Informal Romanization across Languages and Scripts

Maria Ryskina

Carnegie Mellon University
Language Technologies Institute

Non-standard language & NLP

From standardized language…

```
(S (NP-SBJ (NN Compound)
           (NNS yields))
   (VP (VBP assume)
       (UCP (NP (NP (NN reinvestment))
                (PP (IN of)
                    (NP (NNS dividends))))
            (CC and)
            (SBAR (IN that)
                  (S (NP-SBJ (DT the)
                             (JJ current)
                             (NN yield))
                     (VP (VBZ continues)
                         (PP-TMP (IN for)
                                 (NP (DT a)
                                     (NN year))))))))
   (. .)
)
```

e.g. PTB:

- Newswire
- Finance-related
- Formal

…To creative language

- Variety of genres
- Variety of domains
- Spectrum of formality

https://readandsurvive.com/2014/07/09/how-to-stop-evil-malchiks/

Linguistic innovation: What

- Non-standard, novel linguistic items…
  - Lexical: new word forms *(brony)*
  - Morphological: new morphemes *(-gate)* or derivatives *(prolifeness)*
  - Orthographic: non-standard spellings *(2nite)*
- … before they become attested *(tweet)*
- People can infer their meaning, but NLP systems largely treat them as noise

Linguistic innovation: How

- Q1: How do people process non-standard items?
  - Shared knowledge or perception: *2 = 'two'* = /tu/
  - Compositionality: *2nite = 'two' + 'nite'* = /tu/ + /naɪt/ ≈ /tənaɪt/
    antivehicleness= 'anti' + 'vehicle' + 'ness'
- Q2: How can we get our NLP systems to that level?
  - Text normalization: *2nite* → *tonight* (Baldwin et al., 2015: W-NUT shared task)
  - Improving robustness to noise & 'noise' (Li et al., 2019: WMT shared task)
  - Maybe we can encode creative reasoning into them?

Linguistic innovation: Why

- Creativity: expressing extralinguistic information

- Necessity: bypassing constraints or lack of tools

https://www.myfonts.com/fonts/vic-fieger/kremlin-pro-new/

https://ktonanovenkogo.ru/image/krakozyabry-kodirovka.png

https://www.lovelylanguage.ru/for-kids/alphabet/499-english-transliteration

Informal romanization

- *Romanization:* rendering non-Latin-script languages in Latin alphabet
- *Informal:* used online, arises out of Unicode/keyboard issues

Russian	*человек*	*chelovek, 4elovek, ceJIoBek, …*
Arabic	صباح	*saba7, sba7, sabah, …*
Greek	*ξένος*	*xenos, ksenos, 3enos, …*

Informal romanization

- Idiosyncratic representation: character subsititutions up to the user
  - Substitution choices can convey social meaning (Nguen et al., 2021)

Russian	*человек*	*chelovek, 4elovek, ceJloBek, …*
Arabic	صباح	*saba7, sba7, sabah, …*
Greek	*ξένος*	*xenos, ksenos, 3enos, …*

Informal romanization

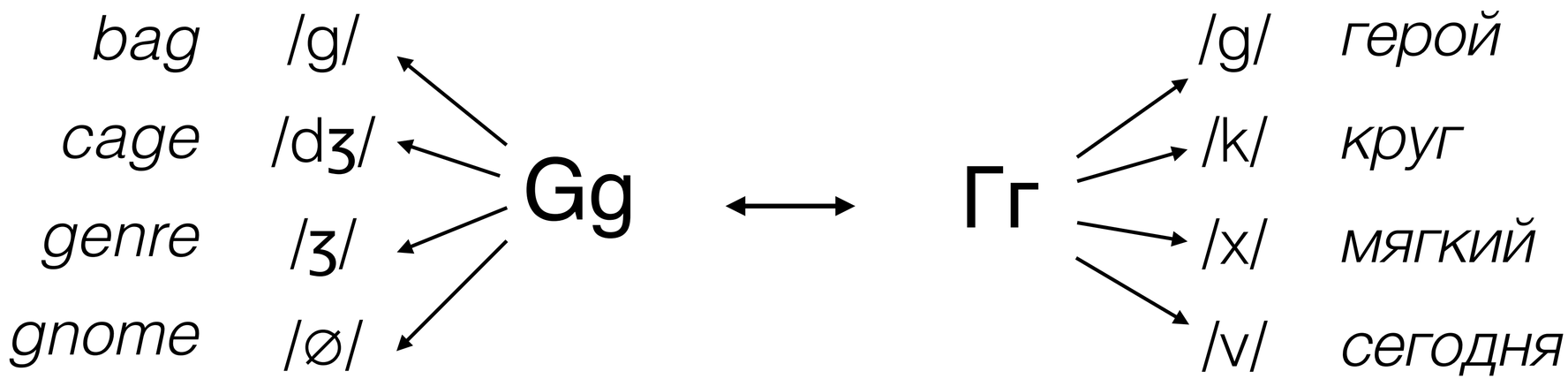
- Idiosyncratic representation: character subsititutions up to the user
  - Substitution choices can convey social meaning (Nguen et al., 2021)

- Most substitutions are based on phonetic or visual similarity
  [but not all: ь→’, θ→u (Chalamandaris et al., 2006)]

Russian	*человек*	*chelovek, 4elovek, сeJloBek, …*
Arabic	صباح	*saba7, sba7, sabah, …*
Greek	*ξένος*	*xenos, ksenos, 3enos, …*

Phonetic romanization

- What does it mean for two characters to be phonetically similar?



- This is just in one language each!

Phonetic romanization

- What does it mean for two characters to be phonetically similar?
- Out-of-context grapheme-phoneme association: ɾ~/g/→g

Every letter makes a sound:
‘A’ says /eɪ/!*

*and /a/

https://www.youtube.com/watch?v=FFhH6576psw

Phonetic romanization

- What does it mean for two characters to be phonetically similar?
- Out-of-context grapheme-phoneme association: ɾ~/g/→g
- Phoneme produced in context: انتي /enti/→enty, صباح /sabaħ/→saba7

Visual romanization

- Broad similarity between glyph shapes
- Single characters can map to bi-/trigraphs
- Can be conditioned on a transformation
- Can be applied to a part of a glyph

а~/a/→a, г~/g/→r

ы→bl, ж→}|{

ɛ →3, л→v

ӈ →2

Character alignment

- Monotonic alignment that depends on the writing system of the language

Alphabet	Abjad (consonantal)	Abugida (alphasyllabary)
хорошо	كريم	ಬೆಳಗಿತು
		
xorosho	kareem	belagitu
~ one-to-one	~ one-to-one + null	~ one-to-many

Task framing

- Convert romanized text to the conventional orthography of the language

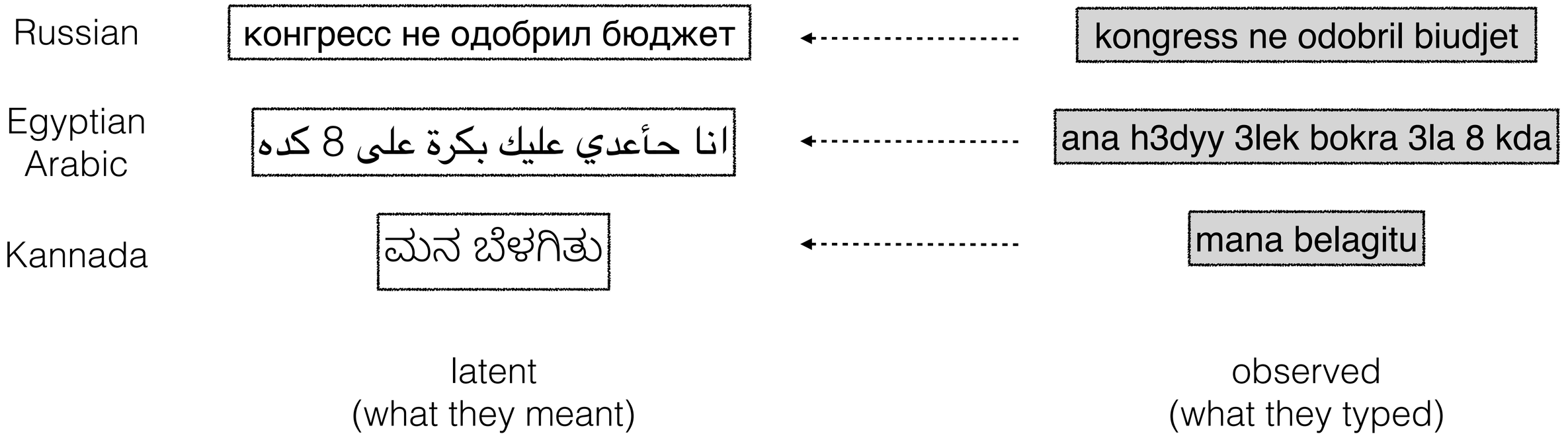
Russian	конгресс не одобрил бюджет	←	kongress ne odobril biudjet
Egyptian Arabic	انا حأعدي عليك بكرة على 8 كده	←	ana h3dyy 3lek bokra 3la 8 kda
Kannada	ಮನ ಬೆಳಗಿತು	←	mana belagitu

Outline

- **Part 2:** Unsupervised WFST (Ryskina et al., ACL 2020)
  - Proposed model and parameterization
  - Similarity-based inductive bias
- **Part 3:** Unsupervised seq2seq, error analysis (Ryskina et al., SIGMORPHON 2021)
  - Seq2seq model and combining it with the WFST
  - Data and experimental results
  - Comparative error analysis

Task framing

- Convert romanized text to the conventional orthography of the language



Task framing

- Parallel data does not occur naturally ⇒ unsupervised learning

- Perceptions of similarity are shared across users and even languages!
  - But does it mean we can train language-independent models? ✗

Linguistically Naïve != Language Independent:
Why NLP Needs Linguistic Typology

Emily M. Bender
University of Washington
Seattle, WA, USA
ebender@u.washington.edu

M Ryskina, MR Gormley, T Berg-Kirkpatrick. Phonetic and Visual Priors for Decipherment of Informal Romanization. ACL 2020.

Task framing

- Parallel data does not occur naturally ⇒ unsupervised learning
- Perceptions of similarity are shared across users and even languages!
  - But does it mean we can train language-independent models? ✗
- **Hypothesis**: inductive bias encoding these similarity notions provides signal that can somewhat approximate human supervision
  - We rely on manually-curated resources to operationalize it

M Ryskina, MR Gormley, T Berg-Kirkpatrick. Phonetic and Visual Priors for Decipherment of Informal Romanization. ACL 2020.

Decipherment

- Can be viewed as a decipherment task (Knight et al., 2006)

> …When I look at an article in Russian, I say: “This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.”
>
> — *Warren Weaver, 1947*

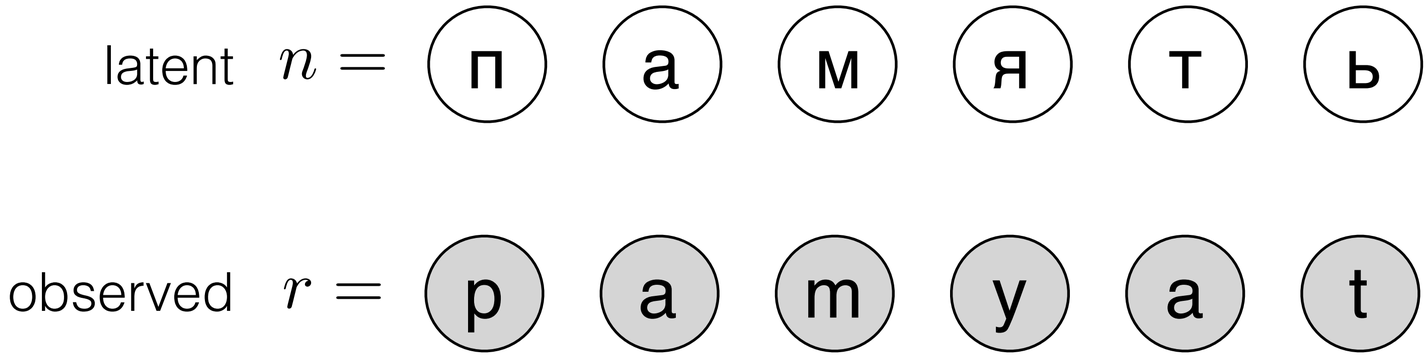
Decipherment

- Can be viewed as a decipherment task (Knight et al., 2006)

…When I look at an article in romanized Russian, I say: “This is really written in Cyrillic, but it has been coded in some strange symbols. I will now proceed to decode.”

— *Warren Weaver, 1947*

Noisy-channel model



$$p(r) = \sum_{n} p(n;\gamma) \cdot p(r|n;\theta) \cdot p_{\text{prior}}(\theta;\alpha)$$

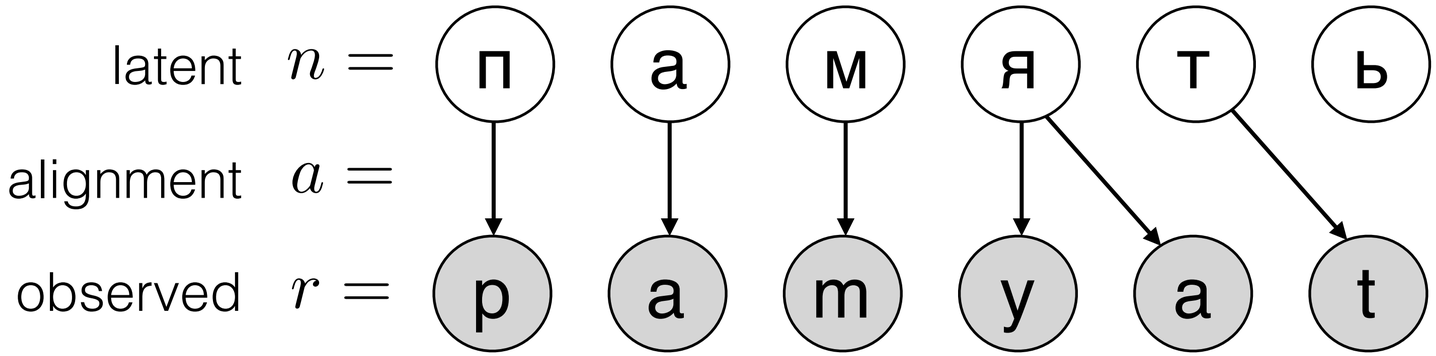
all possible native script sequences

transition probabilities

emission probabilities

prior on parameters

Noisy-channel model



$$p(r) = \sum_{n,a} p(n;\gamma) \cdot p(r|n,a;\theta) \cdot p_{\text{prior}}(\theta;\alpha)$$

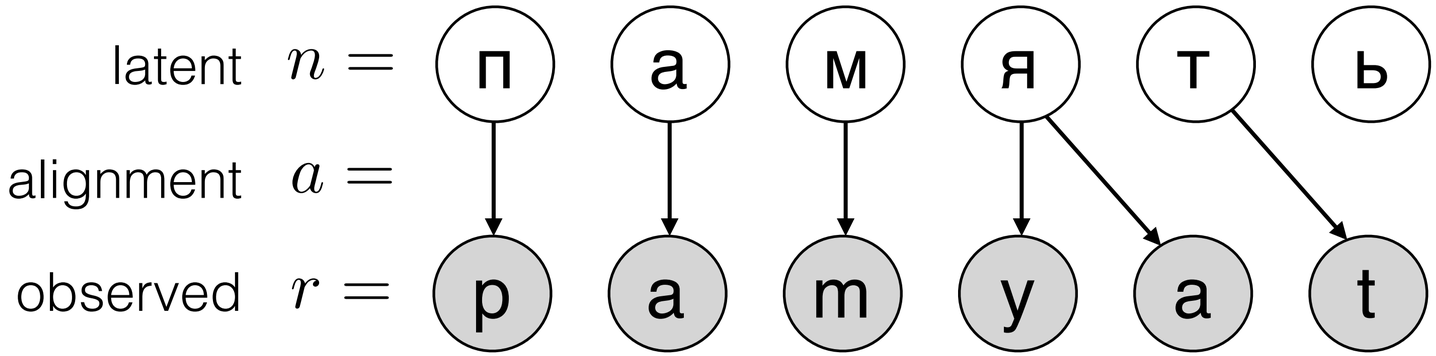
all possible native script sequences and alignments

transition probabilities

emission probabilities

prior on parameters

Noisy-channel model



$$p(r) = \sum_{n,a} p(n;\gamma) \cdot p(r|n,a;\theta) \cdot p_{\text{prior}}(\theta;\alpha)$$

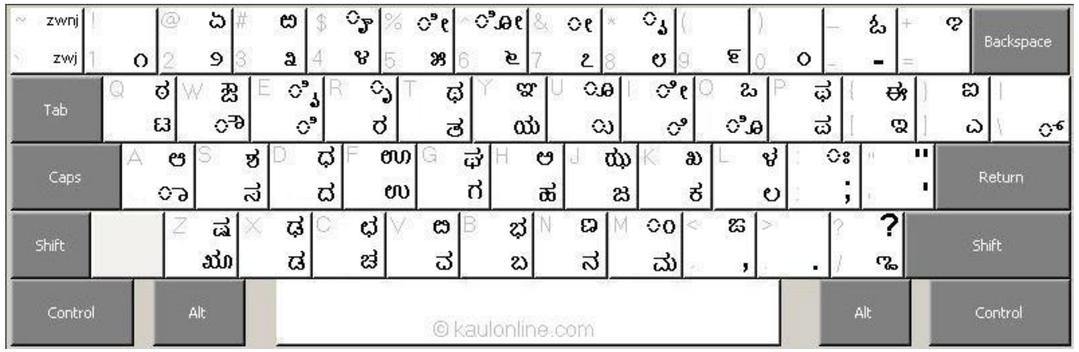
transition probabilities

emission probabilities

prior on parameters

Phonetic bias

- Phonetic priors: mappings off phonetic keyboard layouts

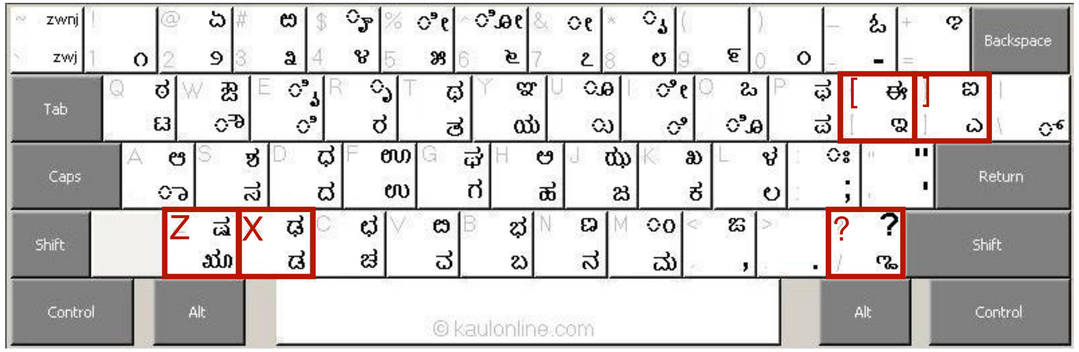


https://en.wikipedia.org/wiki/Phonetic_keyboard_layout,

 https://arabic.omaralzabir.com/, http://kaulonline.com/uninagari/kannada/

Phonetic bias

- Phonetic priors: mappings off phonetic keyboard layouts
  - One-to-one mapping constraints lead to spurious mappings



https://en.wikipedia.org/wiki/Phonetic_keyboard_layout,
https://arabic.omaralzabir.com/, http://kaulonline.com/uninagari/kannada/

Visual bias

- Visual priors: mappings off the Unicode confusables list

- Designed to combat spoofing attacks

y	ɣ	ʏ	γ	у	Ү	ყ
0079 LATIN SMALL LETTER Y	0263 LATIN SMALL LETTER GAMMA	028F LATIN LETTER SMALL CAPITAL Y	03B3 GREEK SMALL LETTER GAMMA	0443 CYRILLIC SMALL LETTER U	04AF CYRILLIC SMALL LETTER STRAIGHT U	10E7 GEORGIAN LETTER QAR
p	ρ	ϱ	р	⍴	ⲣ	𝗽
0070 LATIN SMALL LETTER P	03C1 GREEK SMALL LETTER RHO	03F1 GREEK RHO SYMBOL	0440 CYRILLIC SMALL LETTER ER	2374 APL FUNCTIONAL SYMBOL RHO	2CA3 COPTIC SMALL LETTER RO	1D429 MATHEMATICAL BOLD SMALL P

sigtyp.io

sigtyp.io

https://util.unicode.org/UnicodeJsps/confusables.jsp

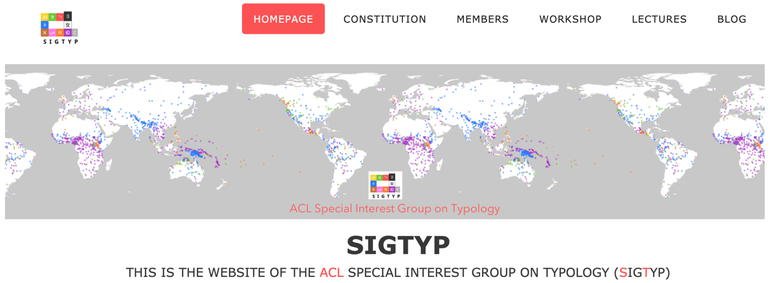
Visual bias

- Visual priors: mappings off the Unicode confusables list

- Designed to combat spoofing attacks

y	ɣ	ʏ	γ	у	Ү	ყ
0079 LATIN SMALL LETTER Y	0263 LATIN SMALL LETTER GAMMA	028F LATIN LETTER SMALL CAPITAL Y	03B3 GREEK SMALL LETTER GAMMA	0443 CYRILLIC SMALL LETTER U	04AF CYRILLIC SMALL LETTER STRAIGHT U	10E7 GEORGIAN LETTER QAR
p	ρ	ϱ	р	⍴	ᑭ	𝗽
0070 LATIN SMALL LETTER P	03C1 GREEK SMALL LETTER RHO	03F1 GREEK RHO SYMBOL	0440 CYRILLIC SMALL LETTER ER	2374 APL FUNCTIONAL SYMBOL RHO	2CA3 COPTIC SMALL LETTER RO	1D429 MATHEMATICAL BOLD SMALL P

sigtyp.io



sigtyp.io

The site you just tried to visit looks fake. Attackers sometimes mimic sites by making small, hard-to-see changes to the URL.

https://util.unicode.org/UnicodeJsps/confusables.jsp

Visual bias

- Visual priors: mappings off the Unicode confusables list

- Designed to combat spoofing attacks

- Hardly any mappings for Arabic and Kannada!

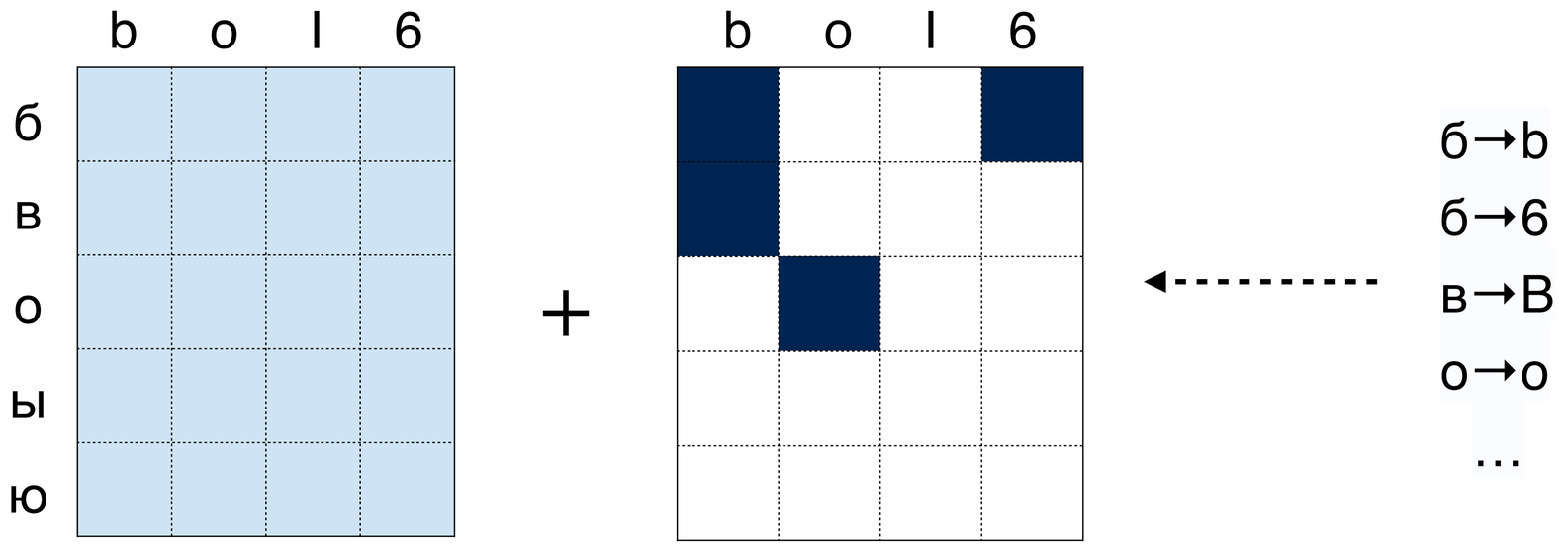
y	ɣ	ʏ	γ	у	Ү	ყ
0079 LATIN SMALL LETTER Y	0263 LATIN SMALL LETTER GAMMA	028F LATIN LETTER SMALL CAPITAL Y	03B3 GREEK SMALL LETTER GAMMA	0443 CYRILLIC SMALL LETTER U	04AF CYRILLIC SMALL LETTER STRAIGHT U	10E7 GEORGIAN LETTER QAR
p	ρ	ϱ	⍴	⍴	⍴	𝗽
0070 LATIN SMALL LETTER P	03C1 GREEK SMALL LETTER RHO	03F1 GREEK RHO SYMBOL	0440 CYRILLIC SMALL LETTER ER	2374 APL FUNCTIONAL SYMBOL RHO	2CA3 COPTIC SMALL LETTER RO	1D429 MATHEMATICAL BOLD SMALL P

https://util.unicode.org/UnicodeJsps/confusables.jsp

Informative priors

- Use mappings of similar characters as priors on emission parameters

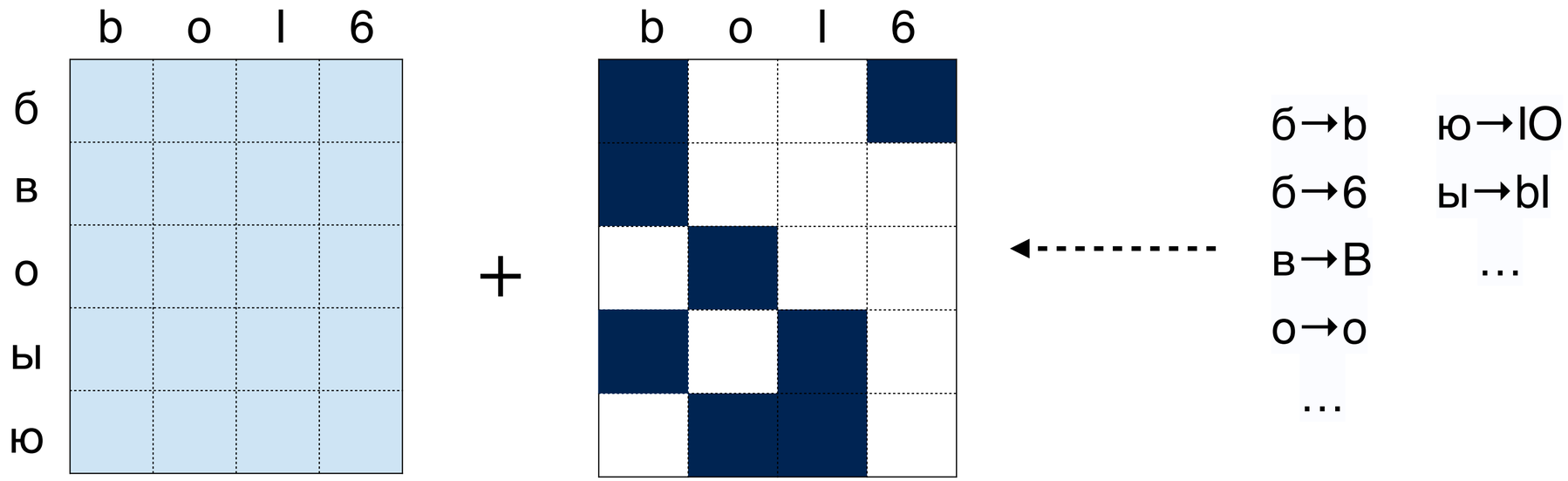
$$c_r | c_n \sim \mathrm{Mult}(\theta_{c_n}) \qquad \theta \sim \mathrm{Dir}(\alpha)$$



Informative priors

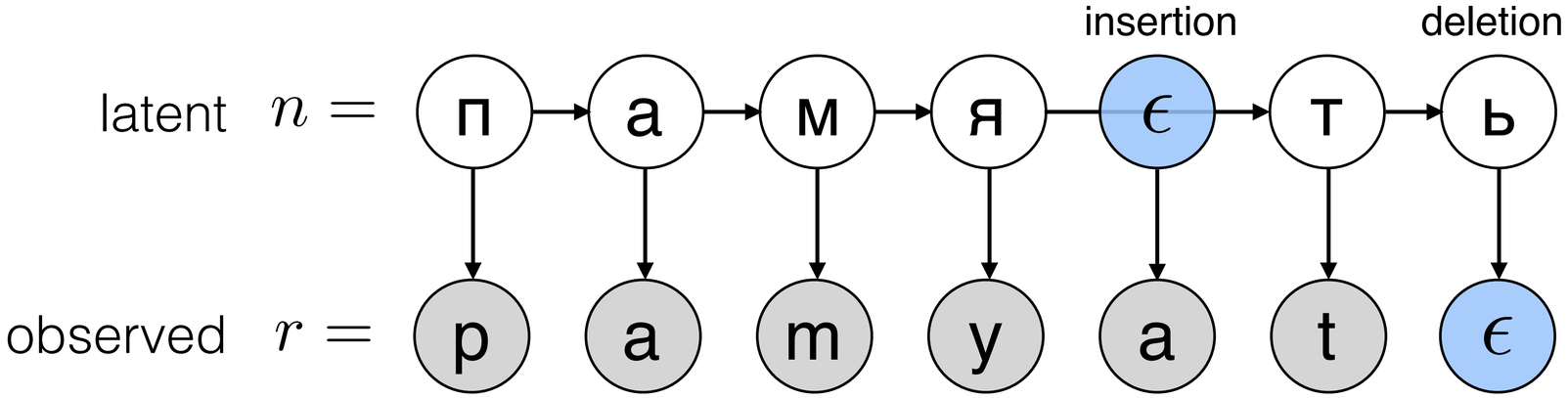
- Use mappings of similar characters as priors on emission parameters

$$c_r | c_n \sim \text{Mult}(\theta_{c_n}) \qquad \theta \sim \text{Dir}(\alpha)$$



Noisy-channel model

- Representing latent alignments via insertions and deletions
  - Natural match for null alignments, but will be used for character n-grams too!



$$p(r) = \sum_n p(n;\gamma) \cdot p(r|n;\theta) \cdot p_{\text{prior}}(\theta;\alpha)$$

transition probabilities

emission probabilities

prior on parameters

Proposed WFST

- Transition WFSA T
  - 6-gram LM built with OpenGrm (Roark et al., 2012)

Proposed WFST

- Transition WFSA T
  - 6-gram LM built with OpenGrm (Roark et al., 2012)
- Emission WFST E
  - Needs to support insertions and deletions

Proposed WFST

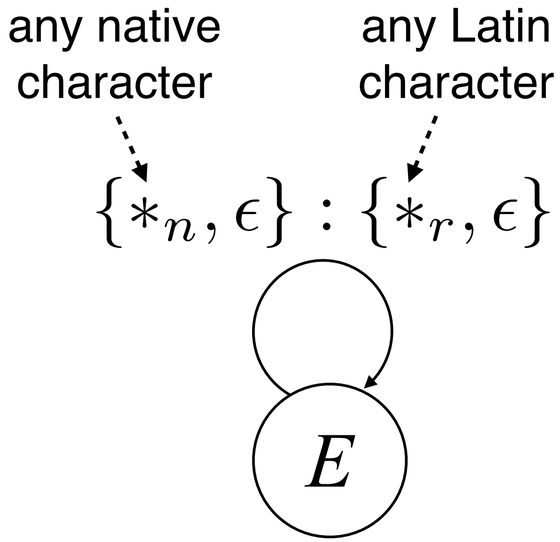
- Transition WFSA T
  - 6-gram LM built with OpenGrm (Roark et al., 2012)
- Emission WFST E
  - Needs to support insertions and deletions
- Input/output acceptors A

Proposed WFST

- Transition WFSA T
  - 6-gram LM built with OpenGrm (Roark et al., 2012)
- Emission WFST E
  - Needs to support insertions and deletions
- Input/output acceptors A
- Training and inference via finite-state methods in $T \circ E \circ A(r)$
  - OpenFst (Allauzen et al., 2007)

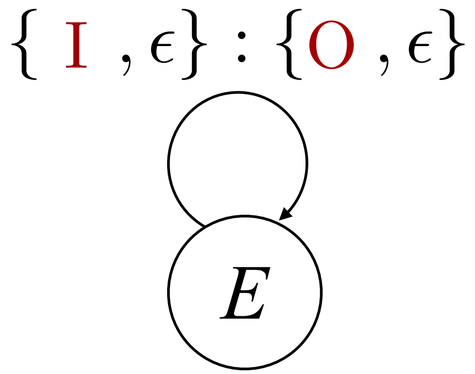
Emission model

- Needs to support substitutions, insertions and deletions



Emission model

- Needs to support substitutions, insertions and deletions



$$E \circ A(\text{`O'})$$

$$\text{I} : \text{O}$$

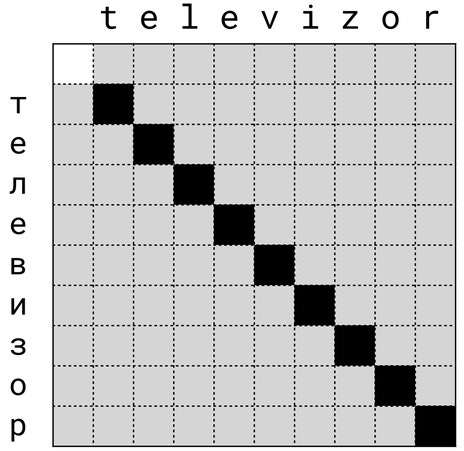
$$\epsilon : \text{O}$$

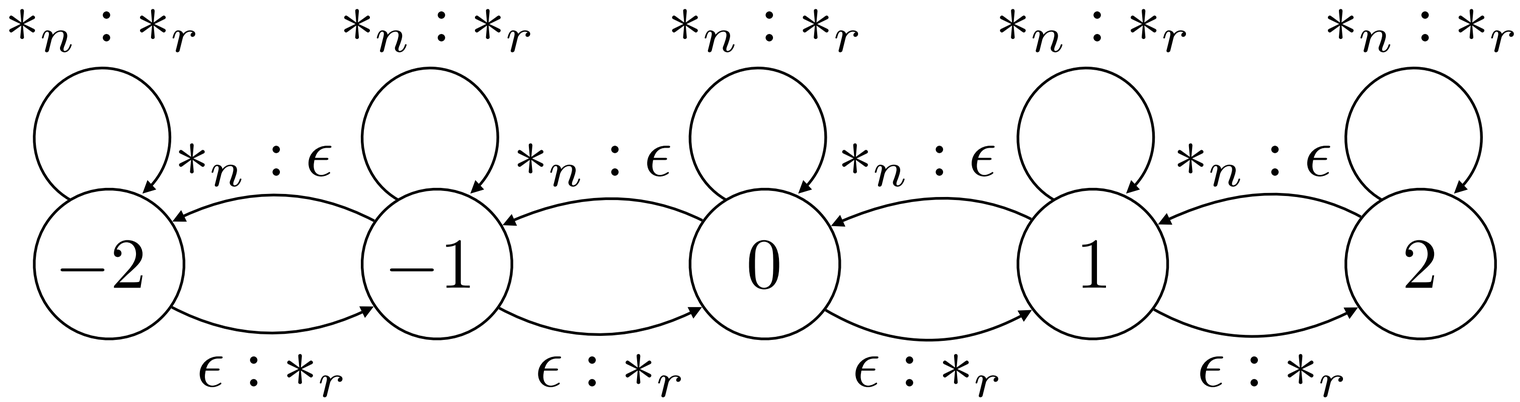
$$\text{I} : \epsilon \quad \epsilon : \text{O}$$

$$\text{I} : \epsilon \quad \text{I} : \epsilon \quad \ldots \quad \epsilon : \text{O}$$

Emission model

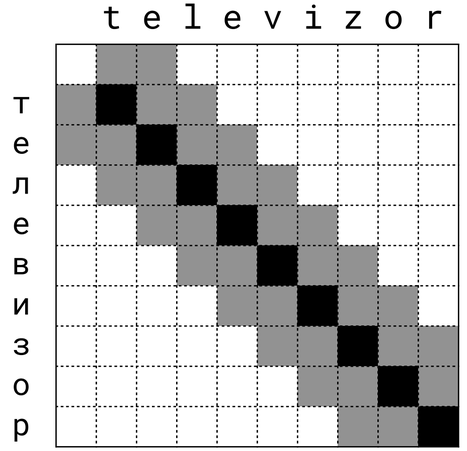
- Needs to support substitutions, insertions and deletions
- Fixed limit on delay: |# of insertions - # of deletions|

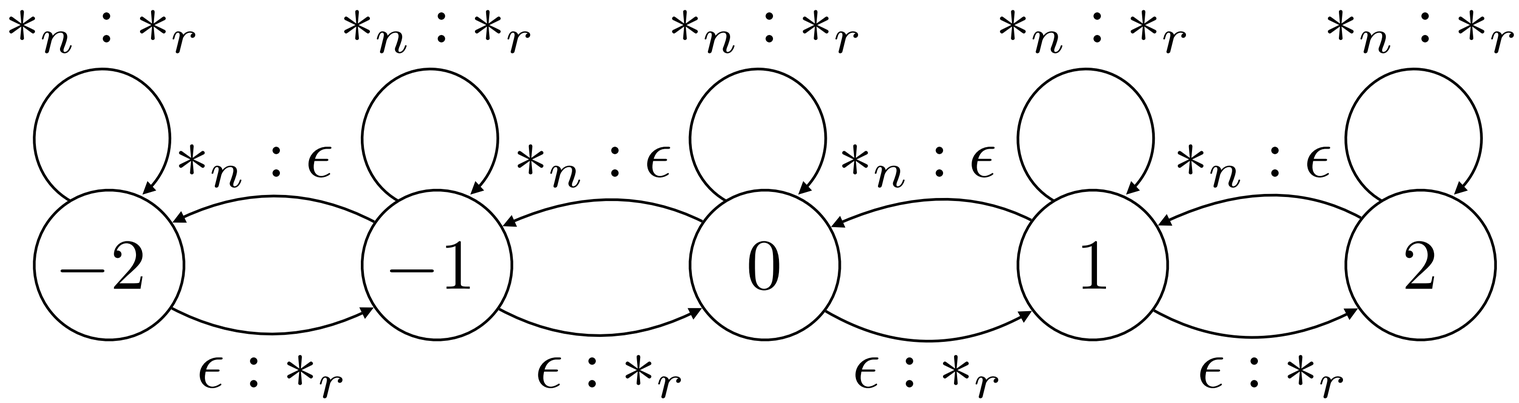




Emission model

- Needs to support substitutions, insertions and deletions
- Fixed limit on delay: |# of insertions - # of deletions|

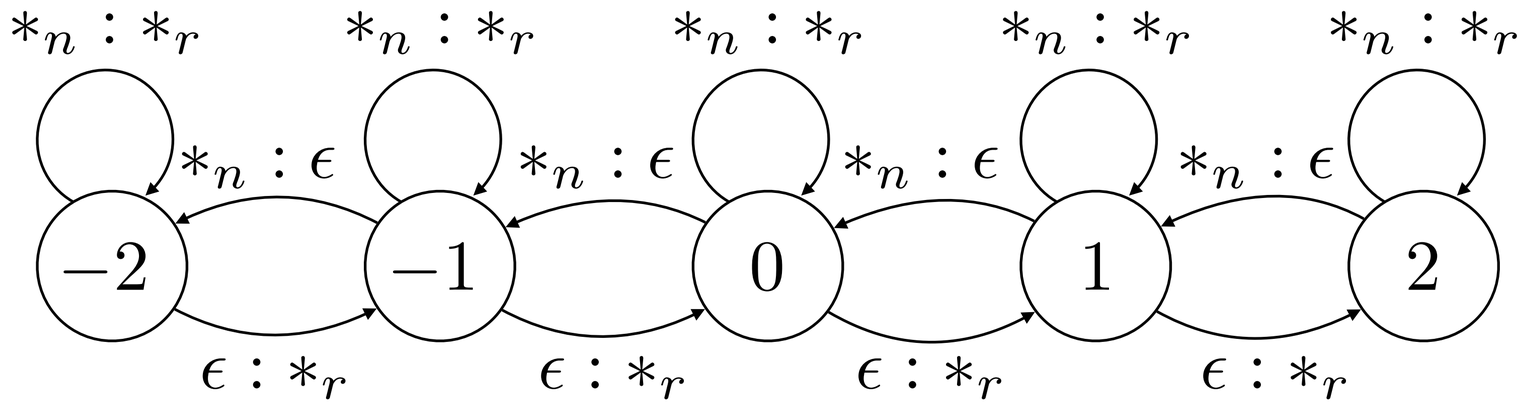




Emission model

- Needs to support substitutions, insertions and deletions
- Fixed limit on delay: |# of insertions - # of deletions|

$E \circ A(\text{‘O’})$

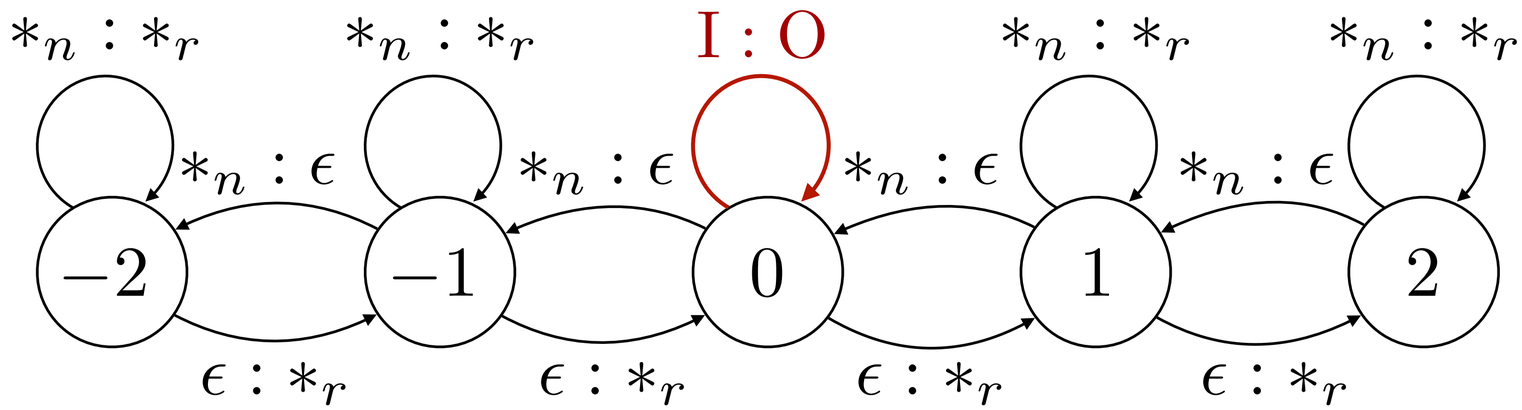


Emission model

- Needs to support substitutions, insertions and deletions
- Fixed limit on delay: |# of insertions - # of deletions|

$E \circ A(\text{‘O’})$

$\mathrm{I} : \mathrm{O}$

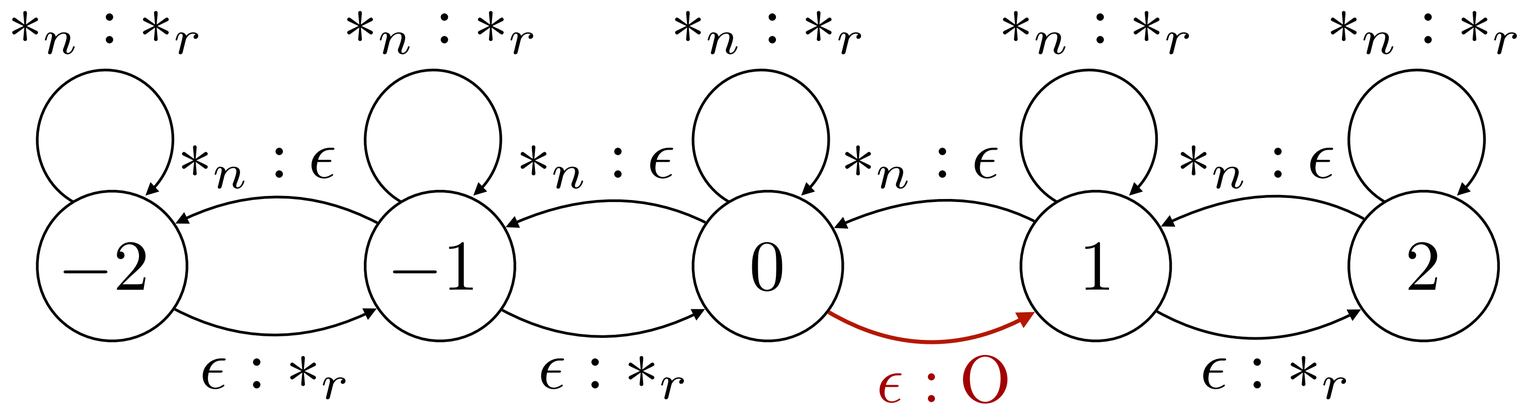


Emission model

- Needs to support substitutions, insertions and deletions
- Fixed limit on delay: |# of insertions - # of deletions|

$E \circ A(\text{‘O’})$

$\epsilon : \text{O}$

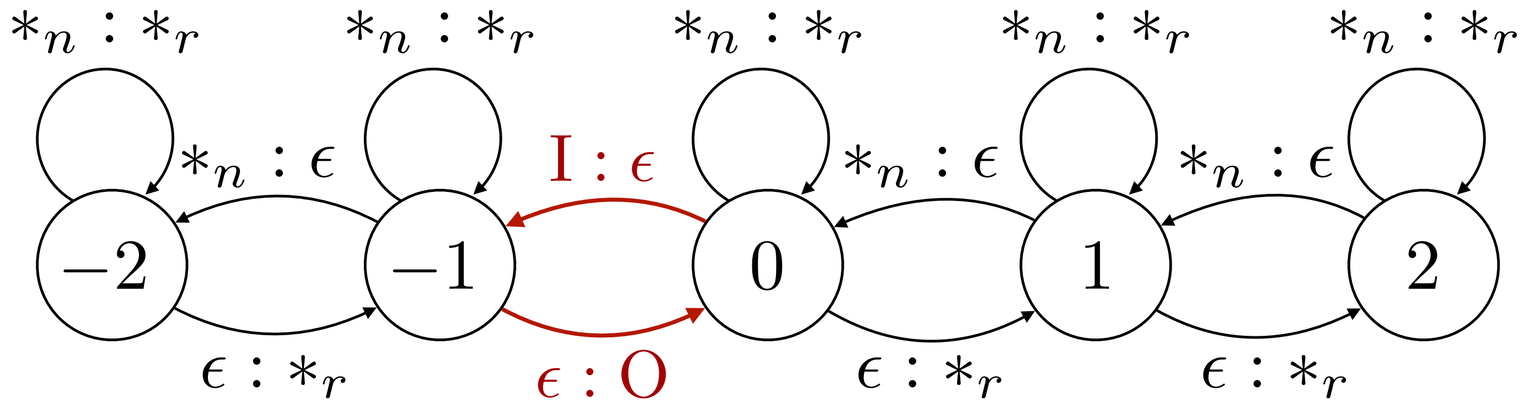


Emission model

- Needs to support substitutions, insertions and deletions
- Fixed limit on delay: |# of insertions - # of deletions|

$E \circ A(\text{‘O’})$

$\mathrm{I} : \epsilon \quad \epsilon : \mathrm{O}$

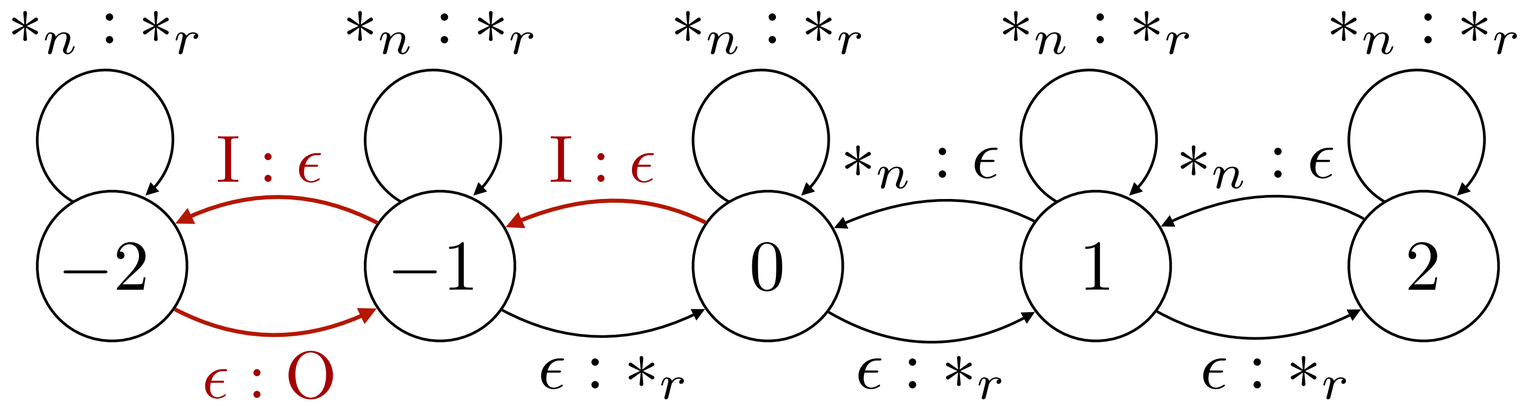


Emission model

- Needs to support substitutions, insertions and deletions
- Fixed limit on delay: |# of insertions - # of deletions|

$$E \circ A(\text{‘O’})$$

$$\text{I} : \epsilon \quad \text{I} : \epsilon \quad \ldots \quad \epsilon : \text{O}$$



Training and inference

- Training with EM algorithm
  - E-step: shortest distance in expectation semiring (Eisner, 2002)
  - M-step: parameter reestimation
- Many tricks to speed up training!
  - Stepwise batched EM (Liang and Klein, 2009)
  - Curriculum learning: shortest sequences first
  - Increasing LM order as training progresses
  - Pruning emission arcs during training
- Inference: shortest path in (max, +) semiring

Datasets

- Arabic: LDC BOLT dataset (Bies et al., 2014)
  - Arabizi SMS/chat dialogs, converted to CODA (Habash et al., 2012)

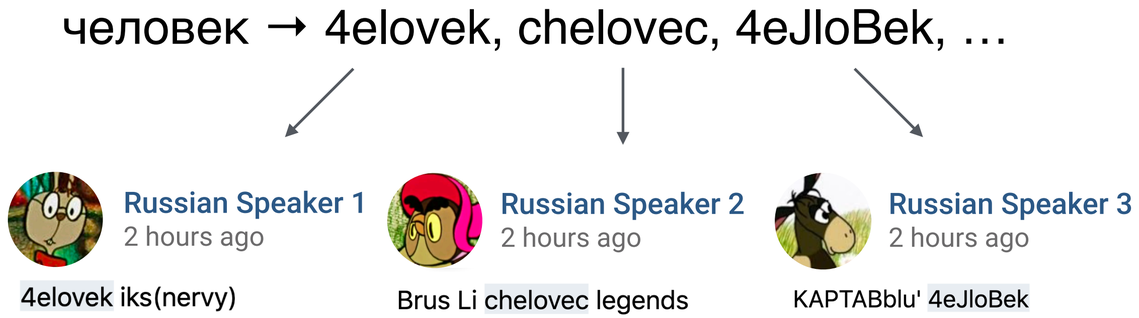
Saba7 el 5eir!
Ezayeeky?

- Kannada: Dakshina dataset (Roark et al., 2020)
  - Kannada Wikipedia, romanizations elicited from native speakers

- Russian:
  - Romanized: collected and partly annotated data from social media
  - Native: Taiga corpus (Shavrina & Shapovalova, 2017), comments in political forums

Russian data

- Romanizations of common words used as queries (Darwish, 2014)



- Manually removed sentences in other languages (e.g. Polish)
- Annotated validation and test with minor error correction

Source: proishodit s prirodoy 4to to very very bad

Filtered: proishodit s prirodoy 4to to <…>

Target: происходит с природой что-то <…>

Defining vocabulary

- Characters ~ Unicode codepoints?
  - Diacritics treated as separate characters
  - Non-printing symbols (e.g. ZWNJ) treated as separate characters

ಕ = ka

ಕ + ು = ku

ಕ + ್ = k

ರಾಜ್‌ಕುಮಾರ್

rāj|kumār

ರ ಾ ಜ ್ ‌ ಕ ು ಮ ಾ ರ ್

Ra Aa Ja ∅ ‌ Ka U Ma Aa Ra ∅

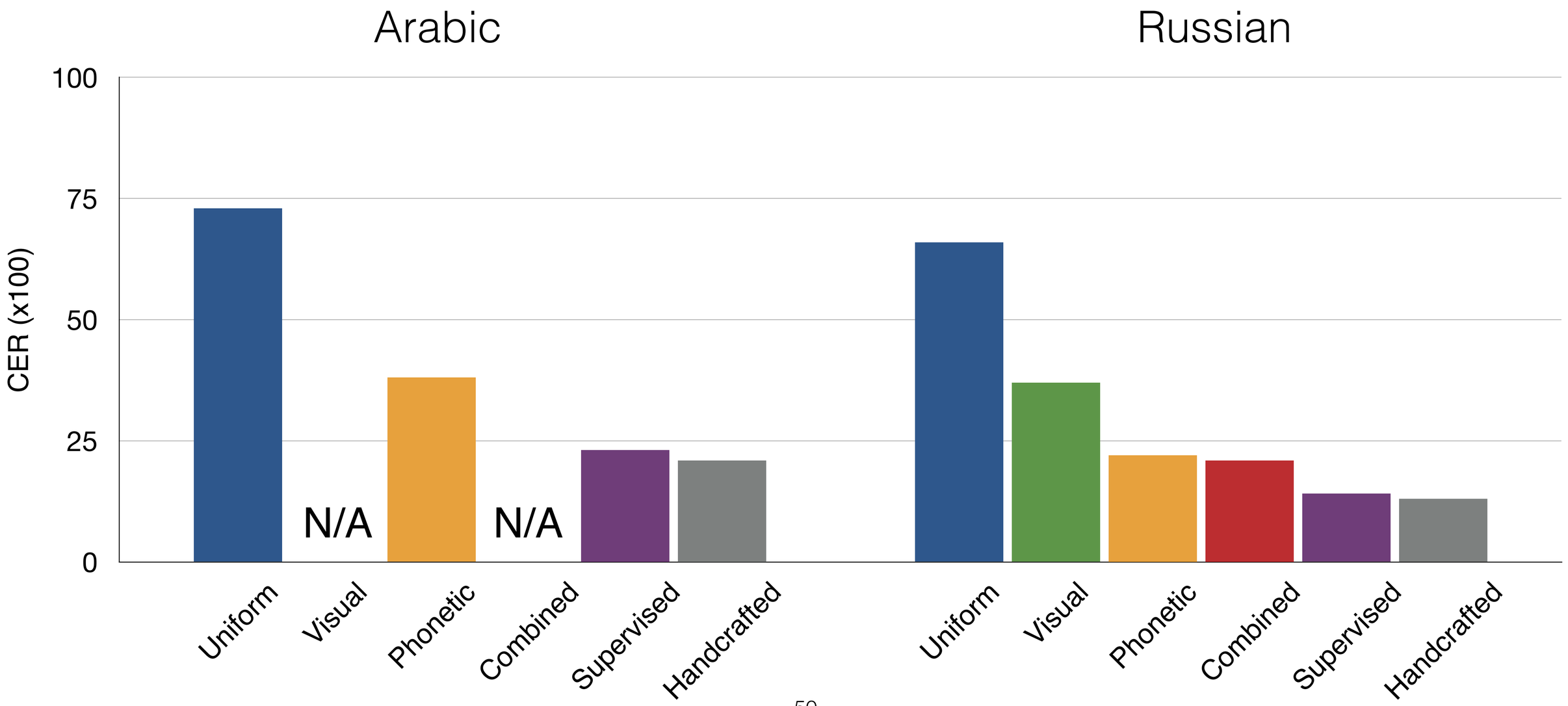
Defining vocabulary

- Characters ~ Unicode codepoints?
  - Diacritics treated as separate characters
  - Non-printing symbols (e.g. ZWNJ) treated as separate characters
  - Combined characters may have multiple Unicode representations

For	Use	Do Not Use
ಊ	0C8A	<0C89, 0CBE>
ೌ	0C94	<0C92, 0CCC>
ೠ	0CE0	<0C8B, 0CBE>

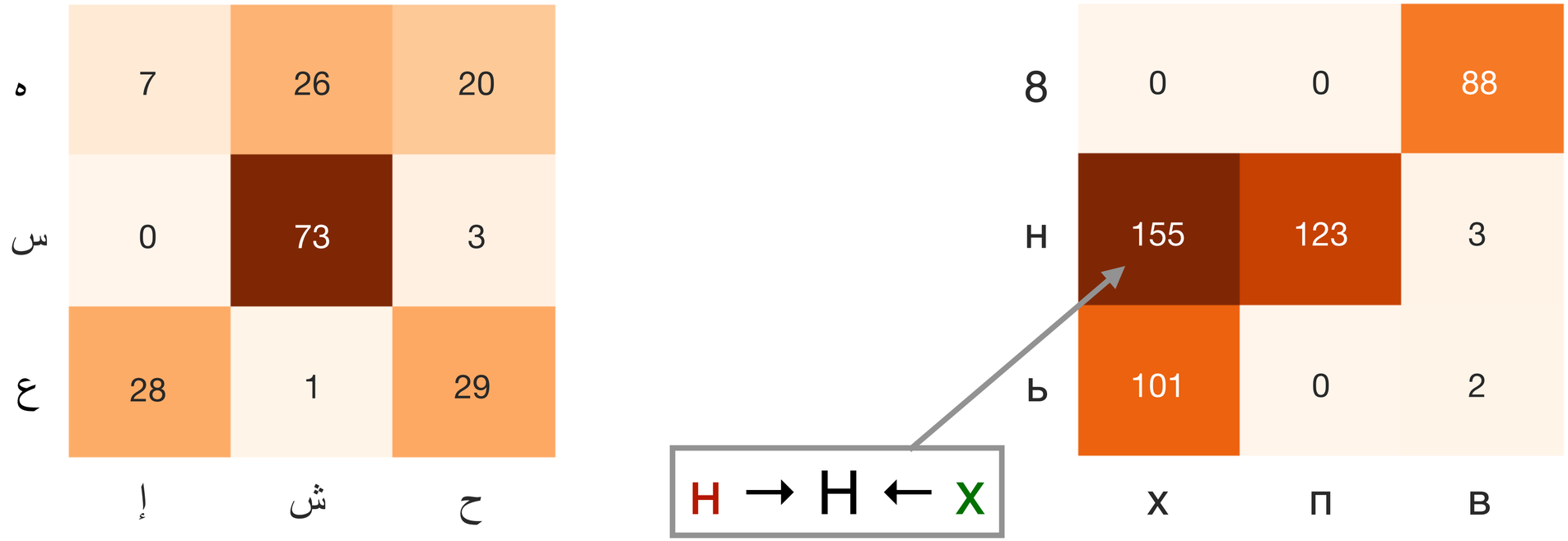
 https://www.unicode.org/versions/Unicode12.0.0/UnicodeStandard-12.0.pdf

WFST results



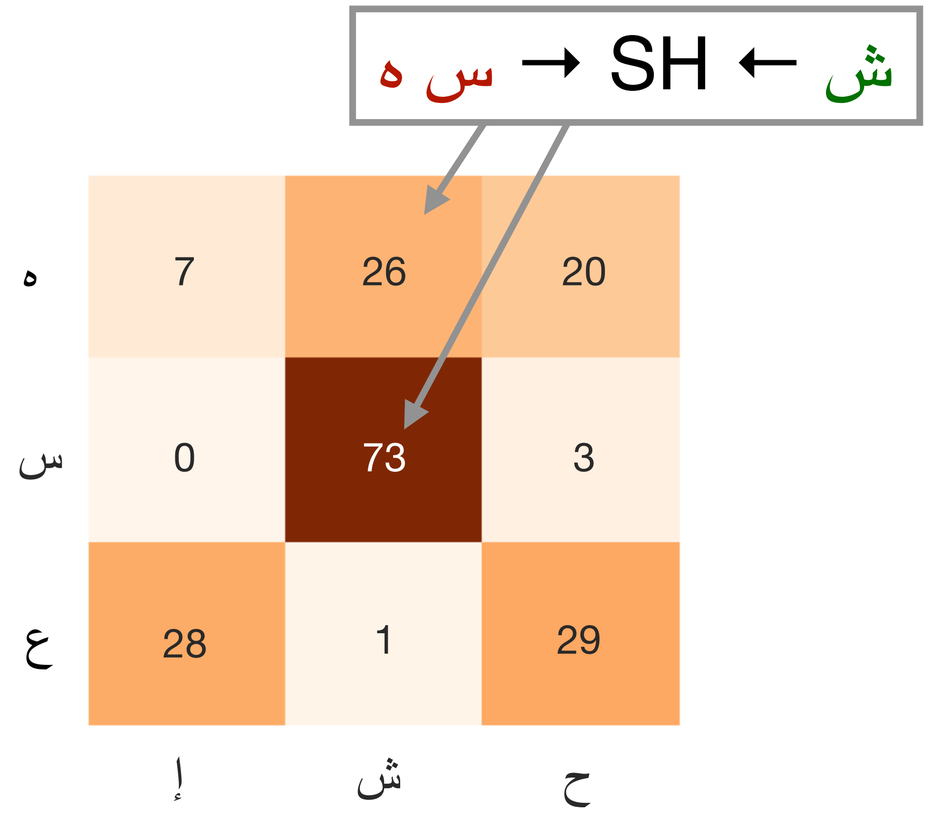
WFST error analysis

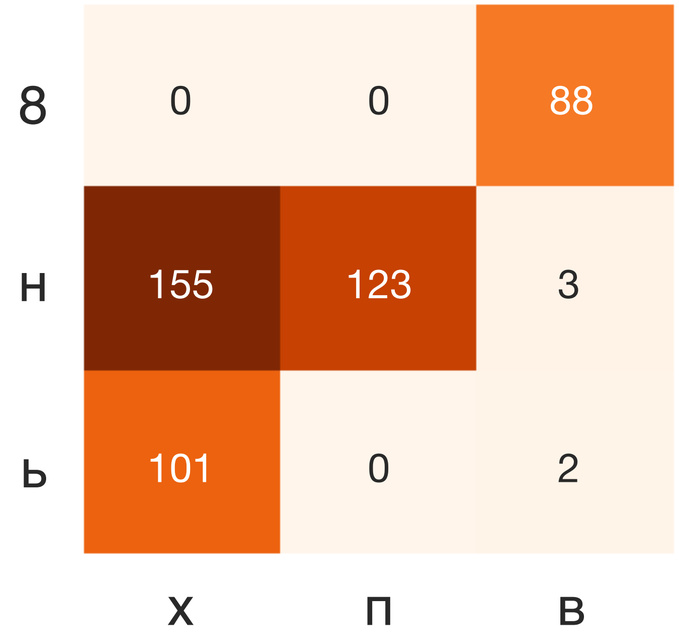
- Incorrect choice of plausible de-romanization (e.g. visual instead of phonetic)



WFST error analysis

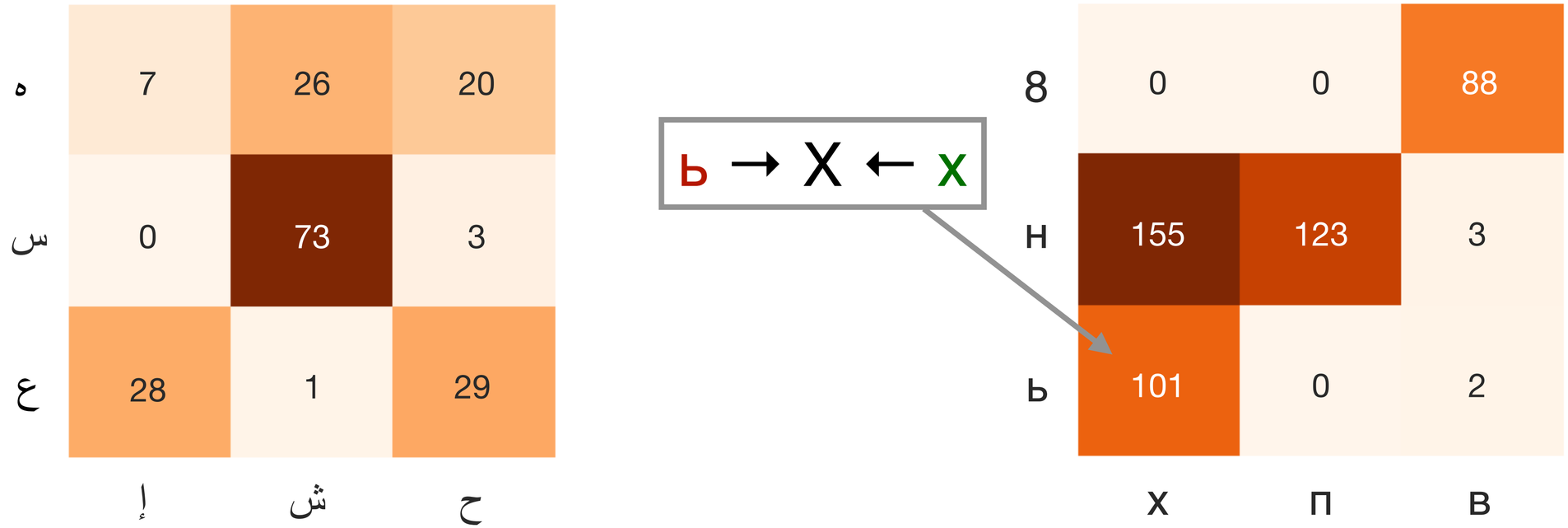
- Inability to handle digraphs like SH





WFST error analysis

- Distracted by spurious mappings in priors



Model classes

WFSTs are structured

- ✓ Easy to encode constraints
- ✓ Can learn from small data
- ✗ Slow exact maximization
- ✗ Weak n-gram language model

Seq2seqs are powerful

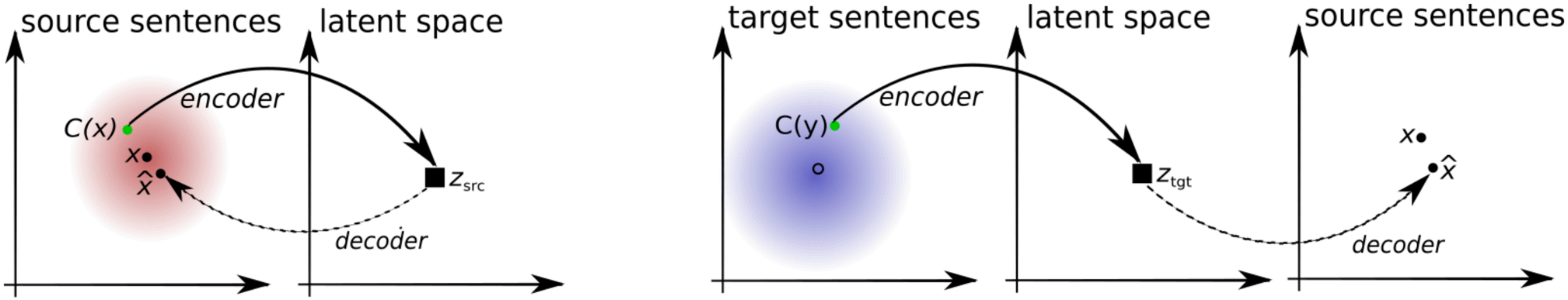
- ✓ Strong language model
- ✓ Faster batch processing
- ✗ Need large training data
- ✗ Hallucinations and search errors

In our case, both are trained unsupervised!

M Ryskina, E Hovy, T Berg-Kirkpatrick, MR Gormley. Comparative Error Analysis in Neural and Finite-state Models for Unsupervised Character-level Transduction. SIGMORPHON 2021.

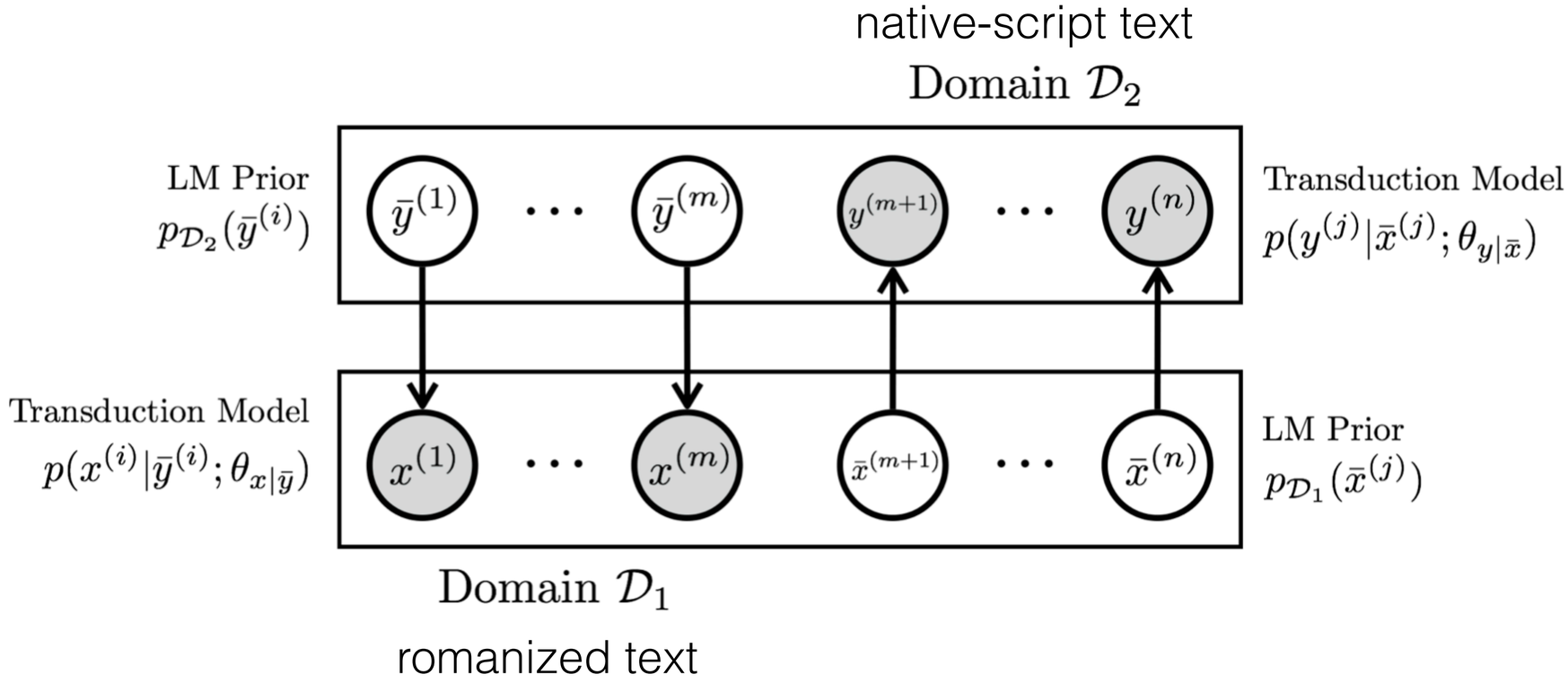
Unsupervised seq2seq

- Unsupervised neural machine translation (UNMT; Lample et al., 2018)
  - Auto-encoding: reconstructing a sentence from its noisy version
  - Back-translation: round trip through the latent space
  - Adversarial: discriminating between sentences in two domains



Unsupervised seq2seq

- Probabilistic formulation of UNMT: deep latent sequence model (He et al., 2020)

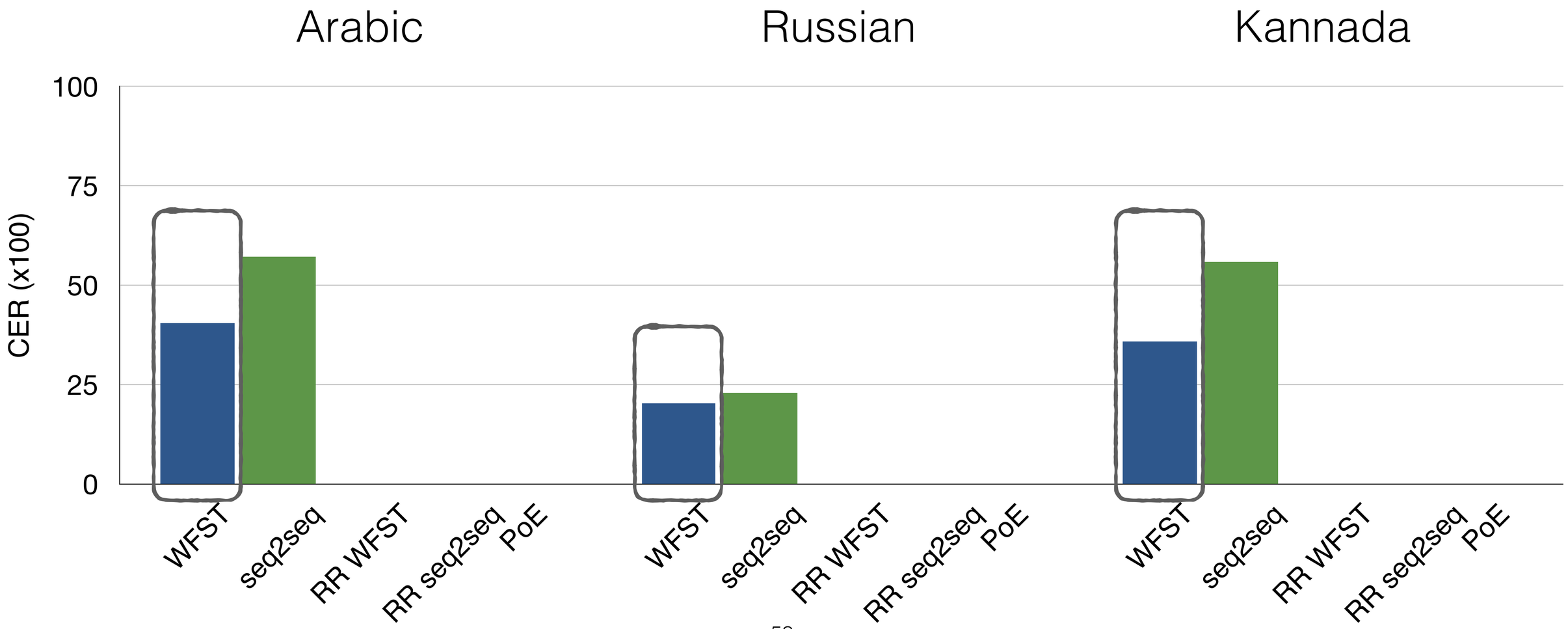


Model combinations

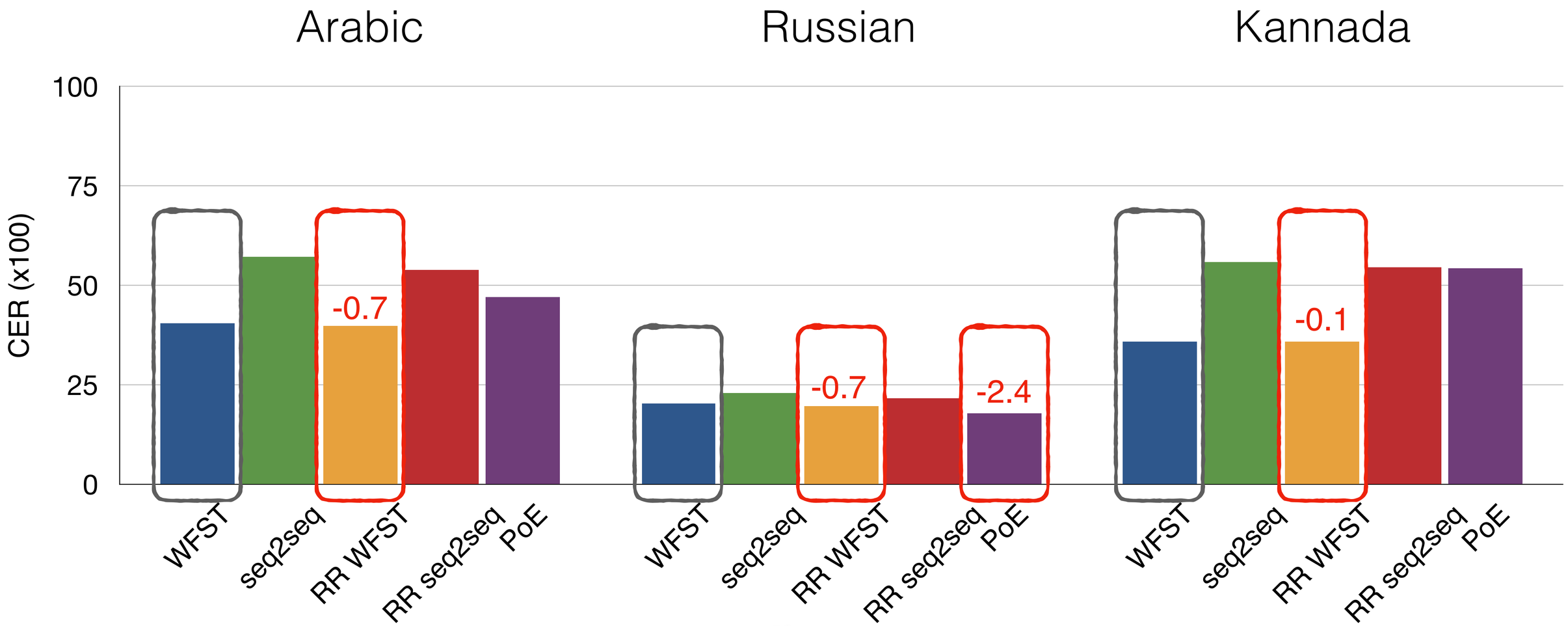
- Reranking
  - M1 generates top k candidate outputs
  - M2 selects the highest-scoring candidate
- Product of experts
  - Beam search on the WFST lattice
  - WFST arcs reweighted with Seq2seq softmax at the corresponding timstep
  - Deletions of input characters are not reweighted
  - Candidates are grouped by consumed input length
- We train the models separately and combine at test time

Results

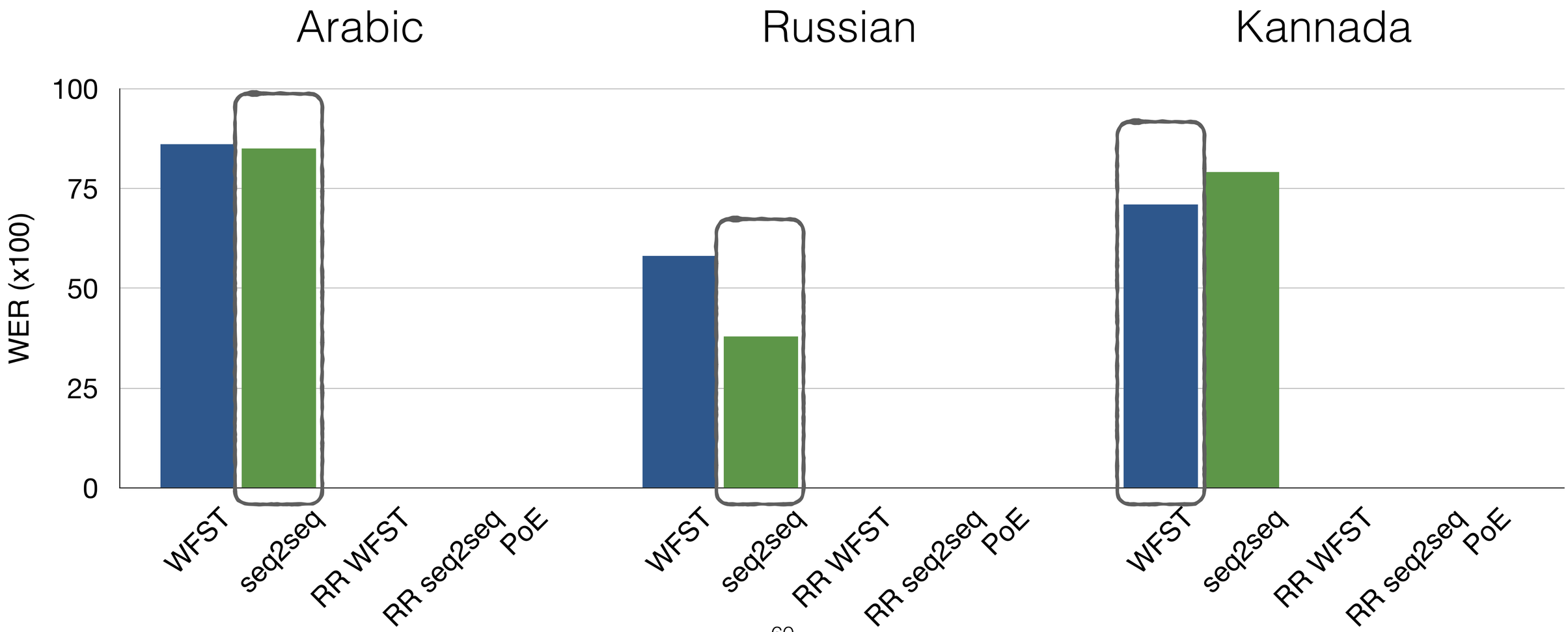
Base models are trained on different amounts of data!



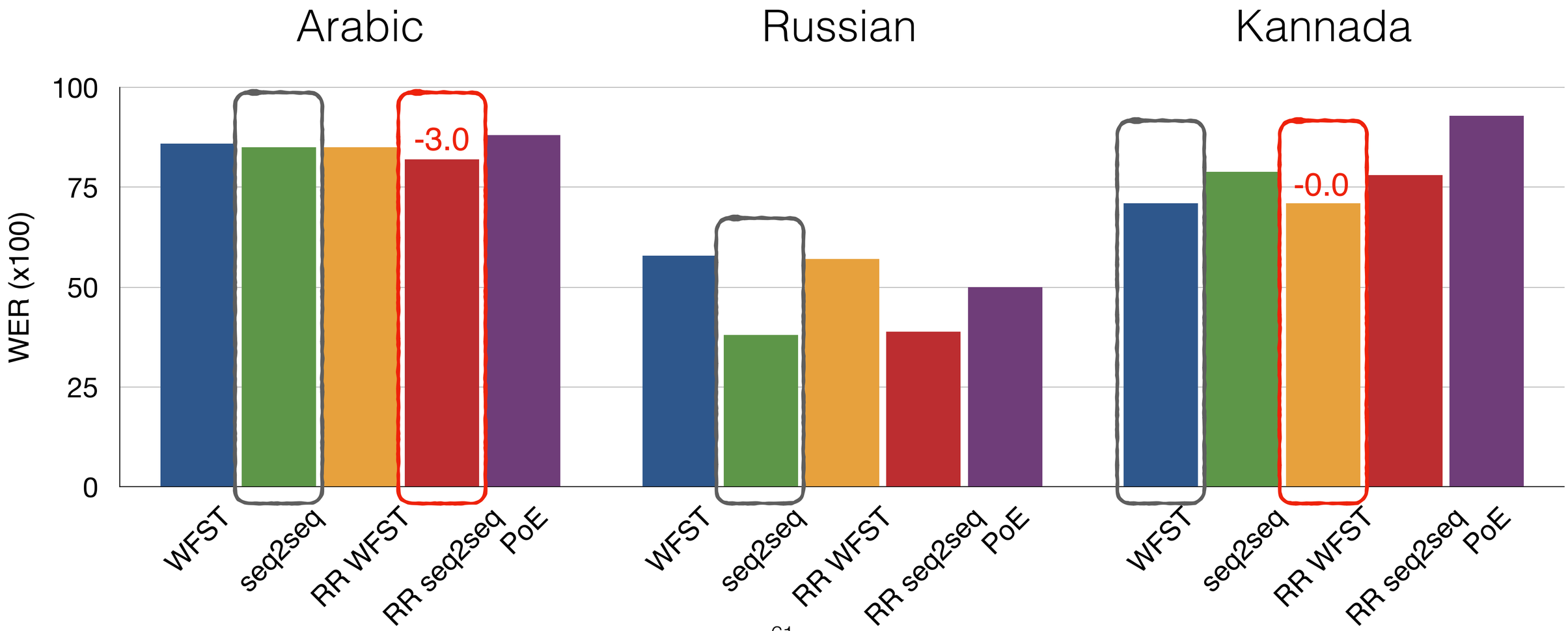
Results


Arabic
Russian
Kannada
CER (x100)
100
75
50
25
0
-0.7
-0.7
-2.4
-0.1
WFST
seq2seq
RR WFST
RR seq2seq
PoE
WFST
seq2seq
RR WFST
RR seq2seq
PoE
WFST
seq2seq
RR WFST
RR seq2seq
PoE

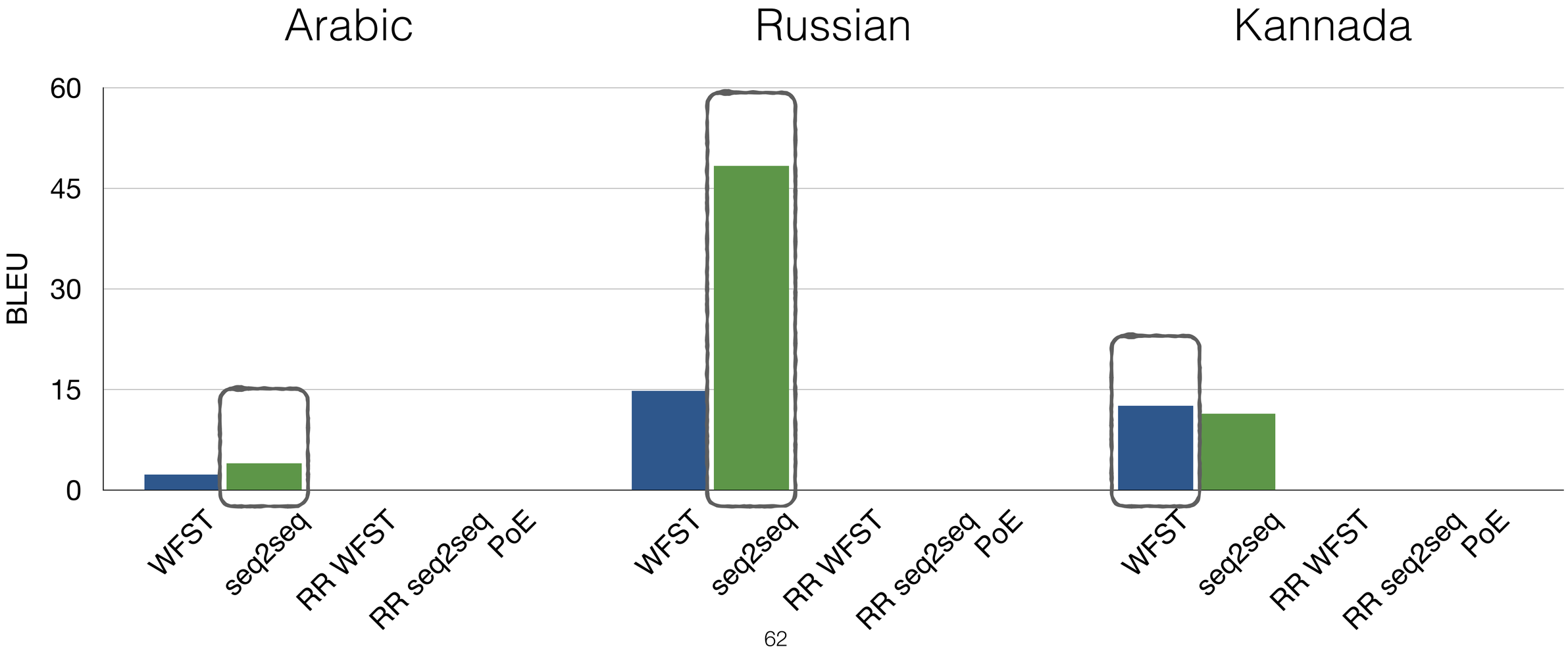
Results


Arabic
Russian
Kannada
WER (x100)
100
75
50
25
0
WFST
seq2seq
RR WFST
RR seq2seq
PoE
WFST
seq2seq
RR WFST
RR seq2seq
PoE
WFST
seq2seq
RR WFST
RR seq2seq
PoE

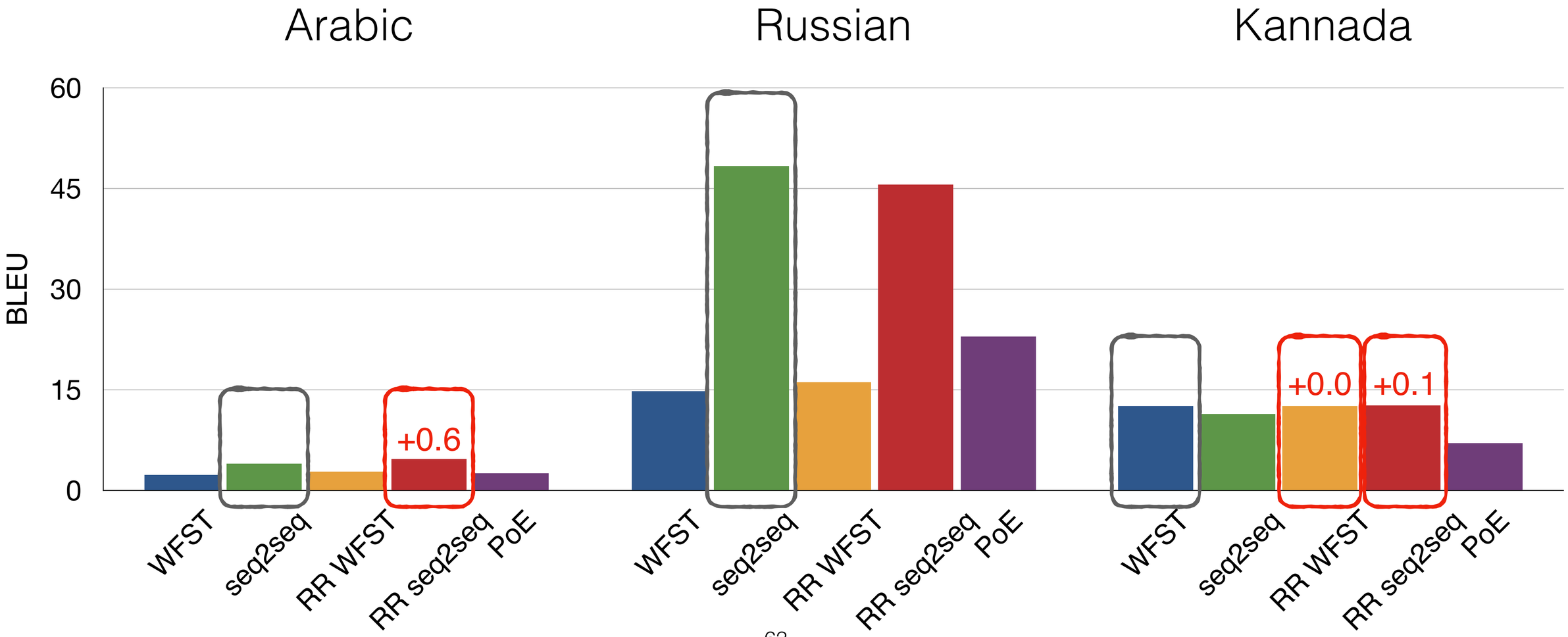
Results


Arabic
Russian
Kannada
WER (x100)
100
75
50
25
0
-3.0
-0.0
WFST
seq2seq
RR WFST
RR seq2seq
PoE

Results


Arabic
Russian
Kannada
BLEU
60
45
30
15
0
WFST
seq2seq
RR WFST
RR seq2seq
PoE
WFST
seq2seq
RR WFST
RR seq2seq
PoE
WFST
seq2seq
RR WFST
RR seq2seq
PoE

Results


Arabic
Russian
Kannada
BLEU
60
45
30
15
0
+0.6
+0.0
+0.1
WFST
seq2seq
RR WFST
RR seq2seq
PoE
WFST
seq2seq
RR WFST
RR seq2seq
PoE
WFST
seq2seq
RR WFST
RR seq2seq
PoE

Error analysis

Input	kongress ne odobril biudjet dlya osuchestvleniye "bor'bi s kommunizmom" v yuzhniy amerike.	
Ground truth	конгресс не одобрил бюджет для осуществления "борьбы с коммунизмом" в южной америке.	kongress ne odobril bjudžet dlja osuščestvlenija "bor'by s kommunizmom" v južnoj amerike.
WFST	конгресс не одобрил виудет дла осусчествлениые "бор#би с коммунизмом" в уузнани америке.	kongress ne odobril viudet dla osusčestvleniye "bor#bi s kommunizmom" v uuznani amerike.
Reranked WFST	конгресс не одобрил видет дела осусчествлениые "бор#би с коммунизмом" в уузнани америке.	kongress ne odobril videt dela osusčestvleniye "bor#bi s kommunizmom" v uuznani amerike.
Seq2Seq	конгресс не одобрил бы удивительно с коммунизмом" в южный америке.	kongress ne odobril by udivitel'no s kommunizmom" v južnyj amerike.
Reranked Seq2Seq	конгресс не одобрил бюджет для осуществление "борьбы с коммунизмом" в южный америке.	kongress ne odobril bjudžet dlja osuščestvlenie "bor'by s kommunizmom" v južnyj amerike.
Product of experts	конгресс не одобрил бидет для а осуществлениые "борьбы с коммунизмом" в уузник амери	kongress ne odobril bidet dlja a osuščestvleniye "bor'by s kommunizmom" v uuznik ameri

#=UNK

Error analysis

Input	kongress ne odobril biudjet dlya osuchestvleniye "bor'bi s kommunizmom" v yuzhniy amerike.	
Ground truth	конгресс не одобрил бюджет для осуществления "борьбы с коммунизмом" в южной америке.	kongress ne odobril bjudžet dlja osuščestvlenija "bor'by s kommunizmom" v južnoj amerike.
WFST	конгресс не одобрил виудет дла осусчествлениые "бор#би с коммунизмом" в уузнани америке.	kongress ne odobril viudet dla osusčestvleniye "bor#bi s kommunizmom" v uuznani amerike.
Reranked WFST	конгресс не одобрил видет дела осусчествлениые "бор#би с коммунизмом" в уузнани америке.	kongress ne odobril videt dela osusčestvleniye "bor#bi s kommunizmom" v uuznani amerike.
Seq2Seq	конгресс не одобрил бы удивительно с коммунизмом" в южный америке.	kongress ne odobril by udivitel'no s kommunizmom" v južnyj amerike.
Reranked Seq2Seq	конгресс не одобрил бюджет для осуществление "борьбы с коммунизмом" в южный америке.	kongress ne odobril bjudžet dlja osuščestvlenie "bor'by s kommunizmom" v južnyj amerike.
Product of experts	конгресс не одобрил бидет для а осуществлениые "борьбы с коммунизмом" в уузник амери	kongress ne odobril bidet dlja a osuščestvleniye "bor'by s kommunizmom" v uuznik ameri

Hallucination

Incorrect but faithful

High-level takeaways

- Model combinations still suffer from search issues

Source: eto uzhe (strashno skazat') stariy rolik.

Target: это уже (страшно сказать) старый ролик

Gloss: ‘By now this is (I’m almost afraid to say it) an old video’

Final beam hypotheses and reranker scores:

456.7, единая россия уже #страшно сказать) старый
502.0, единоросы уже #страшно сказать) старый рол
482.0, единороссы уже #страшно сказать) старый ро
456.8, единую россию уже #страшно сказать) старый
449.8, единой россии уже #страшно сказать) старый

High-level takeaways

- Model combinations still suffer from search issues

Source: eto uzhe (strashno skazat') stariy rolik.

Target: это уже (страшно сказать) старый ролик

Gloss: ‘This’ w this is (I’m almost afraid to say it) an old video’

Final beam hypotheses and reranker scores:

456.7, единая россия уже #страшно сказать) старый
502.0, единоросы уже #страшно сказать) старый рол
482.0, единороссы уже #страшно сказать) старый ро
456.8, единую россию уже #страшно сказать) старый
449.8, единой россии уже #страшно сказать) старый

‘United Russia’

High-level takeaways

- Model combinations still suffer from search issues

Source: eto uzhe (strashno skazat') stariy rolik.

Target: это уже (страшно сказать) старый ролик

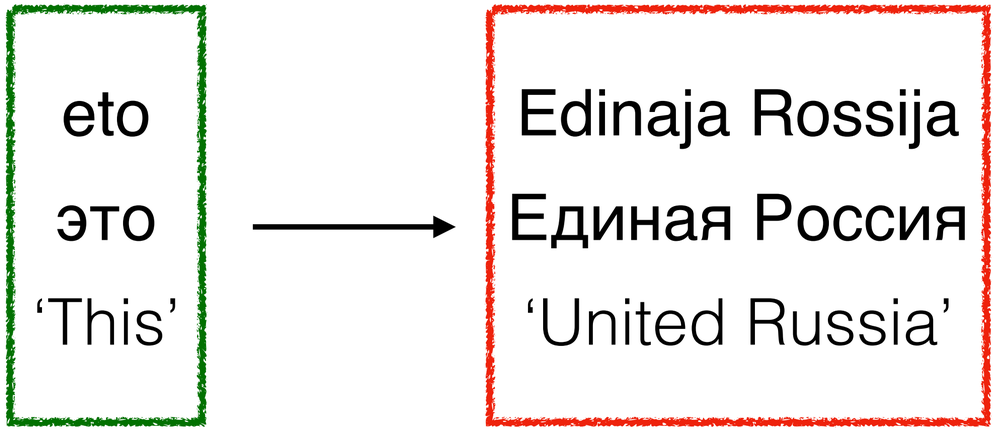
Gloss: ‘By now this is (I’m almost afraid to say it) an old video’

Final beam hypotheses and reranker scores:

456.7, единая россия уже #страшно сказать) старый
502.0, единоросы уже #страшно сказать) старый рол
482.0, единороссы уже #страшно сказать) старый ро
456.8, единую россию уже #страшно сказать) старый
449.8, единой россии уже #страшно сказать) старый

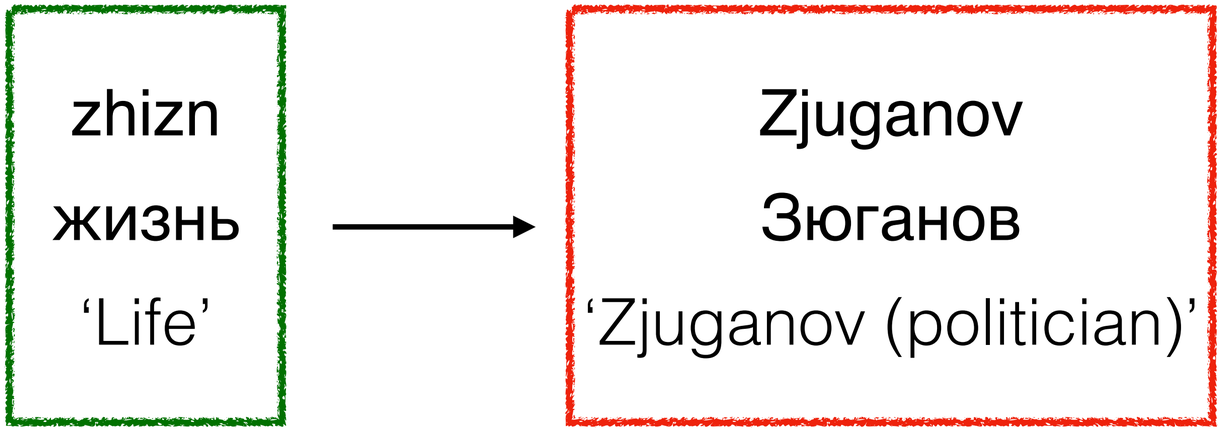
High-level takeaways

- Seq2seq is more sensitive to distributional shifts
  - Remember that our Cyrillic data comes from political discussion groups



High-level takeaways

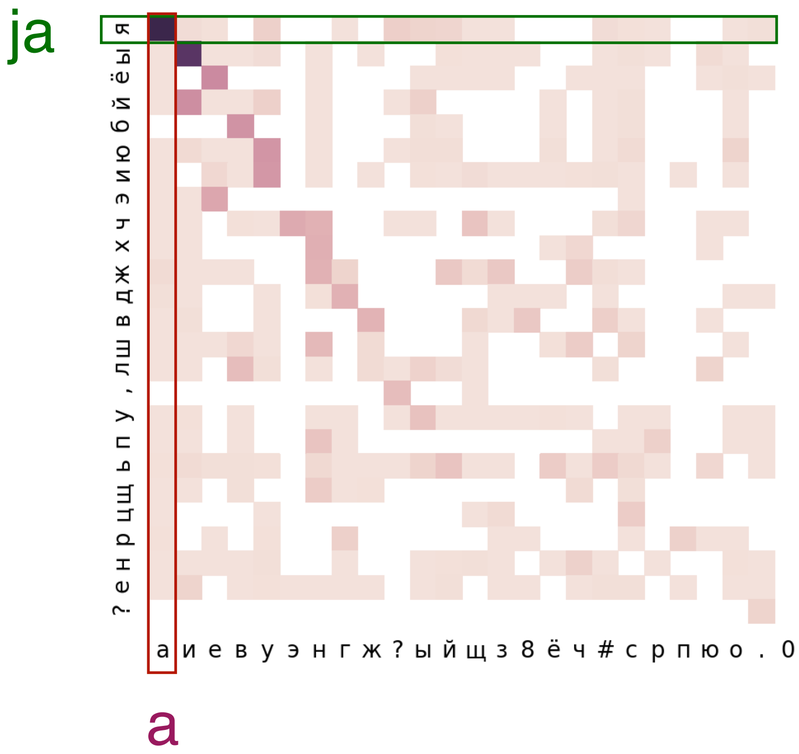
- Seq2seq is more sensitive to distributional shifts
  - Remember that our Cyrillic data comes from political discussion groups
  - 25% of most frequent substitions under the seq2seq are caused by domain mismatch, compared to 3% for WFST!



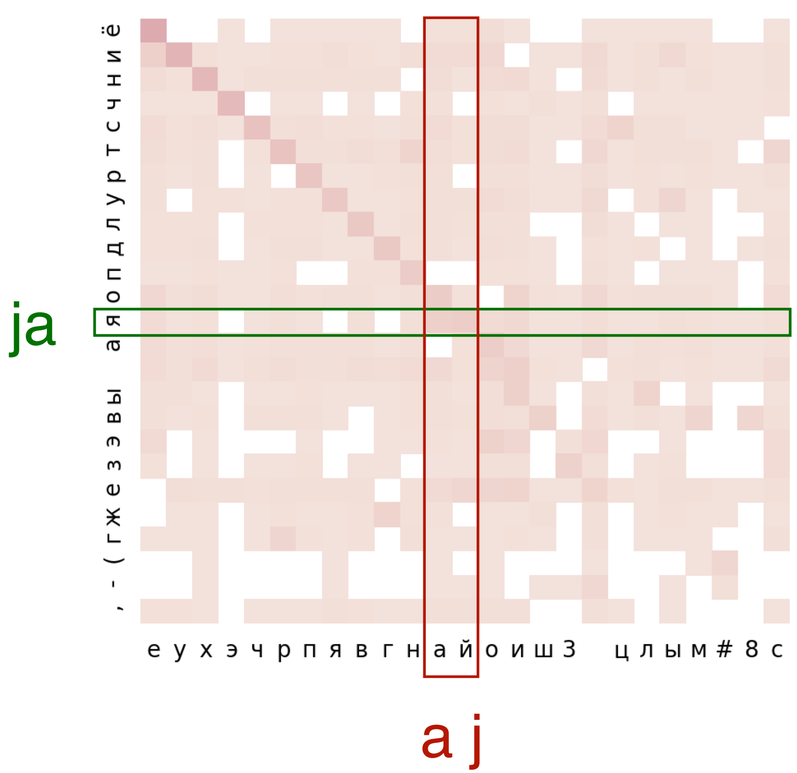
High-level takeaways

- WFST makes more repetitive errors
  - Suggests that WFST outputs might be easier to correct with rule-based postprocessing

WFST



seq2seq



Future work

Combining unsupervised finite-state and neural models

- Our decoding-time combinations didn't help much, but I still believe in combining the two model classes!
- It could be joint training…
- …Or holistic structural combinations…
  - Encoding WFST-like constraints into the attention mechanism (Aharoni & Goldberg, 2017; Makarov et al., 2017; Wu et al., 2018; Wu & Cotterell, 2019)
  - Neural reweighting of WFSTs (Rastogi et al., 2016; Lin et al., 2019)
- …Or 'softer' biasing of one model towards another model's behavior

Future work

Combining unsupervised finite-state and neural models

- Can enough ‘power’ replace ‘structure’?
  - Transformer can learn character-level transduction without structural constraints (Wu et al., 2021)
  - But less likely to be sufficient in unsupervised or low-data settings!

Future work

Making sense of user preferences

- Users tend to have consistent substitution preferences
- These preferences can be correlated with the author's background
  - Where does the author live?

English: B→V

German: B→W

Future work

Making sense of user preferences

- Users tend to have consistent substitution preferences
- These preferences can be correlated with the author's background
  - Where does the author live?
  - What are other dominant languages in that area?

Former British colonies (English)	Former French colonies (French)
ش →sh	ش →ch

Future work

Making sense of user preferences

- Users tend to have consistent substitution preferences
- These preferences can be correlated with the author's background
  - Where does the author live?
  - What are other dominant languages in that area?
  - How old might the author be?

Older standard (~German)	Newer standard (~English)
ц→c, ю→ju	**ц→ts, ю→yu**

Future work

Making sense of user preferences

- Users tend to have consistent substitution preferences
- These preferences can be correlated with the author's background
  - Where does the author live?
  - What are other dominant languages in that area?
  - How old might the author be?
- Could be used to study creative language variation and change
- But also might be a privacy risk!

Final thoughts

- Typology is important!
  - One needs to carefully account for differences between languages
  - These differences extend to seemingly ‘technical’ side as well
- Rules can be better than data
  - Handcrafted converters, even simple ones, outperformed the supervised WFST
- Normalization of creative phenomena is lossy
  - Ideally, annotation should be performed with feedback from the author
  - Developing non-destructive normalization methods to preserve social meaning

References

R. Aharoni, Y. Goldberg. Morphological inflection generation with hard monotonic attention. ACL 2017.

C. Allauzen, M. Riley, J. Schalkwyk, W. Skut, M. Mohri. OpenFst: A general and efficient weighted finite-state transducer library. CIAA 2007.

A. Bies, Z. Song, M. Maamouri, S. Grimes, H. Lee, J. Wright, S. Strassel, N. Habash, R. Eskander, O. Rambow. Transliteration of Arabizi into Arabic orthography: Developing a parallel annotated Arabizi-Arabic script SMS/chat corpus. ANLP 2014.

T. Baldwin, M. C. de Marneffe, B. Han, Y.-B. Kim, A. Ritter, W. Xu. Shared tasks of the 2015 Workshop on Noisy User-generated Text: Twitter lexical normalization and named entity recognition. W-NUT 2015.

E. M. Bender. Linguistically naïve != Language independent: Why NLP needs linguistic typology. ILCL 2009

A. Chalamandaris, A. Protopapas, P. Tsiakoulis, S. Raptis. All Greek to me! An automatic Greeklish to Greek transliteration system. LREC 2006

K. Darwish. Arabizi detection and conversion to Arabic. ANLP 2014.

J. Eisner. Parameter estimation for probabilistic finite-state transducers. ACL 2002.

N. Habash, M. Diab, O. Rambow. Conventional orthography for dialectal Arabic. LREC 2012.

J. He, X. Wang, G. Neubig, T. Berg-Kirkpatrick. A probabilistic formulation of unsupervised text style transfer. ICLR 2020.

K. Knight, A. Nair, N. Rathod, K. Yamada. Unsupervised analysis for decipherment problems. COLING/ACL 2006.

G. Lample, A. Conneau, L. Denoyer, M. Ranzato. Unsupervised machine translation using monolingual corpora only. ICLR 2018.

X. Li, P. Michel, A. Anastasopoulos, Y. Belinkov, N. Durrani, O. Firat, P. Koehn, G. Neubig, J. Pino, H. Sajjad. Findings of the first shared task on machine translation robustness. WMT 2019

References

P. Liang, D. Klein. Online EM for unsupervised models. NAACL 2009.

C.-C. Lin, H. Zhu, M. R. Gormley, J. Eisner. Neural finite-state transducers: Beyond rational relations. NAACL 2019.

P. Makarov, T. Ruzsics, S. Clematide. Align and copy: UZH at SIGMORPHON 2017 shared task for morphological reinflection. SIGMORPHON 2017.

D. Nguyen, L. Rosseel, J. Grieve. On learning and representing social meaning in NLP: A sociolinguistic perspective. NAACL 2021.

P. Rastogi, R. Cotterell, J. Eisner. Weighting finite-state transductions with neural context. NAACL 2016.

B. Roark, R. Sproat, C. Allauzen, M. Riley, J. Sorensen, T. Tai. The OpenGrm open-source finite-state grammar software libraries. ACL 2012.

B. Roark, L. Wolf-Sonkin, C. Kirov, S. J. Mielke, C. Johny, I. Demirsahin, K. Hall. Processing South Asian languages written in the Latin script: The Dakshina dataset. LREC 2020

M. Ryskina, M. R. Gormley, T. Berg-Kirkpatrick. Phonetic and visual priors for decipherment of informal romanization. ACL 2020.

M. Ryskina, E. Hovy, T. Berg-Kirkpatrick, M. R. Gormley. Comparative error analysis in neural and finite-state models for unsupervised character-level transduction. To appear at SIGMORPHON 2021.

T. Shavrina, O. Shapovalova. To the methodology of corpus construction for machine learning: Taiga syntax tree corpus and parser. CORPORA 2017.

S. Wu, P. Shapiro, R. Cotterell. Hard non-monotonic attention for character-level transduction. EMNLP 2018.

S. Wu, R. Cotterell. Exact hard monotonic attention for character-level transduction. ACL 2019.

S. Wu, R. Cotterell, M. Hulden. Applying the Transformer to character-level transduction. EACL 2021.