Translating through Transliteration
The Case of Tajik Persian

Rayyan Merchant[1], Kevin Tang[21]
[1]University of Florida, [2]Heinrich Heine Universität
rayyan.merchant@ufl.edu, kevin.tang@hhu.de

Background

- The Persian language is written in two different scripts

فارسی ⟷ форсӣ

Perso-Arabic **Tajik-Cyrillic**

- Mutual intelligibility between standard dialects is high in spoken form, but falls to zero in written form
- Tajikistan, a country of ~10 million, cannot access written media from the greater Persian-speaking world (~100 million people
- Roughly 2.2% of the Internet is written in Persian
  - Less than 0.1% is written in Cyrillic, the rest is in Arabic
- The scripts do not have a simple one-to-one correspondence, obfuscating typical transliteration
- **Can a model be trained to "translate" between the two dialects through transliteration?**

Method

- **Previous Work:**
  - Proposed a statistical model for machine transliteration, but lacked a true parallel corpus with which to fully verify model performance (Davis, 2012)
- **Model:**
  - Neural network-based Grapheme-to-Phoneme (G2P)
- **Why G2P:**
  - G2P models are typically used in Text-to-Speech (TTS) systems, converting graphemes (letters) to phonemes (pronunciations)
  - Typical transliteration models do exist, but G2P may be more suited to this task, as it greatly resembles TTS
  - The Arabic standard does not accurately represent pronunciation, but the Cyrillic standard does
  - We seek to apply such a model (Yolchuyeva et al., 2020) in one direction: Arabic (Grapheme) to Cyrillic (Phoneme)
- **Corpus:** the **very first** aligned digraphic Persian corpus, manually collected from blogs and articles online
  - ~5400 sentences, ~42,000 words

Challenges

- **Script Comparison**
  - The Perso-Arabic script is an abjad
    - Vowels are often unwritten, and sometimes ambiguous
    - The speaker must know how to pronounce the word already
  - The Tajik-Cyrillic script is an alphabet
    - All sounds are (generally) written as they are pronounced
    - The speaker does not require prior knowledge to learn how to pronounce a word
- **Case Sensitivity**
  - The Arabic script does not implement case, while Cyrillic does
  - When converting from Arabic to Cyrillic, case must be inferred
- **Unwritten Grammatical Particle: "Ezâfe"**
  - The "Ezâfe" links two words together, and can be used to denote: possession, adjective-noun relationships, noun linkage, and given name
  - Despite being so common, it is often unwritten in Perso-Arabic text, but always written in Tajik-Cyrillic
  - When transliterating from Arabic to Cyrillic, the location of the"Ezâfe"'must be inferred and inserted where necessary
- **Non-bijective Alignment and Letter Ambiguity**
  - Several syllables and letters have one rendering in Cyrillic, but several in Arabic
  - When transliterating from Cyrillic to Arabic, the correct option must be chosen

Perso-Arabic **Tajik-Cyrillic**

من کتاب را خواندم — ман китобро хондам

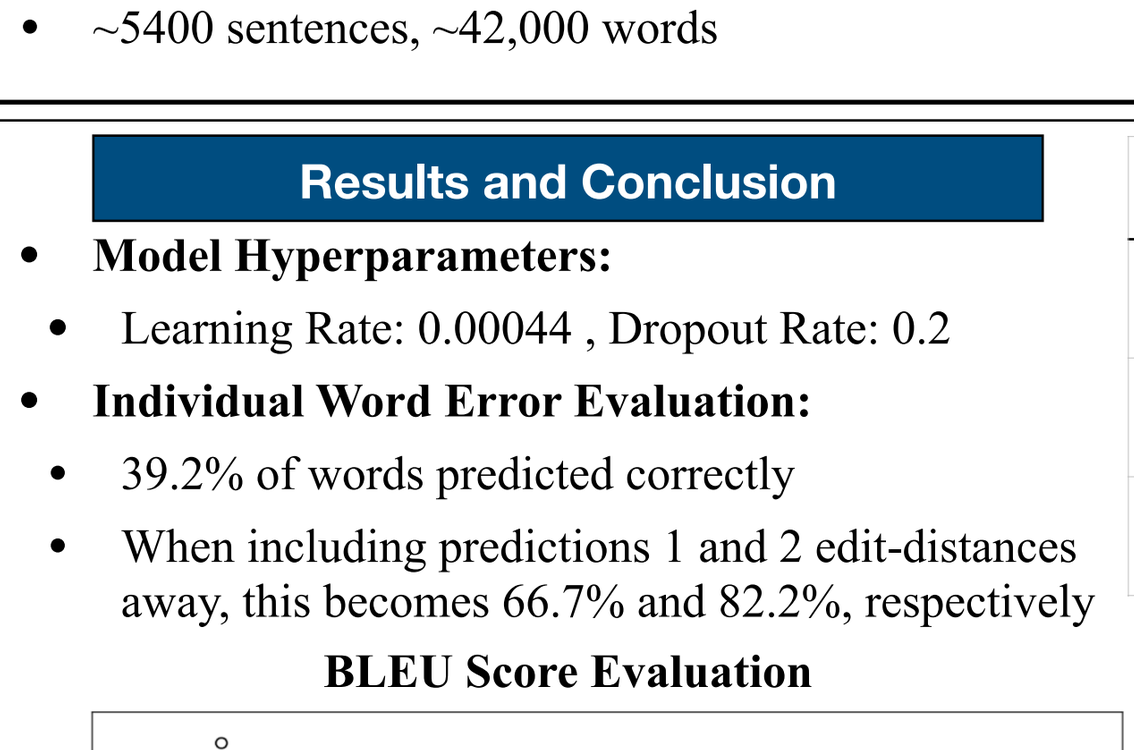
↓ ↓

mn ktob ro xwondm — man kitobro xondam

Results and Conclusion

- **Model Hyperparameters:**
  - Learning Rate: 0.00044 , Dropout Rate: 0.2
- **Individual Word Error Evaluation:**
  - 39.2% of words predicted correctly
  - When including predictions 1 and 2 edit-distances away, this becomes 66.7% and 82.2%, respectively

BLEU Score Evaluation



Script	Example Sentence (errors marked in red)
Arabic	امروز ناظران بین المللی کنفرانس مطبوعاتی در شهر دوشنبه برگزار می‌نمایند
Cyrillic (Expected)	Имрӯз нозирони байналмиллалӣ конфронси матбуотӣ дар шаҳри Душанбе баргузор менамоянд
Cyrillic (Predicted)	имрӯз нозарон беин лмлӣ канфаронс матбуъотӣ дар шаҳр душанбе баргузор маинаминд

- **BLEU Score: A Corpus-Based Metric (Papineni et al., 2002)**
  - Number from 0 to 1 that measures similarity of machine-translated text to reference translations
  - Phrase length of N determines how many N-grams match their counterpart in the reference translation
- **Analysis**
  - Some vowels successfully predicted, others unsuccessfully
  - Vowel insertion partially successful
  - Model proves unable to detect ezâfe
- **Conclusion**
  - G2P approach presents a viable approach to transliterating Persian from Arabic to Cyrillic
  - Further improvements required before our model becomes usable
- **Future Work**
  - Increase model attention to account for case sensitivity
  - Supplement corpus with manually-added "ezâfe" tags
  - Continue hyperparameter testing

References:

Chris Irwin Davis. 2012. Tajik-Farsi Persian Transliteration Using Statistical Machine Translation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3988–3995, Istanbul, Turkey. European Language Resources Association (ELRA).

Yolchuyeva, S., Németh, G., & Gyires-Tóth, B. 2020. Transformer based grapheme-to-phoneme conversion.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. 2002. BLEU: A Method For Automatic Evaluation Of Machine Translation. In Proceedings of the 40th annual meeting on Association for Computational Linguistics.

Acknowledgements: First and foremost, I would like to thank Dr. Tang for his guidance, along with Akhilesh Ramarao, Chris Geissler, and all other members of our laboratory for their advice and feedback in conducting and presenting this research project. I would also like to express my gratitude to the Iranian and Tajikistani writers whose transliterated writings allowed for this corpus' (and subsequently this model's) creation.