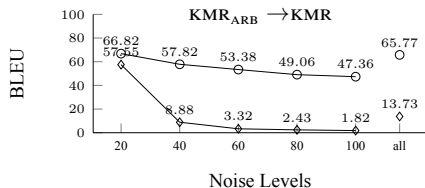
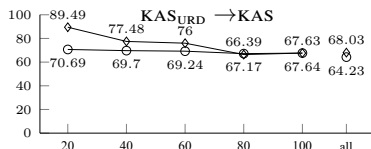
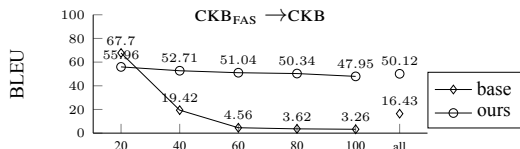
Script Normalization: Approach

- 1 Data collection – Not easy!
- 2 Script mapping
 - Common characters
 - Visual resemblance
 - Orthographic rules
- 3 Character-alignment matrix
 - sequence alignment based on dictionaries
- 4 Synthetic data generation
 - randomly generate pairs
 - inject noise
- 5 Model
 - encoder-decoder with self-attention



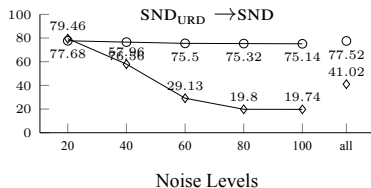
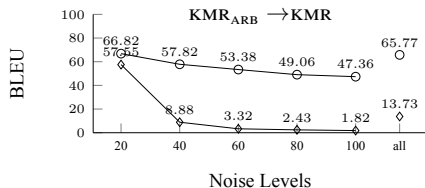
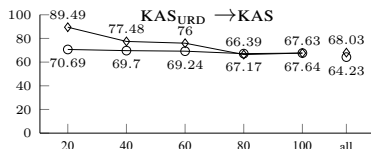
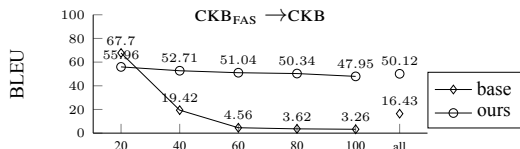
Script Normalization: Intrinsic Experiments

- Baseline: a naive “copy” system



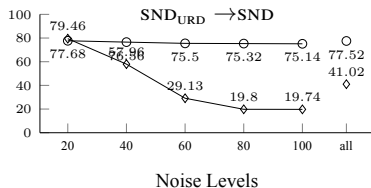
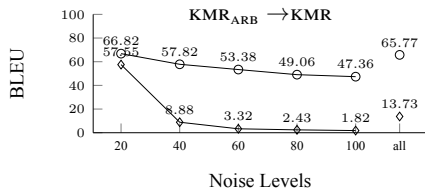
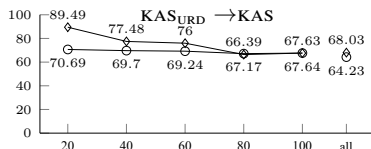
Script Normalization: Intrinsic Experiments

- Baseline: a naive “copy” system
- Ours: trained models on different levels of noise



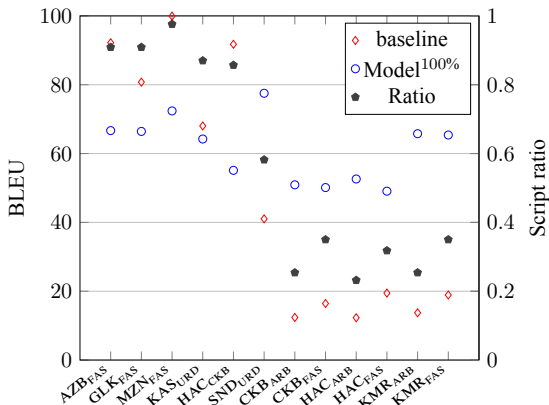
Script Normalization: Intrinsic Experiments

- Baseline: a naive “copy” system
- Ours: trained models on different levels of noise
- **Our models dramatically improve over the baseline**



Script Normalization: Intrinsic Experiments

- Baseline: a naive “copy” system
- Ours: trained models on different levels of noise
- **Our models dramatically improve over the baseline**
- The more similar the scripts, the more difficult the normalization!



Script Normalization: Extrinsic Experiments

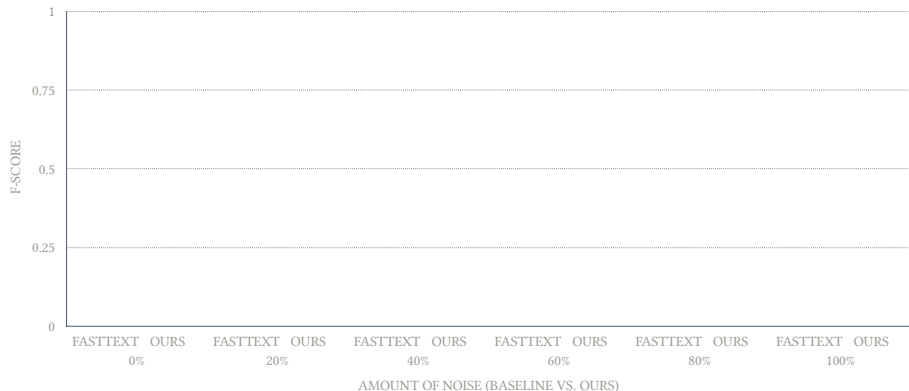
1 Language identification (LID)

Sentence	Language
اور لادینیت واشتراکیت کو جمہوریت کے حسین لبادہ میں پیش کردیا گیا۔	Punjabi
کہیں وی زبان وادب تے تحقیق زیادہ تر کیفیتی	Saraiki
گھٹا دفعا ھک عورت سائیاٹی جنھن سان کوئی افلاطونی	Sindhi
آیانی رابا کہ تئی مھر بوتگ أنت گنج گوار	Balochi
قوزئی و دوغو سوریه موختار ایداره ائتمه سی	Azeri
شوراب ایسم ایته روستا ایسه جه راستوی دهستان	Gilaki
جوانی زمان فرا گرفتن دانایی است. پیری زمان تمرین کردن آن است.	Persian

Script Normalization: Extrinsic Experiments

1 Language identification (LID)

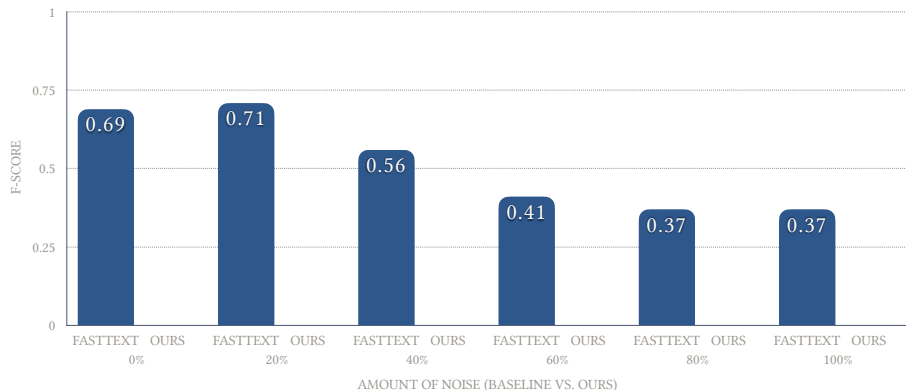
- Compare LID with and without normalization



Script Normalization: Extrinsic Experiments

① Language identification (LID)

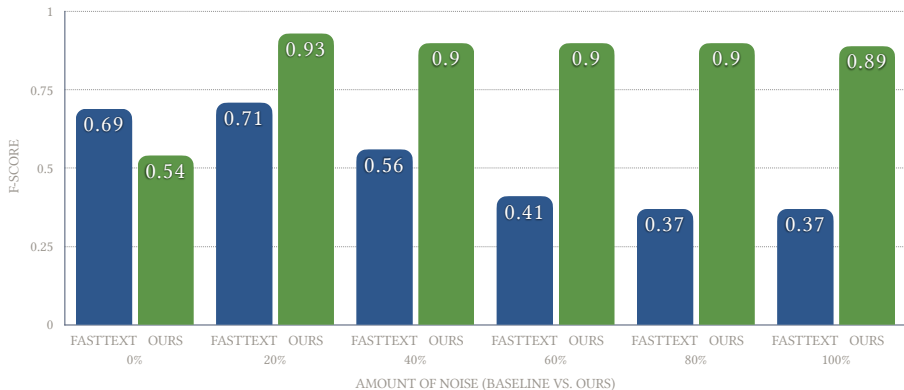
- Compare LID with and without normalization
- Terrible performance by any existing model



Script Normalization: Extrinsic Experiments

1 Language identification (LID)

- Compare LID with and without normalization
- Terrible performance by any existing model
- Models trained on normalized datasets improve the F-scores



Script Normalization: Extrinsic Experiments

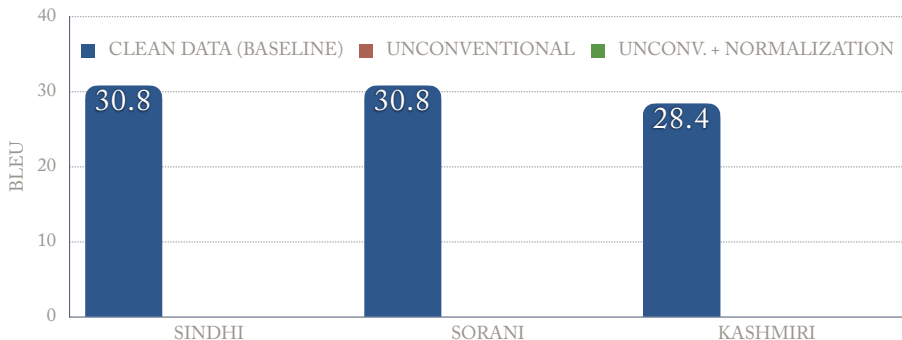
1 Language identification (LID)

- Compare LID with and without normalization
- Terrible performance by any existing model
- Models trained on normalized datasets improve the F-scores
- Closely-related languages (scripts) are confused!

Azeri	1383	312	301	0	0	98	114	112	6	2	0
Gilaki	98	1116	107	1	1	75	98	83	6	26	0
Mazanderani	126	149	1199	1	19	85	89	70	18	11	0
Arabic	1	2	1	2368	10	3	0	0	0	4	5
Persian	0	5	13	10	2366	0	0	0	2	0	3
Gorani	215	222	199	9	4	1327	330	321	0	12	0
Sorani	320	327	325	3	0	483	1336	477	0	8	0
Kurmanji	170	170	156	0	0	239	301	1223	2	10	2
Kashmiri	4	4	5	2	0	2	5	5	2086	46	1
Sindhi	83	93	94	5	0	87	125	108	18	2271	5
Urdu	0	0	0	1	0	0	2	1	2	10	2383
	Azeri	Gilaki	Mazanderani	Arabic	Persian	Gorani	Sorani	Kurmanji	Kashmiri	Sindhi	Urdu

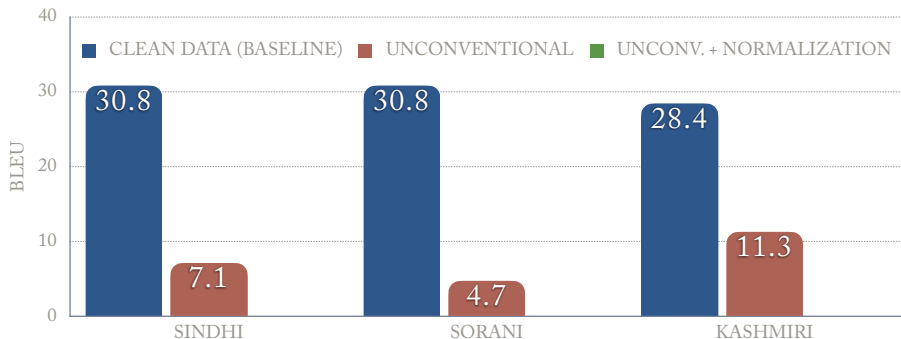
Script Normalization: Extrinsic Experiments

- 1 Language identification (LID)
- 2 **Machine Translation (MT)**
 - Evaluate MT with and without normalization



Script Normalization: Extrinsic Experiments

- 1 Language identification (LID)
- 2 **Machine Translation (MT)**
 - Evaluate MT with and without normalization
 - Terrible performance on noisy data (NLLB as baseline)

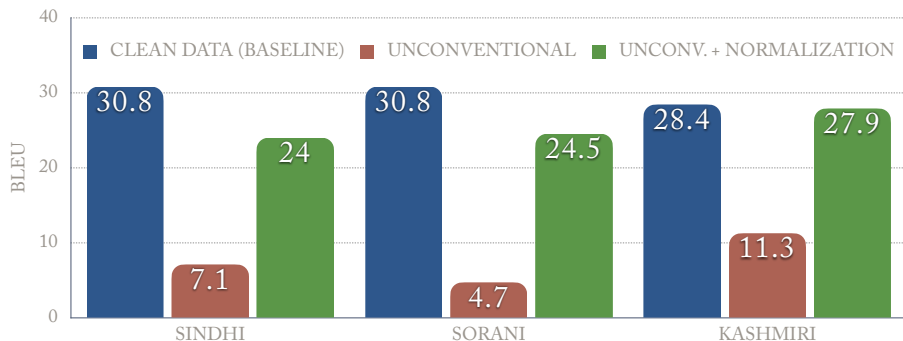


Script Normalization: Extrinsic Experiments

1 Language identification (LID)

2 Machine Translation (MT)

- Evaluate MT with and without normalization
- Terrible performance on noisy data (NLLB as baseline)
- Models trained on normalized datasets improve the F-scores



Conclusion

Key Takeaways

- 1 Unconventional writing is more widespread than thought of.

Key Takeaways

- 1 Unconventional writing is more widespread than thought of.
- 2 It is an open problem and non-trivial to solve.

Key Takeaways

- 1 Unconventional writing is more widespread than thought of.
- 2 It is an open problem and non-trivial to solve.
- 3 It negatively affects NLP for low-resourced languages.

Key Takeaways

- 1 Unconventional writing is more widespread than thought of.
- 2 It is an open problem and non-trivial to solve.
- 3 It negatively affects NLP for low-resourced languages.
- 4 We can effectively remediate it, *but only to some extent...*

Key Takeaways

- 1 Unconventional writing is more widespread than thought of.
- 2 It is an open problem and non-trivial to solve.
- 3 It negatively affects NLP for low-resourced languages.
- 4 We can effectively remediate it, *but only to some extent...*
- 5 Do we always need to write a language?
 - Multi-modal NLP
 - Multi-lingual NLP
 - Multi-task NLP
 - Better adaptation in NLP

We don't know yet.

Any questions?

Kiitos धन्यवाद Köszönöm
Rahmat Tak

شكراً 谢谢 Gracias

Спасибо

Adúpé

Hvala

Mulțumesc

Paldies

Bedankt

Dzięki

ขอบคุณ

Obrigado

koe

Shurkan

Danke Diolch

Danko

Daalu

ơn

Cảm

Grazie

Teşekkürler

감사합니다

Tānan

Merci



ευχαριστώ Mahalo

References I

-  Ahmadi, Sina, Milind Agarwal, and Antonios Anastasopoulos (May 2023). “PALI: A Language Identification Benchmark for Perso-Arabic Scripts”. In: *Proceedings of the 10th Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*. Dubrovnik, Croatia: The 17th Conference of the European Chapter of the Association for Computational Linguistics.
-  Ahmadi, Sina and Antonios Anastasopoulos (2023). “Script Normalization for Unconventional Writing of Under-Resourced Languages in Bilingual Communities”. In: Toronto, Canada: The 61st Annual Meeting of the Association for Computational Linguistics.
-  Sheyholislami, Jaffer (2012). “Kurdish in Iran: A case of restricted and controlled tolerance”. In: *International Journal of the Sociology of Language* 2012.217, pp. 19–47.