Phonetic and Visual Priors for Decipherment of Informal Romanization

Maria Ryskina, Matthew R. Gormley, Taylor Berg-Kirkpatrick

Carnegie Mellon University
Language Technologies Institute

ML
MACHINE LEARNING
DEPARTMENT

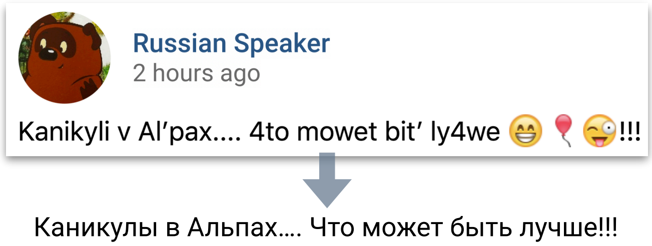
UCSDCSE
Computer Science and Engineering

In a Nutshell…

- Decoding transliteration in social media
- Inductive bias: character similarity
- Unsupervised finite-state approach
- New dataset of romanized Russian
- Performance on par with supervised!

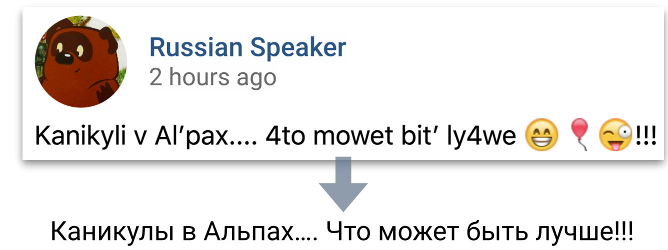
In a Nutshell…

- Decoding transliteration in social media

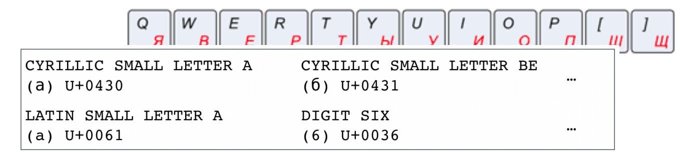


- Inductive bias: character similarity
- Unsupervised finite-state approach
- New dataset of romanized Russian
- Performance on par with supervised!

In a Nutshell…



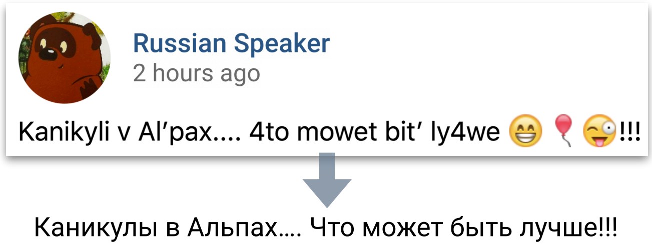
- Decoding transliteration in social media

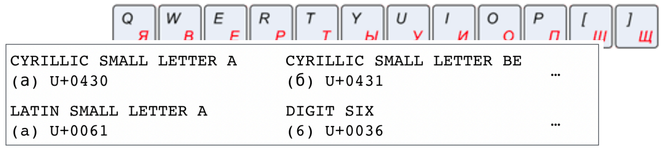


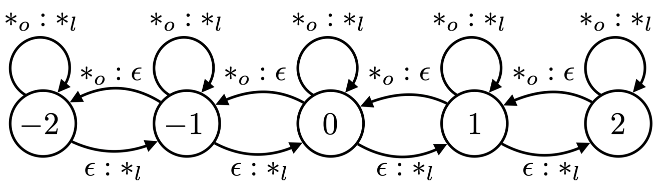
- Inductive bias: character similarity
- Unsupervised finite-state approach
- New dataset of romanized Russian
- Performance on par with supervised!

In a Nutshell…

- Decoding transliteration in social media
- Inductive bias: character similarity
- Unsupervised finite-state approach
- New dataset of romanized Russian
- Performance on par with supervised!

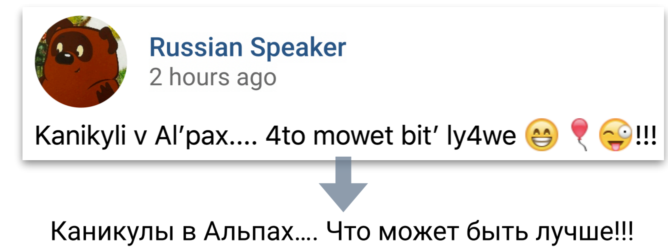




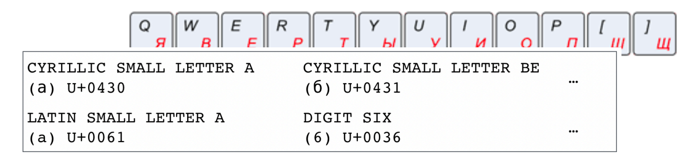


In a Nutshell…

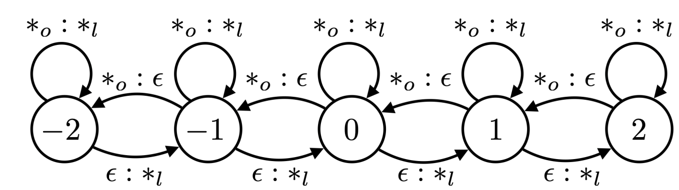
- Decoding transliteration in social media



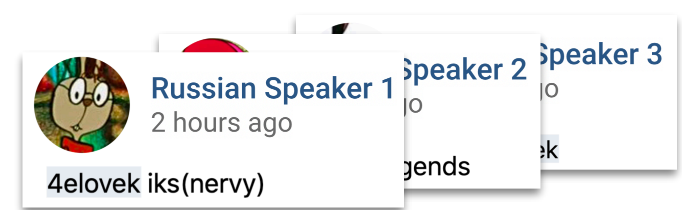
- Inductive bias: character similarity



- Unsupervised finite-state approach



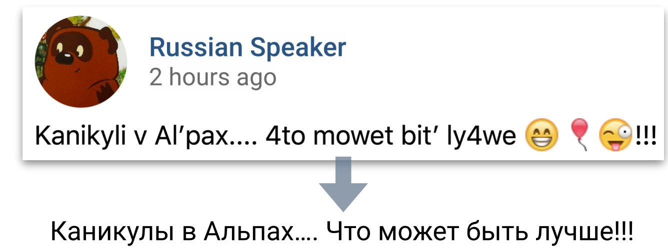
- New dataset of romanized Russian



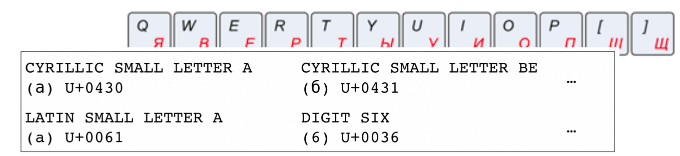
- Performance on par with supervised!

In a Nutshell…

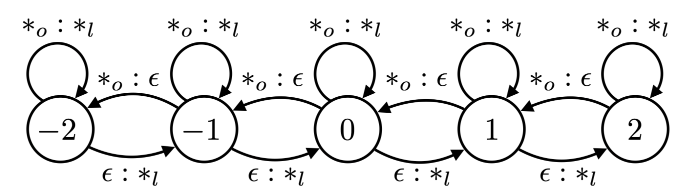
- Decoding transliteration in social media



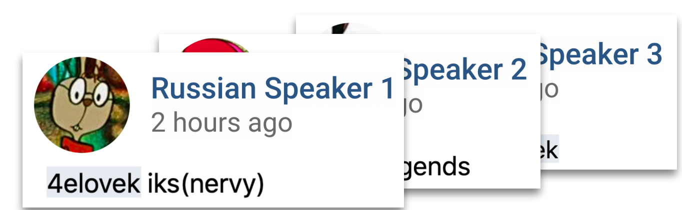
- Inductive bias: character similarity



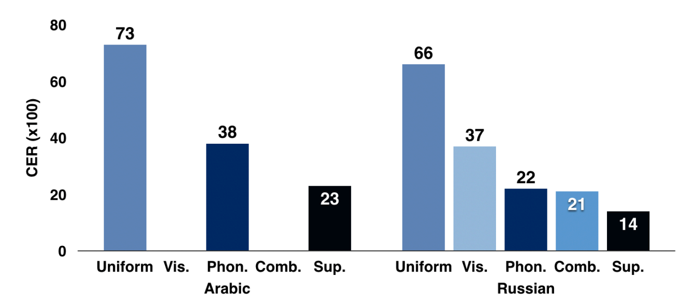
- Unsupervised finite-state approach



- New dataset of romanized Russian



- Performance on par with supervised!



Informal Romanization

Russian Speaker
2 hours ago

Kanikyli v Al’pax.... 4to mowet bit’ ly4we 😁 🎈 😜!!!

Informal Romanization

Russian Speaker
2 hours ago

Kanikyli v Al’pax.... 4to mowet bit’ ly4we 😁 🎈 😜!!!

What they type (Latin):	`Kanikyli v Al’pax`
What they mean (Cyrillic):	`Каникулы в Альпах`
English translation:	Vacation in the Alps

Informal Romanization

Russian Speaker
2 hours ago

Kanikyli v Al’pax…. 4to mowet bit’ ly4we !!!

observed

What they type (Latin):	`Kanikyli v Al’pax`
What they mean (Cyrillic):	`Каникулы в Альпах`
English translation:	Vacation in the Alps

goal

Visual and Phonetic Patterns

- Some character substitutions are phonetic…

Latin:	Kanikyli v Al’pax
Cyrillic:	Каникулы в Альпах
	/n/ /p/

+ Arabic ش /ʃ/ → **sh**, Russian **м /m/** → **m**, etc.

Visual and Phonetic Patterns

- Some character substitutions are phonetic…
- …and some are visual

Latin:	Kanikyli v Al’pax
Cyrillic:	Каникулы в Альпах
	/u/ /x/

+ Arabic ع /ʕ/ → 3, Russian в /v/ → B, etc.

Visual and Phonetic Patterns

- Some character substitutions are phonetic…
- …and some are visual
- It is a many-to-many cipher that also varies across users

Latin:	`Kanikyli v Al'pax`
Cyrillic:	`Каникулы в Альпах`
	/i/ /ɨ/

+ Arabic ص → s ← س, Russian 8 ← в → B, etc.

Problem

- Parallel data does not occur naturally ⇒ unsupervised learning

Problem

- Parallel data does not occur naturally ⇒ unsupervised learning
- Despite user variation, we can assume that the notions of similarity are shared

Problem

- Parallel data does not occur naturally ⇒ unsupervised learning
- Despite user variation, we can assume that the notions of similarity are shared

- **Hypothesis**: inductive bias encoding these similarity notions provides signal that can approximate human supervision

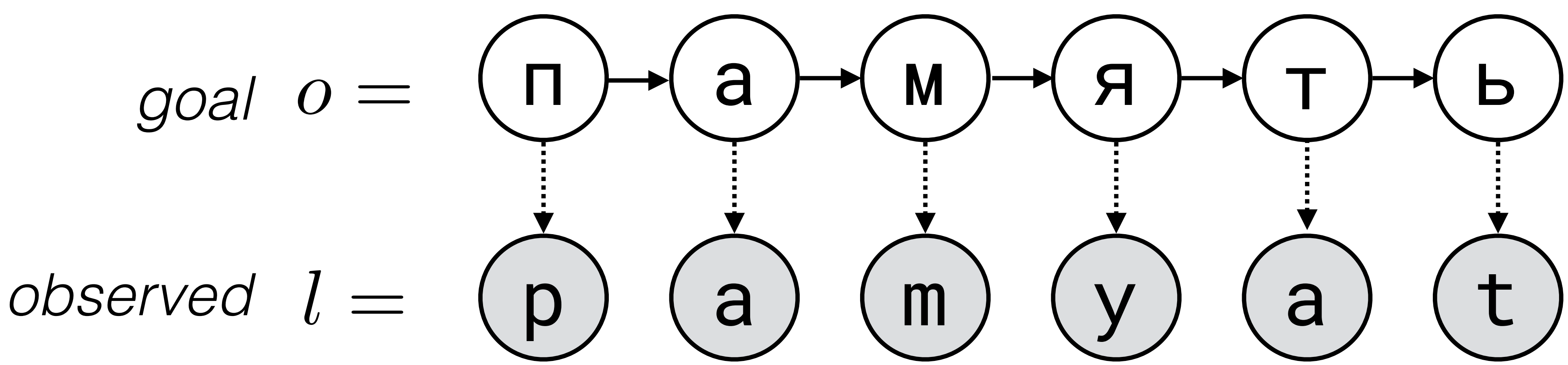
Noisy Channel Model

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

transition probabilities

emission probabilities

prior on parameters



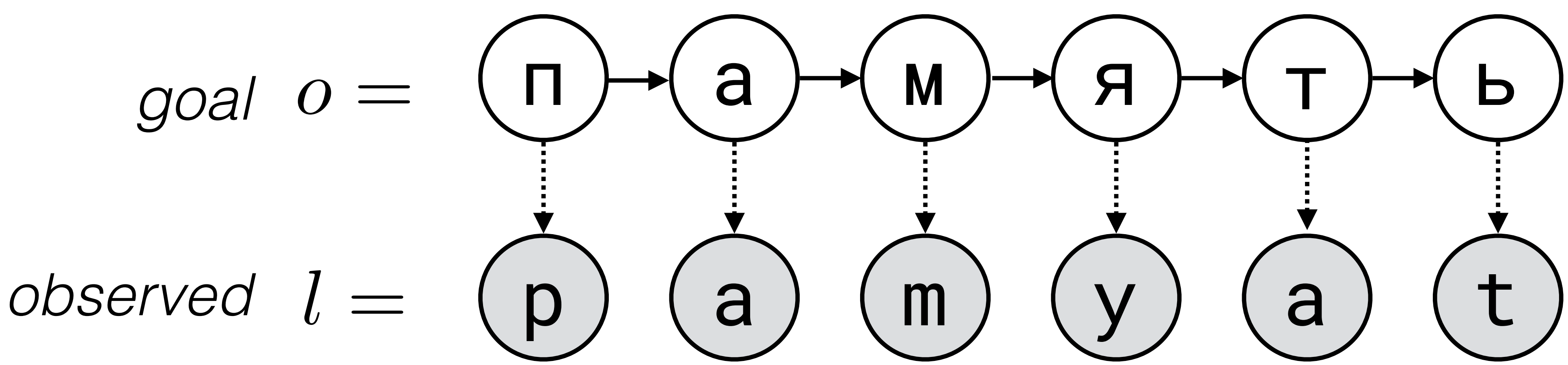
Noisy Channel Model

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

emission probabilities

transition probabilities

prior on parameters



Inductive Bias

- Phonetic prior: read mappings off phonetic keyboard layouts

Inductive Bias

- Phonetic prior: read mappings off phonetic keyboard layouts

Inductive Bias

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

Inductive Bias

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

Inductive Bias

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

Inductive Bias

- Visual prior: read mappings off Unicode confusable symbols list

CYRILLIC SMALL LETTER A (а) U+0430	CYRILLIC SMALL LETTER BE (б) U+0431	…
LATIN SMALL LETTER A (a) U+0061	DIGIT SIX (6) U+0036	…

Inductive Bias

- Visual prior: read mappings off Unicode confusable symbols list

CYRILLIC SMALL LETTER A (а) U+0430	CYRILLIC SMALL LETTER BE (б) U+0431	...
LATIN SMALL LETTER A (a) U+0061	DIGIT SIX (6) U+0036	...

  - Designed to combat spoofing attacks:

Inductive Bias

- Visual prior: read mappings off Unicode confusable symbols list

CYRILLIC SMALL LETTER A (а) U+0430	CYRILLIC SMALL LETTER BE (б) U+0431	…
LATIN SMALL LETTER A (a) U+0061	DIGIT SIX (6) U+0036	…

  - Designed to combat spoofing attacks:

acl2020.org → ACL 2020
U+0061 U+0063 …

асl2020.org → ???
U+0430 U+0441 …

Inductive Bias

- Visual prior: read mappings off Unicode confusable symbols list

CYRILLIC SMALL LETTER A (а) U+0430	CYRILLIC SMALL LETTER BE (б) U+0431	…
LATIN SMALL LETTER A (a) U+0061	DIGIT SIX (6) U+0036	…

  - Designed to combat spoofing attacks:

acl2020.org
U+0061 U+0063 …
→ ACL 2020

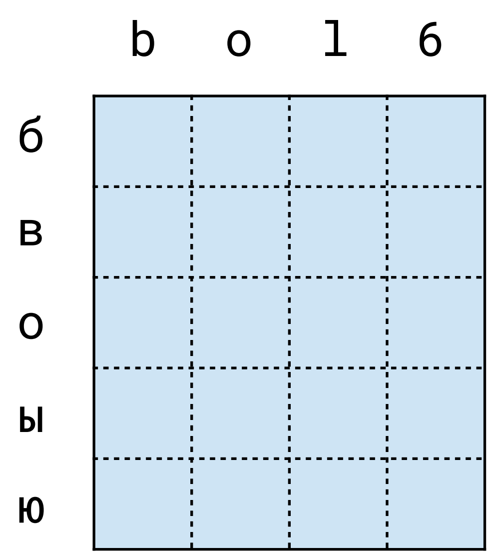
асl2020.org
U+0430 U+0441 …
→ ???

  - No Arabic—Latin mappings due to script dissimilarity

Inductive Bias

- Use mappings of similar characters as priors on emission parameters

$$c_l | c_o \sim \mathrm{Mult}(\theta_{c_o}) \qquad \theta \sim \mathrm{Dir}(\alpha)$$



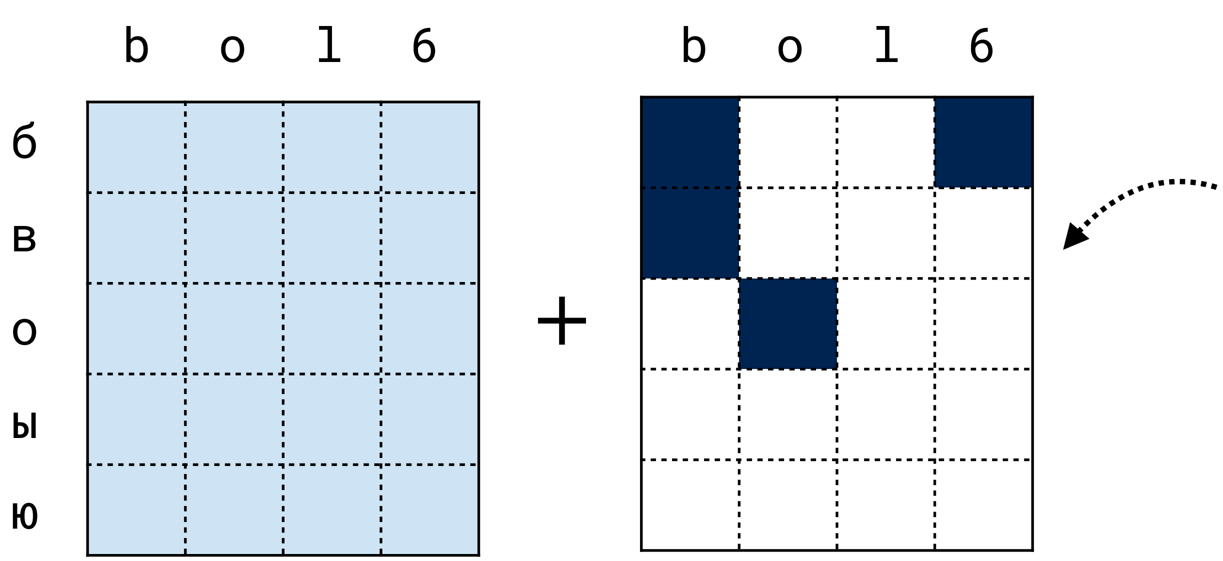
б→b
б→6
в→B
о→o
...

ю→IO
ы→bI
...

Inductive Bias

- Use mappings of similar characters as priors on emission parameters

$$c_l | c_o \sim \mathrm{Mult}(\theta_{c_o}) \qquad \theta \sim \mathrm{Dir}(\alpha)$$

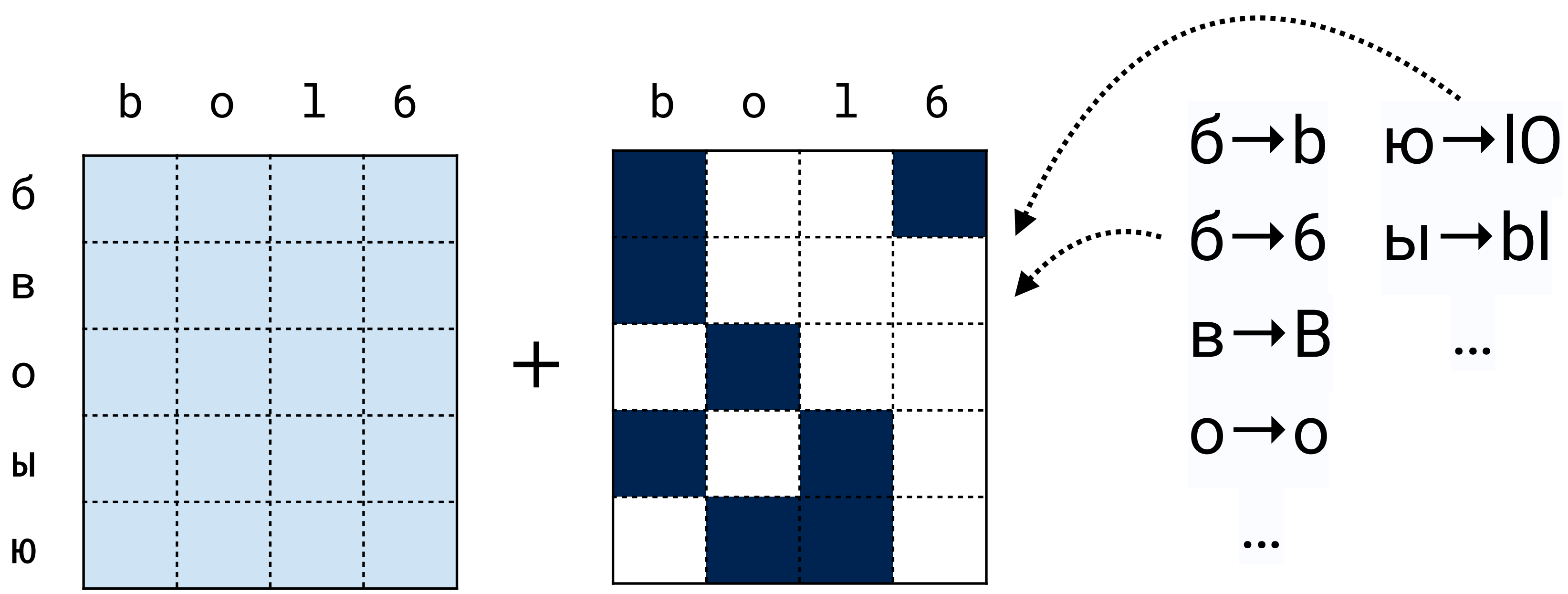


б→b ю→IO
б→6 ы→bI
в→B ...
о→o
...

Inductive Bias

- Use mappings of similar characters as priors on emission parameters

$$c_l | c_o \sim \mathrm{Mult}(\theta_{c_o}) \qquad \theta \sim \mathrm{Dir}(\alpha)$$



Noisy Channel Model

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

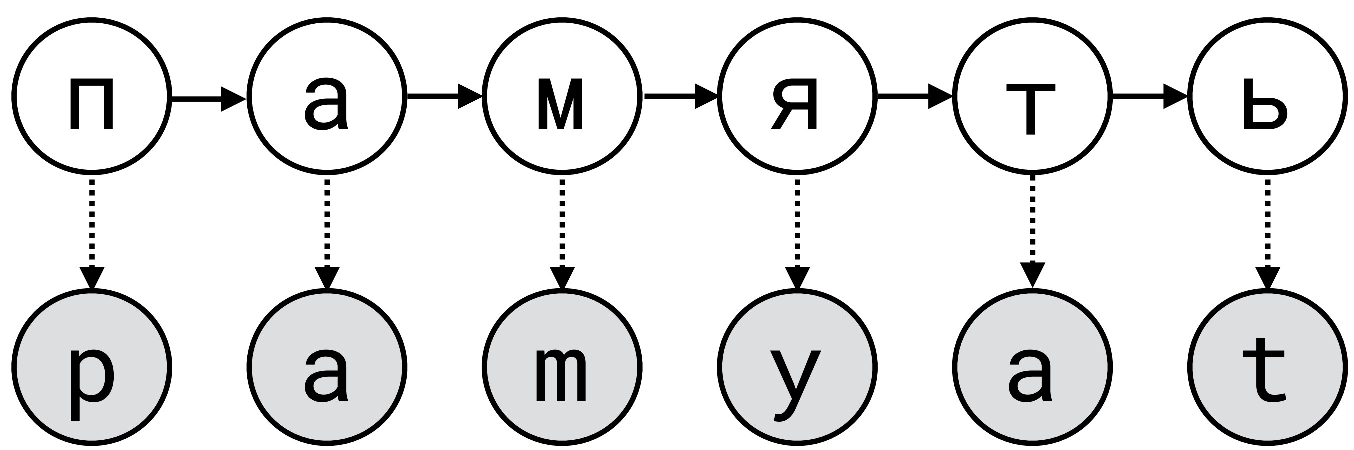
emission probabilities

transition probabilities

prior on parameters

goal $o =$

observed $l =$



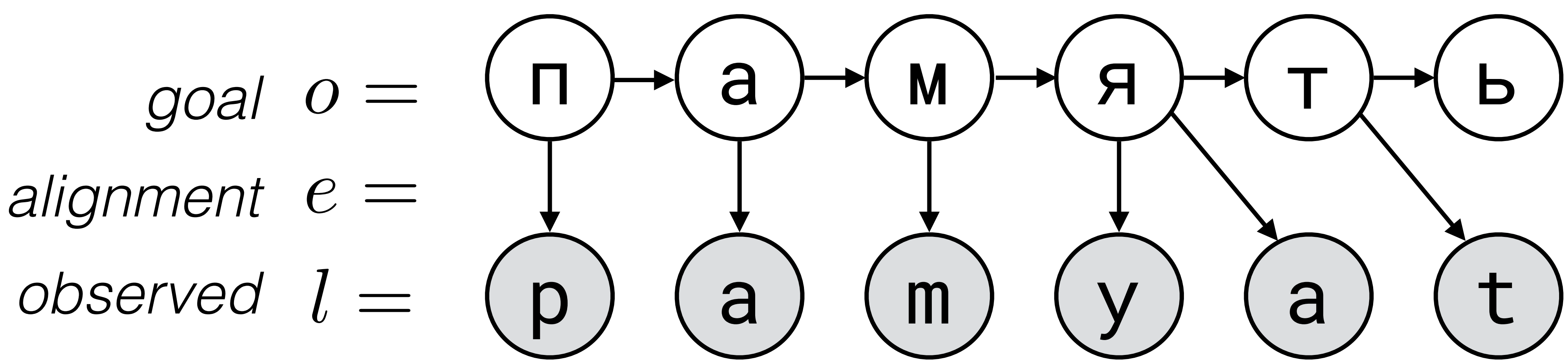
Noisy Channel Model

$$p(l) = \sum_{o,e} p(o;\gamma) \cdot p(l, e | o; \theta) \cdot p_{\text{prior}}(\theta; \alpha)$$

emission probabilities

transition probabilities

prior on parameters



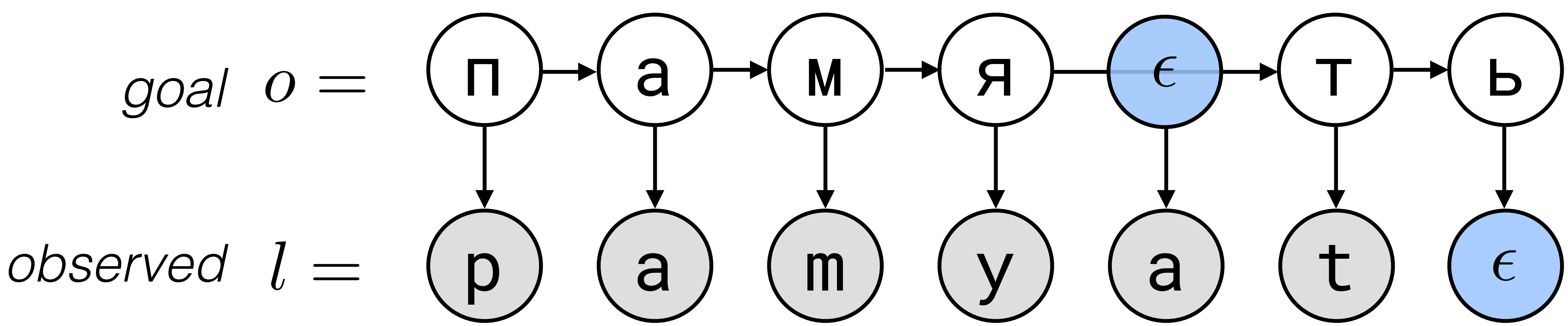
Noisy Channel Model

$$p(l) = \sum_{o,e} p(o;\gamma) \cdot p(l, e | o; \theta) \cdot p_{\mathrm{prior}}(\theta; \alpha)$$

emission probabilities

transition probabilities

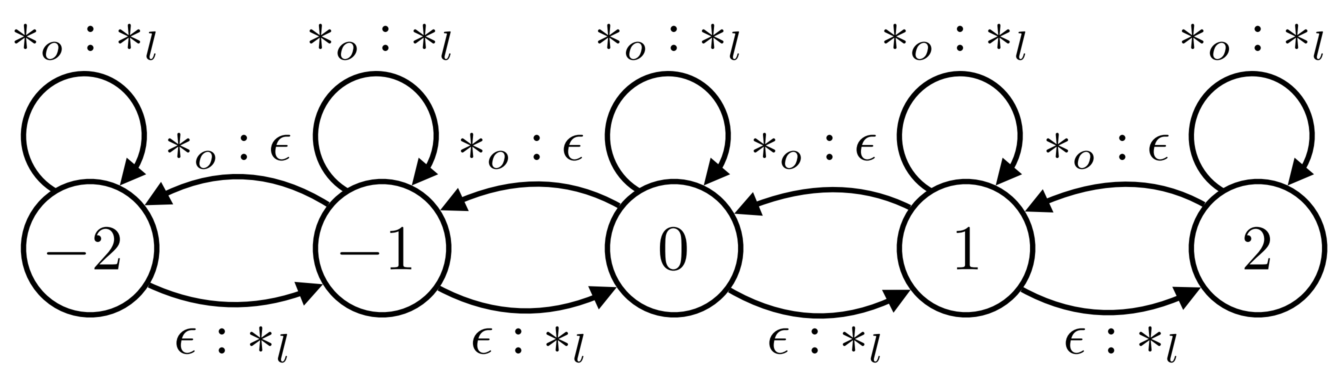
prior on parameters



Representing latent alignment via insertions and deletions

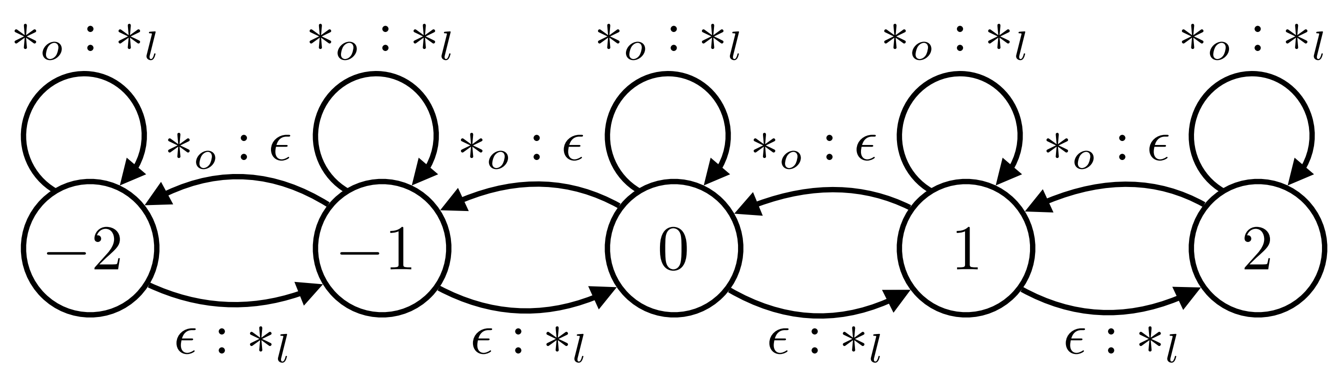
WFST Cascade

- Transition model: original script n-gram LM
- Emission model: WFST supporting deletions and insertions, with limited <u>delay</u>



WFST Cascade

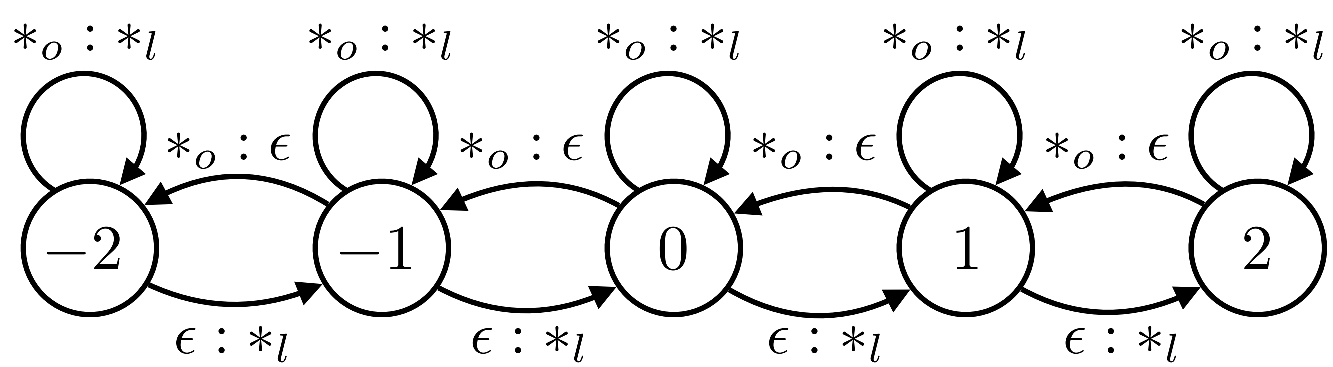
- Transition model: original script n-gram LM
- Emission model: WFST supporting deletions and insertions, with limited delay



difference between the numbers of insertions and deletions at each timestep

WFST Cascade

- Transition model: original script n-gram LM
- Emission model: WFST supporting deletions and insertions, with limited <u>delay</u>

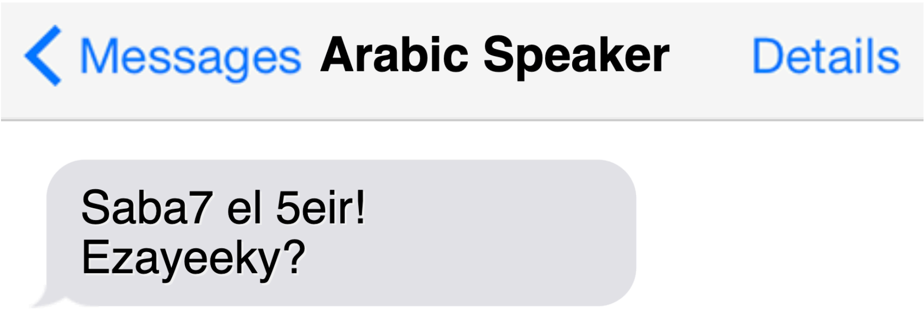


difference between the numbers of insertions and deletions at each timestep

- Trained with EM algorithm + stepwise training, curriculum learning, pruning…

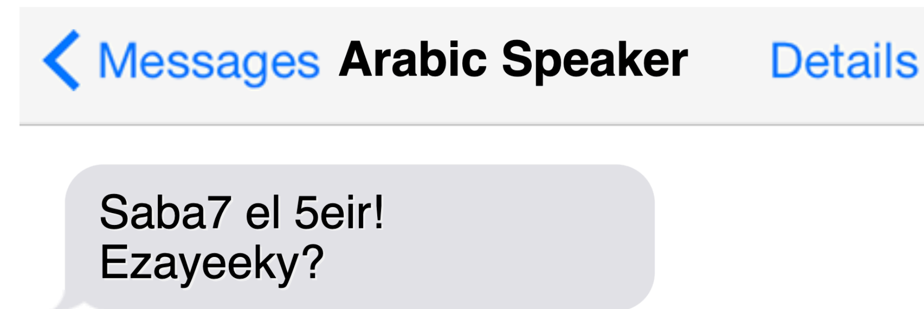
Data

- Arabic: LDC BOLT dataset (SMS / chat dialogs)



Data

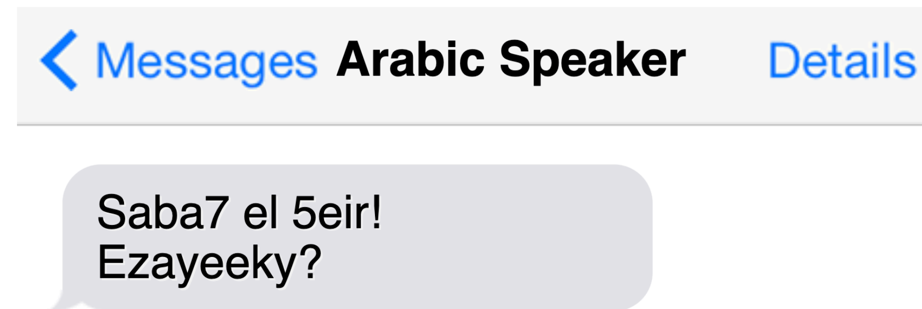
- Arabic: LDC BOLT dataset (SMS / chat dialogs)



- Russian: collect social media data using transliterations of frequent words as queries

Data

- Arabic: LDC BOLT dataset (SMS / chat dialogs)

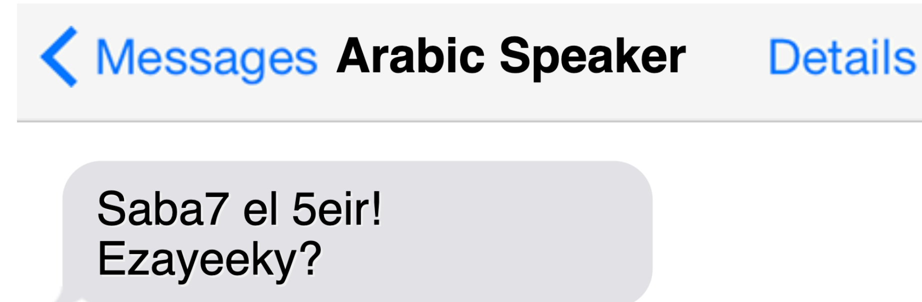


- Russian: collect social media data using transliterations of frequent words as queries

человек → 4elovek, chelovec, 4eJloBek, …

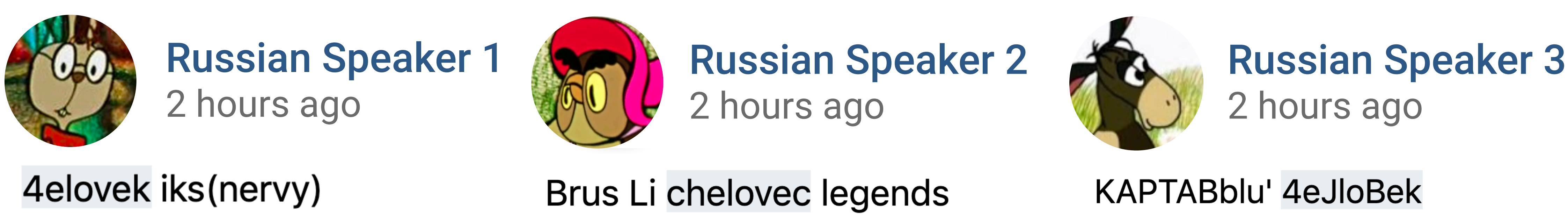
Data

- Arabic: LDC BOLT dataset (SMS / chat dialogs)



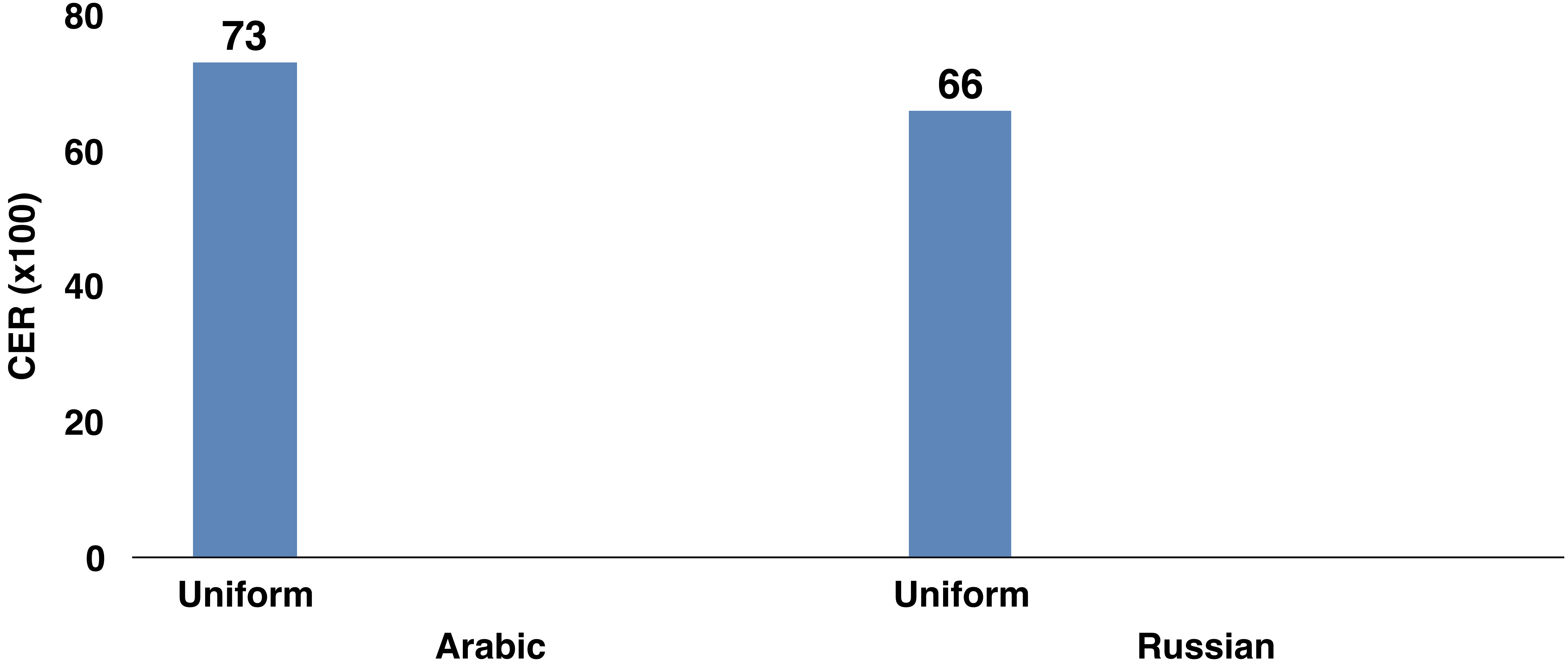
- Russian: collect social media data using transliterations of frequent words as queries

человек → 4elovek, chelovec, 4eJloBek, ...



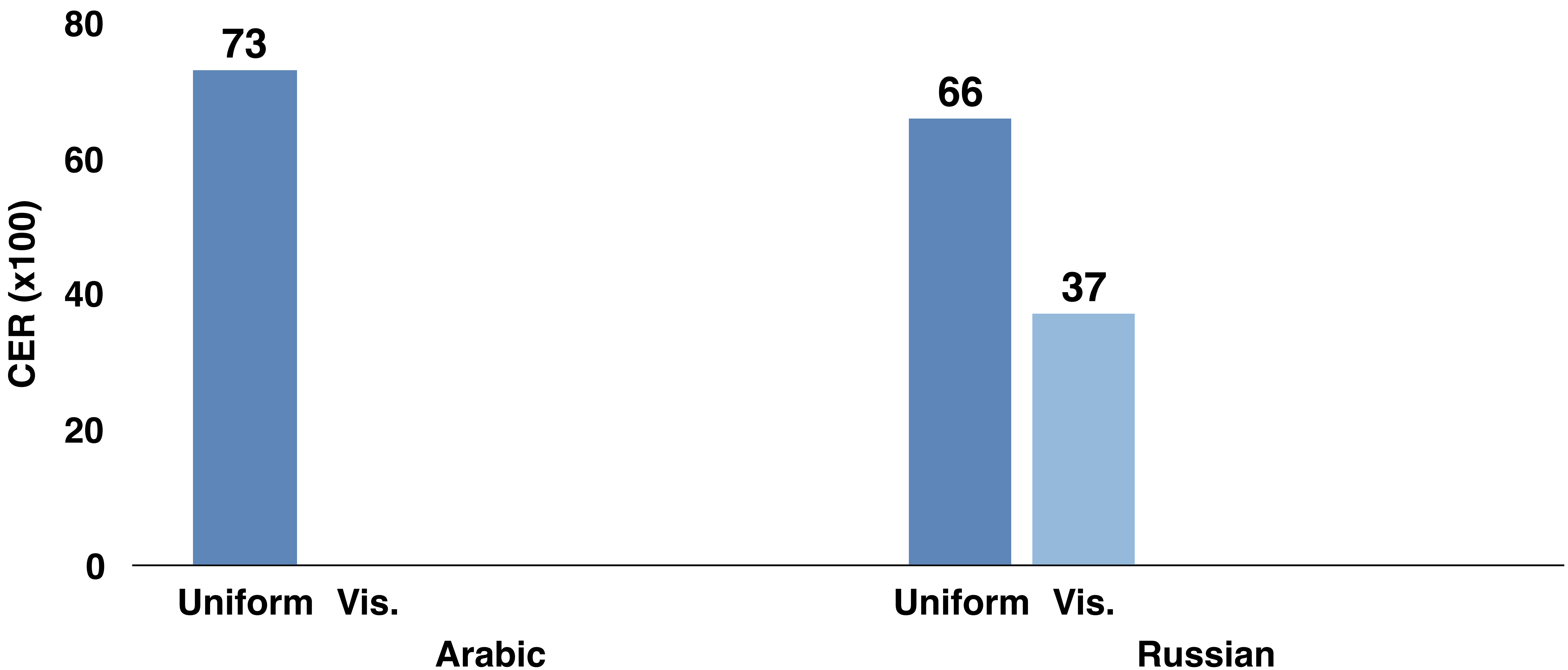
Results

- Unsupervised model without inductive bias has high error rate



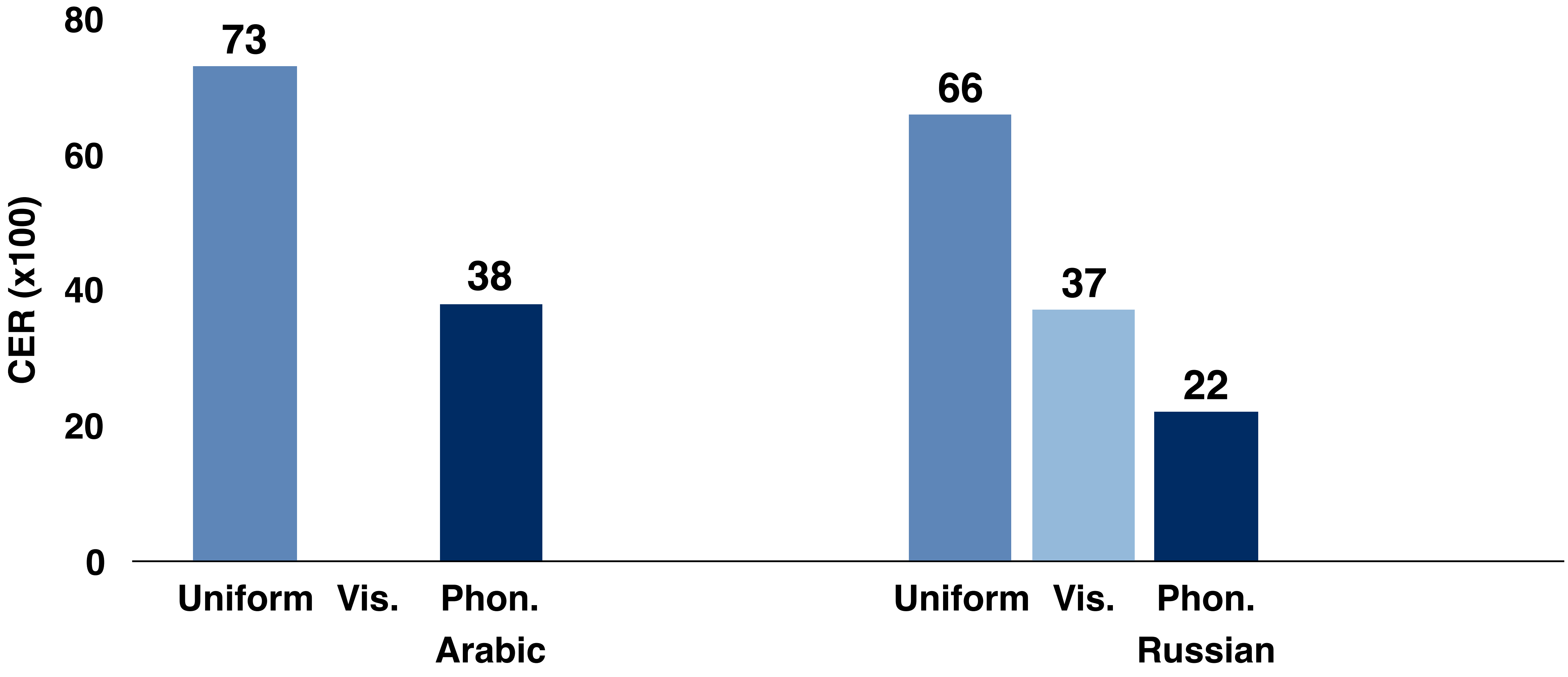
Results

- Unsupervised model without inductive bias has high error rate
- Even a sparse visual prior reduces error rate by almost a half



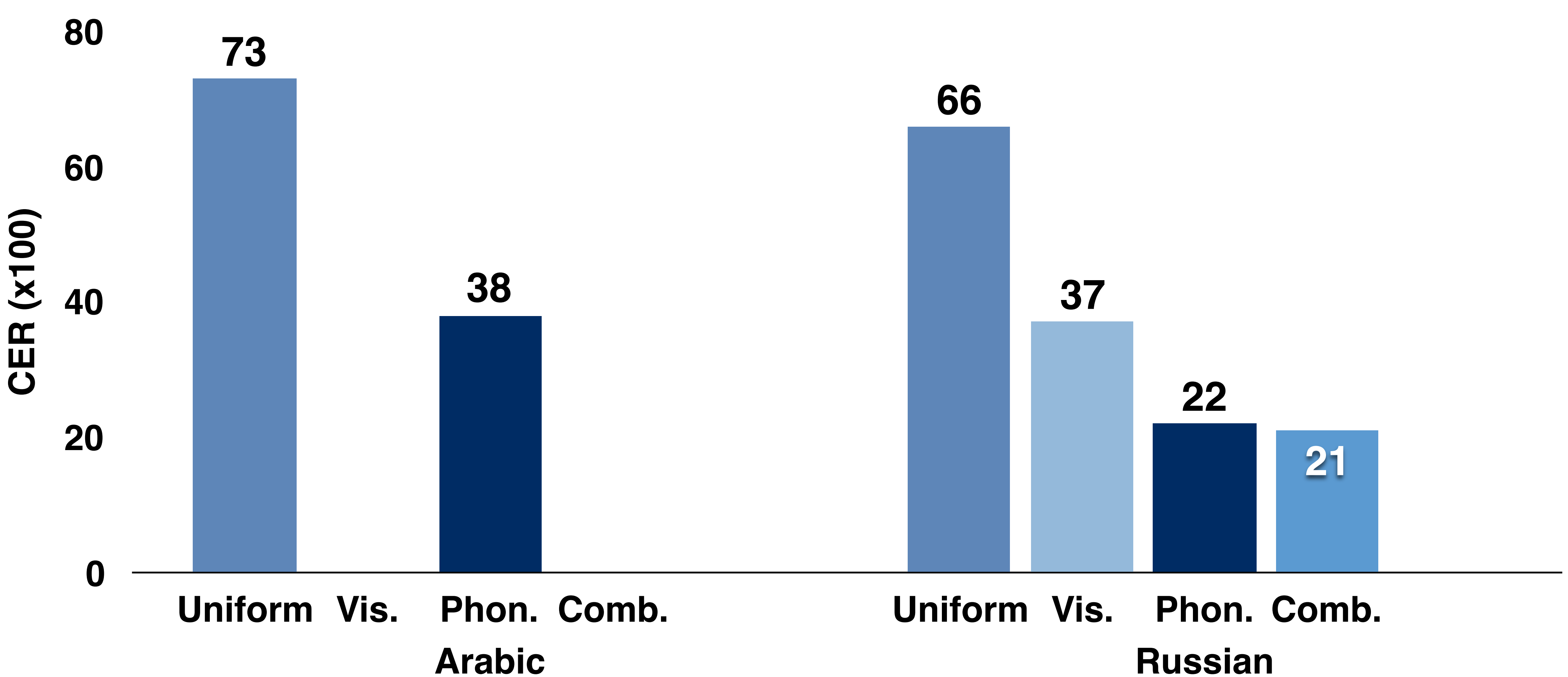
Results

- Unsupervised model without inductive bias has high error rate
- Even a sparse visual prior reduces error rate by almost a half
- Phonetic prior is better (more mappings + pattern more frequent)



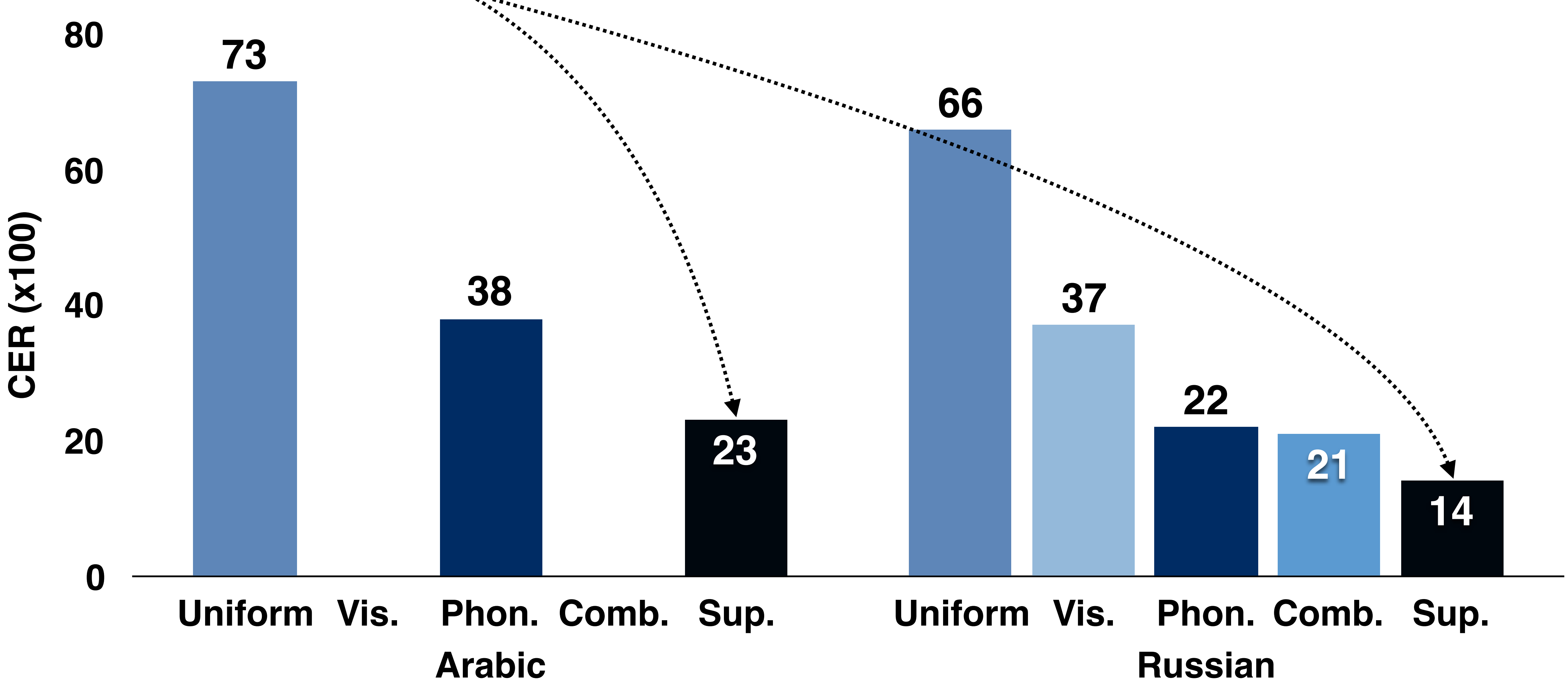
Results

- Unsupervised model without inductive bias has high error rate
- Even a sparse visual prior reduces error rate by almost a half
- Phonetic prior is better (more mappings + pattern more frequent)
- Combining phonetic and visual mappings yields best CER



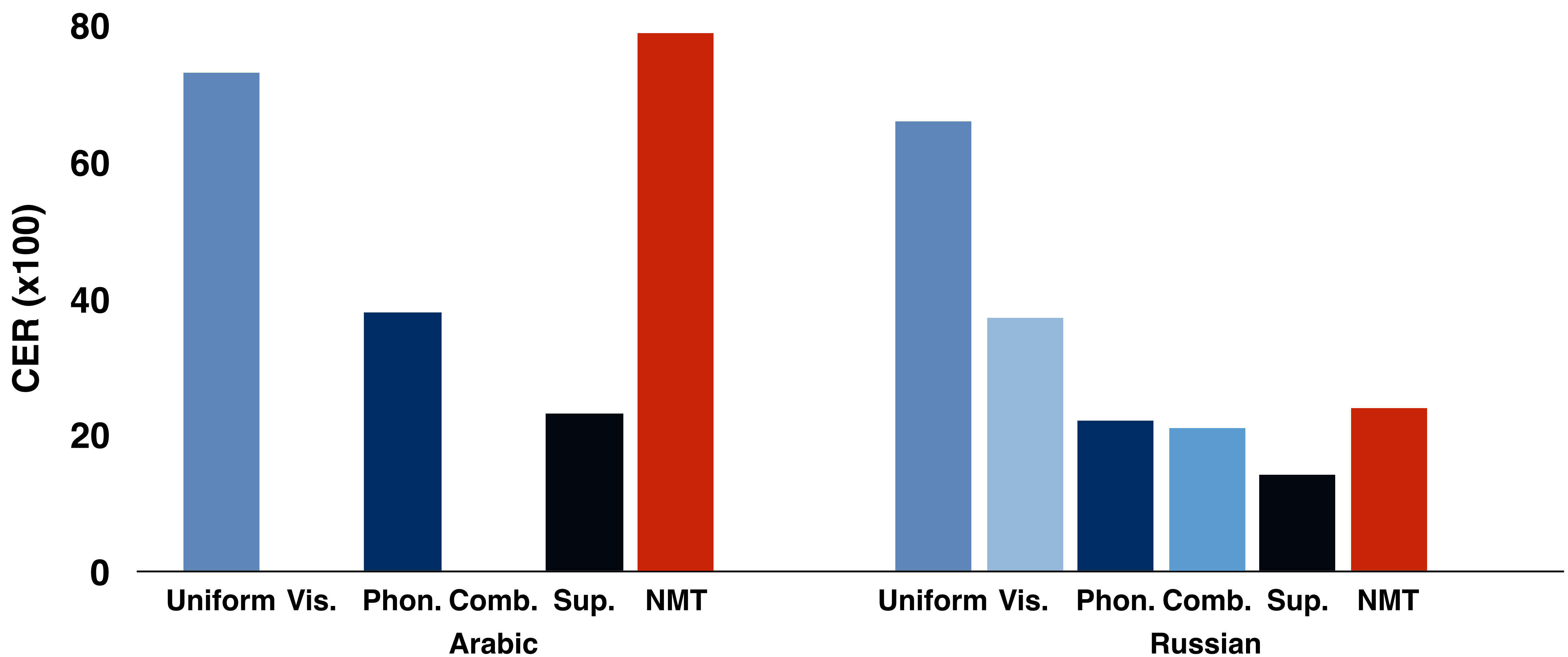
Results

- Unsupervised model without inductive bias has high error rate
- Even a sparse visual prior reduces error rate by almost a half
- Phonetic prior is better (more mappings + pattern more frequent)
- Combining phonetic and visual mappings yields best CER
- Supervised skyline compares effect of annotation vs. priors



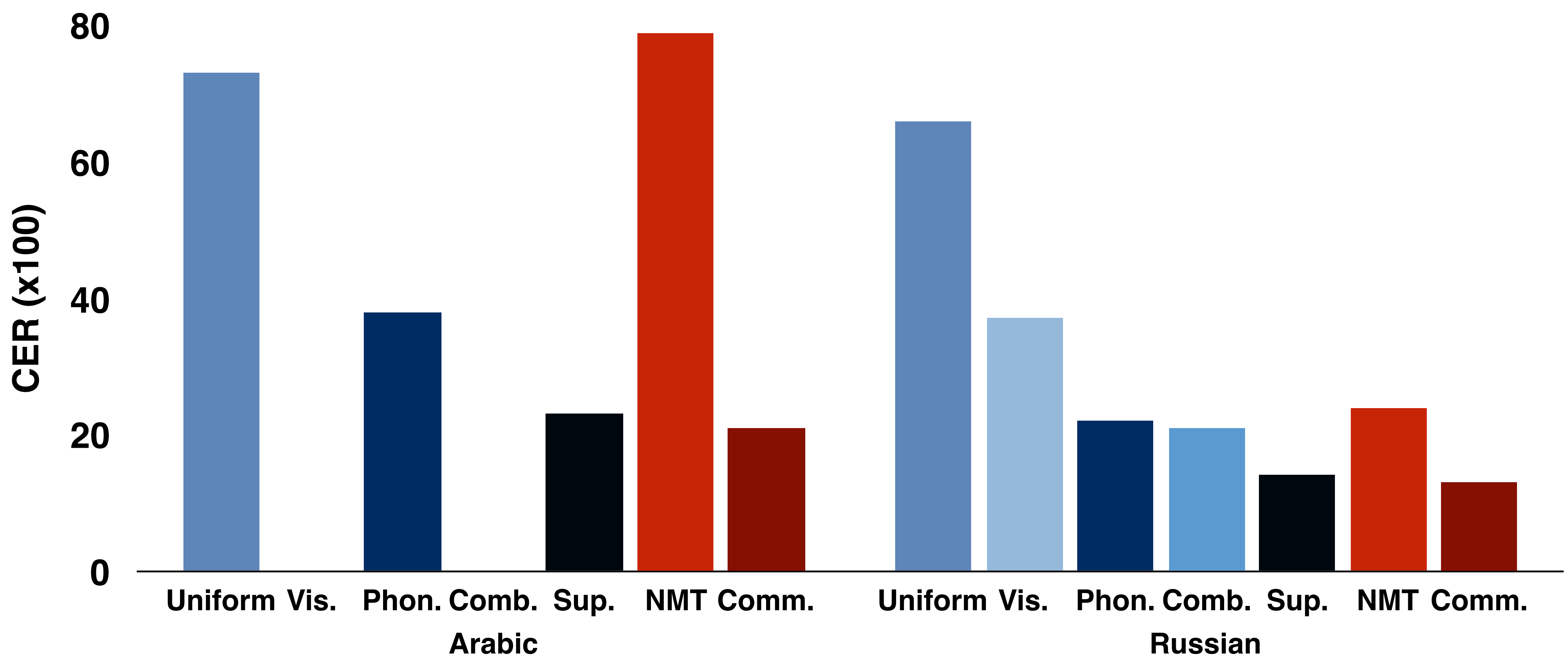
Other Baselines

- Unsupervised neural MT (Lample et al.) trained on characters works for Russian



Other Baselines

- Unsupervised neural MT (Lample et al.) trained on characters works for Russian
- Nothing beats commercial hand-built systems



Conclusion

Conclusion

- We propose a noisy-channel unsupervised WFST model to convert informal romanization into original script

Conclusion

- We propose a noisy-channel unsupervised WFST model to convert informal romanization into original script
- We present a dataset of informally romanized Russian

Conclusion

- We propose a noisy-channel unsupervised WFST model to convert informal romanization into original script
- We present a dataset of informally romanized Russian
- We show that similarity priors induce a substantial amount of supervision contained in human annotation

Conclusion

- We propose a noisy-channel unsupervised WFST model to convert informal romanization into original script
- We present a dataset of informally romanized Russian
- We show that similarity priors induce a substantial amount of supervision contained in human annotation
- Future work
  - Explicitly operationalizing character similarity
  - User-specific substitution preferences

Questions?

Q&A sessions:

- July 8, 17:00 UTC+0 (1pm EDT)
- July 8, 21:00 UTC+0 (5pm EDT)

github.com/ryskina/romanization-decipherment

mryskina@cs.cmu.edu

@maria_ryskina