Inria-ALMAnaCH at the WMT 2022 shared task: Does Transcription Help Cross-Script Machine Translation?

Jesujoba O. Alabi[1] Lydia Nishimwe[2] Benjamin Muller[2,3] Camille Rey[2]
Benoît Sagot[2] Rachel Bawden[2]

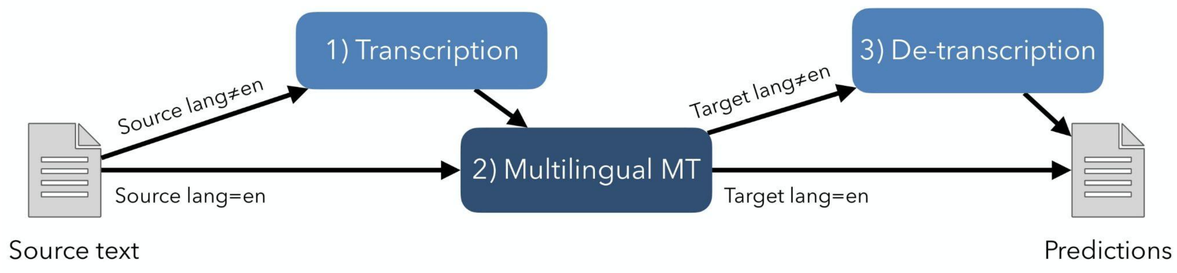
1. Spoken Language Systems (LSV), Saarland University, Saarland Informatics Campus, Germany
2. Inria, Paris, France 3. Sorbonne Université, France

Approach: linguistically motivated transcription

- Multilingual translation between Slavic languages: **{cs,ru,uk}↔en and cs↔uk**
- Encourage cross-lingual sharing and transfer using linguistically inspired cross-script transcription

cs	original	Sníh pokryl stromy vedle zámku.
cs	transcribed (cl)	Snig pokril stromi vedle zamku.
uk	original	Сніг вкрив дерева біля замку.
uk	transliterated	Snih vkryv dereva bilja zamku.
uk	transcribed (ul)	Sneg vkriv dereva bela zamku.
ru	original	Снег покрыл деревья возле замка.
ru	transliterated	Sneg pokryl derev'ja vozle zamka.
ru	transcribed (rl)	Sneg pokril dereva vozle zamka.
en	original	The snow has covered the trees next to the castle.

Translation process

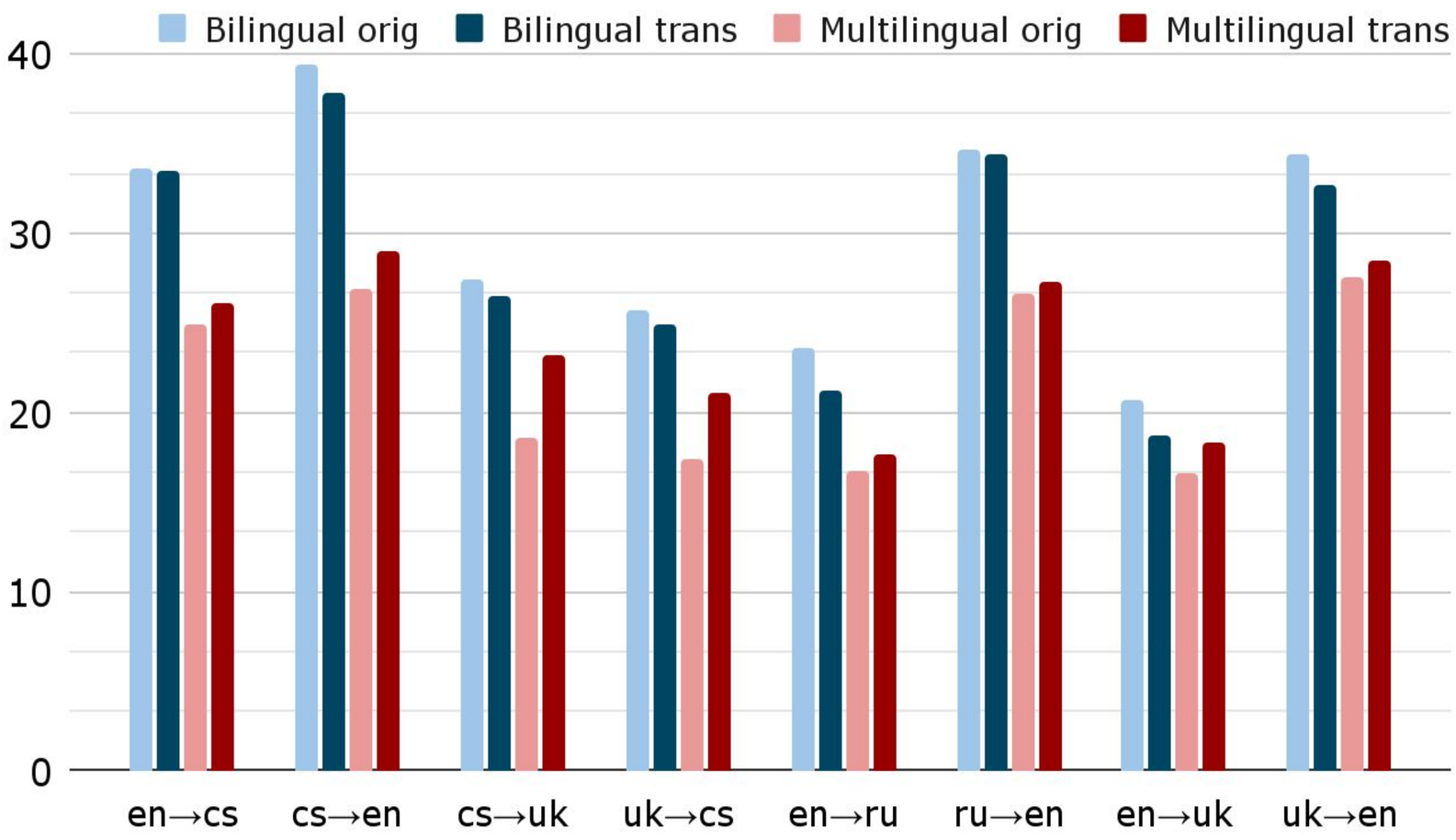


Dataset and training

Lang. pair	#sents
cs-en	54,495,258
cs-uk	2,490,622
en-ru	25,584,007
en-uk	22,322,394

- Constrained setting
- Bilingual and multilingual MT models (Fairseq, parallel data)
- Detranscription models for each Slavic language (Fairseq, ~20M sentences of monolingual data)

Results: comparison between original and transcribed texts (bilingual and multilingual settings)



Bilingual results:
- Generally worse results when using transcription
- Less problematic for en→cs

Multilingual results:
- Much worse than bilingual results
- Transcription helps to improve the multilingual setting somewhat

Why these results?
- Possibly the limited vocabulary size for multilingual models (64k)
- Increased sharing helps the vocabulary size issue
- Any gain from transfer does not counterbalance this issue
- Potential noise from transcription

Detranscription model results

	cl→cs	rl→ru	ul→uk
$\text{FLORES}_{\text{devtest}}$	97.49	94.74	96.29
WMT 2022 (src)	96.47	95.34	94.70
WMT 2022 (ref)	97.33	96.24	97.12

Take-aways

- Transcription harms translation performance between Slavic languages
  - regardless of the side where it is applied
- Multilingual models perform worse than monolingual models when vocabulary sizes are identical

Pre-transcription results

Compare the raw output of:
- the bilingual model with transcription (i.e. before detranscription)
- the bilingual baseline model with transcription applied

	en→cl	en→rl	en→ul
Bilingual with transcription			
$\text{FLORES}_{\text{devtest}}$	29.53	22.90	22.62
WMT 2022	34.06	22.56	19.27
Transcribing bilingual baseline's output			
$\text{FLORES}_{\text{devtest}}$	29.37	25.35	24.60
WMT 2022	33.87	23.75	20.94

- →{rl,ul}: transcription = worse
- →cl: transcription = better

Transcribing the source or target or both

Compare the effect of transcription on cs↔uk pair
- transcribe source, target or both
- transcription made **uk** look like **cs**

	none	source	target	both
cs→uk				
$\text{FLORES}_{\text{devtest}}$	19.76	19.64	19.32	19.05
WMT 2022	27.40	27.01	26.82	26.43
uk→cs				
$\text{FLORES}_{\text{devtest}}$	20.86	20.41	18.50	20.09
WMT 2022	25.65	25.15	25.41	24.96

- using no transcription on either side gave better result
- in some cases, transcribing source side works better than both sides