Try me! (Demo page)

Transliteration for Low-Resource Code-Switching Texts: Building an Automatic Cyrillic-to-Latin Converter for Tatar

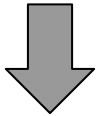
Chihiro Taguchi, Yusuke Sakai, and Taro Watanabe
Nara Institute of Science and Techonology
{taguchi.chihiro.td0, sakai.yusuke.sr9, taro}@is.naist.jp

Introduction

Try me! (Demo page)

Automatic transliteration tool from Cyrillic Tatar to Latin Tatar

Коронавирус инфекциясеннән прививкага ясатырга килгәндә үзең белән паспорт, полис, СНИЛС булырга тиеш.



Koronavirus infektsiyäsennän privivkağa yasatırğa kilgändä üzeñ belän pasport, polis, SNİLS bulırğa tiyeş.

* Words highlighted in blue are Russian. Source: https://tatar-inform.tatar/news/health/31-05-2021/kazanda-yuzhnyy-s-d-z-gend-vaktsinatsiya-punkty-kabat-achyldy-5825091

Two difficulties:

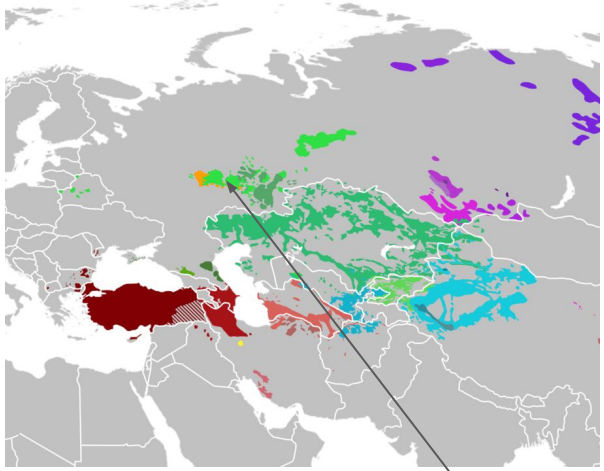
1. **Different transliteration rules** for Tatar-origin words and Russian words
2. **Low resource** for training a language identifier

Tatar: Linguistic Background

Try me! (Demo page)

Tatar < Kipchak < Turkic

- Spoken by 5 million people
- Mainly in the Republic of **Tatarstan**, Russia
- Commonly written in **Cyrillic**, while some diaspora communities write with the **Latin** orthography
- SOV, head-final, agglutinative



Distribution of Turkic languages. Light green indicates the Tatar-speaking area.
Source: Wikimedia Commons

Long history of language contact with Russian

- Most speakers are bilingual with Russian
- Frequent Russian loanwords and **code switching**

Two Orthographies in Modern Tatar

Try me! (Demo page)

Mainstream: Cyrillic

Russian alphabet + 6 extended letters (ә, ө, ү, ң, җ, һ)

Some diaspora communities: Latin

Based on the Common Turkic alphabet (English alphabet + ä, ı, ö, ü, ñ, ğ, ş, ç)

History

- Before 1928: Arabic-based orthography
- 1920–39: *Yañalif* Latin alphabet
- **1939–: Current Cyrillic alphabet**
- 1999: New Latin orthography proposed, rejected in 2002
- **2013–: Current Latin alphabet**

Tatar: Code Switching

Try me! (Demo page)

Language contact, bilingualism, language shift in the Tatar-speaking society

- Frequent code-switching
  e.g., (blue words are Russian)

Башка	урын	табарга	да	җиңел	түгел.	Алдан	забронировать	надо.
Başqa	urın	tabarğa	da	ciñel	tügel.	Aldan	zabronirovat	nado.
other	place	to find	also	easy	not	beforehand	to book	necessary

“It is not easy to find another place. It is necessary to book beforehand.”

- Intra-word code-switching (mixed morpheme)
  e.g., Әхмәтдиновка (Äxmätdinovqa) “To Äxmätdinov”

Related Work: Code Switching

Try me! (Demo page)

CS: Recent trends in NLP

- Srivastava+ 2020, Singh&Lefever 2020, etc.: Indic languages – English
- Alvarez-Mellado 2020, Claeser+ 2018, etc.: Spanish – English
- Hamed+ 2019, Samih&Maier 2016, etc.: Colloquial Arabic – MSA

Intra-word CS

- Mager+ 2019: German–Turkish, Spanish–Wixarika (segRNN: Lu+ 2016)
- Nguyen&Cornips 2016: Dutch–Limburgish (Morfessor: Creutz&Lagus 2006)
- Yirmibeşoğlu&Eryiğit 2018: Turkish–English (Character n-gram, CRF)

Related Work: Tatar Transliteration

Try me! (Demo page)

Transliteration tools:

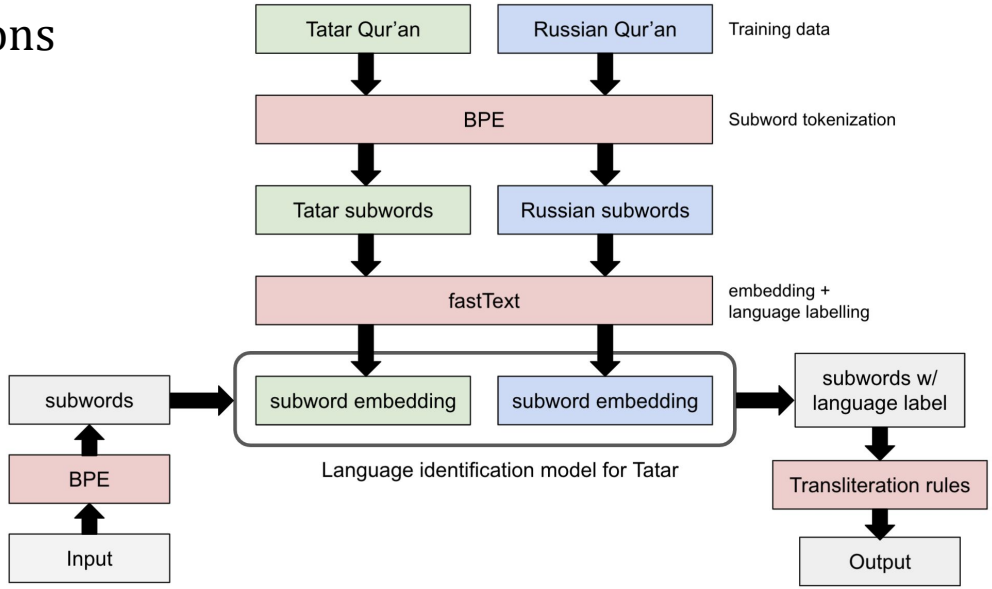
- Tatar Transcription Tool (TTT) (Bradley 2014)
  Rule-based, Tatar words only
- speak.tatar (anonymous)
  Rule-based, Tatar words only
- FinTat (Saykhunov+ 2019)
  Rule-based, using Finnish-Tatars' orthography
- Aylandirow (Korbanov, n.d.)
  Rule-based, both Tatar and Russian words

Method

Try me! (Demo page)

Train a binary language classifier: Tatar or Russian

- Monolingual corpus: **Qur'an** translations
  - Tatar: 19,691 words w/ dup.
  - Russian: 21,256 words w/ dup.
- Subword tokenization
  - **BPE** (Sennrich 2016)
  - Robust to the OOV problem
- Embedding for classification
  - **FastText** (Bojanowski+ 2017, Joulin+ 2017)



Experimental Setup

Try me! (Demo page)

Evaluation data

- 700 sentences from the Corpus of Written Tatar (Saykhunov+ 2019)
  - shuffled; 8,466 words w/ duplication
  - not public
- Corresponding sentences manually transcribed into the Latin alphabet
  - Russian morphemes are tagged

Evaluation metrics

- Character BLEU, longest common sequence (LCS) F-measure (Chen+ 2018), word accuracy (ACC), character error rate (CER)
- Baselines: speak.tatar, TTT, Aylandirow
- Proposed method: tt-ru hybrid
  + our Tatar-monolingual rule-based transliteration (tt-based)

Results

Try me! (Demo page)

	BLEU	LCS F-score	CER	ACC	# correct sentence	# error word
speak.tatar	0.869	0.953	0.049	0.952	67	1,747
TTT	0.879	0.956	0.054	0.946	121	1,505
Aylandirow	0.971	**0.994**	0.009	0.991	362	526
tt-based	0.968	0.989	0.011	0.989	365	552
tt-ru hybrid	**0.981**	**0.994**	**0.007**	**0.993**	**437**	**332**

tt-ru hybrid outscores the baselines

- Aylandirow's high score in the **LCS F-score**
  Aylandirow's transliteration is closer to the gold data character-wise
- Low scores in monolingual transliterators (speak.tatar, TTT, tt-based)
  tt-based's high score is merely a result of the orthographic consistency

Analysis (1)

Try me! (Demo page)

	BLEU	LCS F-score	CER	ACC	# correct sentence	# error word
speak.tatar	0.869	0.953	0.049	0.952	67	1,747
TTT	0.879	0.956	0.054	0.946	121	1,505
Aylandirow	0.971	**0.994**	0.009	0.991	362	526
tt-based	0.968	0.989	0.011	0.989	365	552
tt-ru hybrid	**0.981**	**0.994**	**0.007**	**0.993**	**437**	**332**

Difference among monolingual transliterators:

- Some rules in speak.tatar and TTT do not follow the current Latin orthography
- Orthographic variations are actually observed among users

Analysis (2)

Try me! (Demo page)

- Improvement from tt-based to tt-ru hybrid

  Accuracy:

  Russian words: 46.7% to **78.1**%

  intra-CS words: 48.6% to **76.7**%

tt-ru hybrid can detect some Russian words

	ru words		CS words	
	#	accuracy	#	accuracy
speak.tatar	**805**	**0.798**	455	0.752
TTT	258	0.256	175	0.283
Aylandirow	738	0.731	461	0.762
tt-based	471	0.467	294	0.486
tt-ru hybrid	788	0.781	**464**	**0.767**

A performance comparison for Russian CS words. High accuracy of speak.tatar arises from the similarity of its translit. rules to Russian words. Instead, it fails to transliterate Tatar words correctly.

Source	РФ Закон чыгаручылар советы президиумы утырышында катнашты.
tt-based	RF Zaqon çığaruçılar sowetı prezidiumı utırışında qatnaştı.
tt-ru hybrid	RF Zakon çığaruçılar sovetı prezidiumı utırışında qatnaştı.

The underlined words are wrong in tt-based but are successfully transliterated in tt-ru hybrid

Analysis (3): Negative result

Try me! (Demo page)

- Some Tatar words were mistakenly identified as Russian
  tt-ru hybrid mistakenly transliterated 116 words
  that were correct in tt-based

- Similar character sequences may cause confusion
  in language classification

set (total 5,261 words)	# words
$V(T)$	552
$V(H)$	332
$V(T) \cap V(H)$	216
$V(T) \setminus V(H)$	336
$V(H) \setminus V(T)$	116

V(T): set of error words in tt-based
V(H): set of error words in tt-ru hybrid

Source	Һәр юлчы туктап, аның хозурлыгына сокланып китә.
tt-based	Här yulçı tuqtap, anıñ xozurlığına soqlanıp kitä.
tt-ru hybrid	Här yulçı tuqtap, anıñ xozurlığına soklanıp kitä.

The underlined word is wrong in tt-ru hybrid but is correctly transliterated in tt-based

Conclusion

Try me! (Demo page)

Transliteration quality

- our tool has the **highest accuracy** in transliteration overall
- it **detects Russian** morphemes successfully to some extent

Advantages of our method

- The model is trained only on **low-resource** corpus
- It employs **language-agnostic** approaches (BPE, fastText)

  Applicable to other CS language pairs

References

- Elena Alvarez-Mellado. 2020. An annotated corpus of emerging anglicisms in Spanish newspaper headlines.
- Tamali Banerjee and Pushpak Bhattacharyya. 2018. Meaningless yet meaningful: Morphology grounded subword-level NMT.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information.
- Jeremy Bradley. 2014. Tatar transcription tool. URL: https://www.univie.ac.at/maridict/site-2014
- Nancy Chen, Rafael E. Banchs, Min Zhang, Xiangyu Duan, and Haizhou Li. 2018. Report of NEWS 2018 named entity transliteration shared task.
- Daniel Claeser, Samantha Kent, and Dennis Felske. 2018. Multilingual named entity recognition on Spanish–English code-switched tweets using support vector machines.
- Mathias Creutz and Krista Lagus. 2006. Morfessor in the morpho challenge.
- Injy Hamed, Moritz Zhu, Mohamed Elmahdy, Slim Abdennadher, and Ngoc Thang Vu. 2019. Code-switching language modeling with bilingual word embeddings: A case study for Egyptian Arabic–English.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification.
- Dinar Korbanov. n.d. Aylandirow. URL: http://aylandirow.tmf.org.ru
- Liang Lu, Lingpeng Kong, Chris Dyer, Noah A. Smith, and Steve Renals. 2016. Segmental recurrent neural networks for end-to-end speech recognition.
- Manuel Mager, Özlem Çetinoğlu, and Katharina Kann. 2019. Subword-level language identification for intra-word code-switching.
- Dong Nguyen and Leonie Cornips. 2016. Automatic detection of intra-word code-switching.
- Mansur R. Saykhunov, R. R. Khusainov, T. I. Ibragimov, J. Luutonen, I. F. Salimzyanov, G. Y. Shaydullina, and A. M. Khusainova. 2019. Corpus of Written Tatar.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units.
- Pranaydeep Singh and Els Lefever. 2020. Sentiment analysis for Hinglish code-mixed tweets by means of cross-lingual word embeddings.
- Abishek Srivastava, Kalika Bali, and Monojit Choudhury. 2020. Understanding script-mixing: A case study of Hindi–English bilingual Twitter users.
- Zeynep Yirmibeşoğlu and Gülşen Eryiğit. 2018. Detecting code-switching between Turkish–English language pair.