Transliteration for Cross-Lingual Morphological Inflection

Nikitha Murikinati
nmurikin@andrew.cmu.edu

Antonis Anastasopoulos
aanastas@andrew.cmu.edu

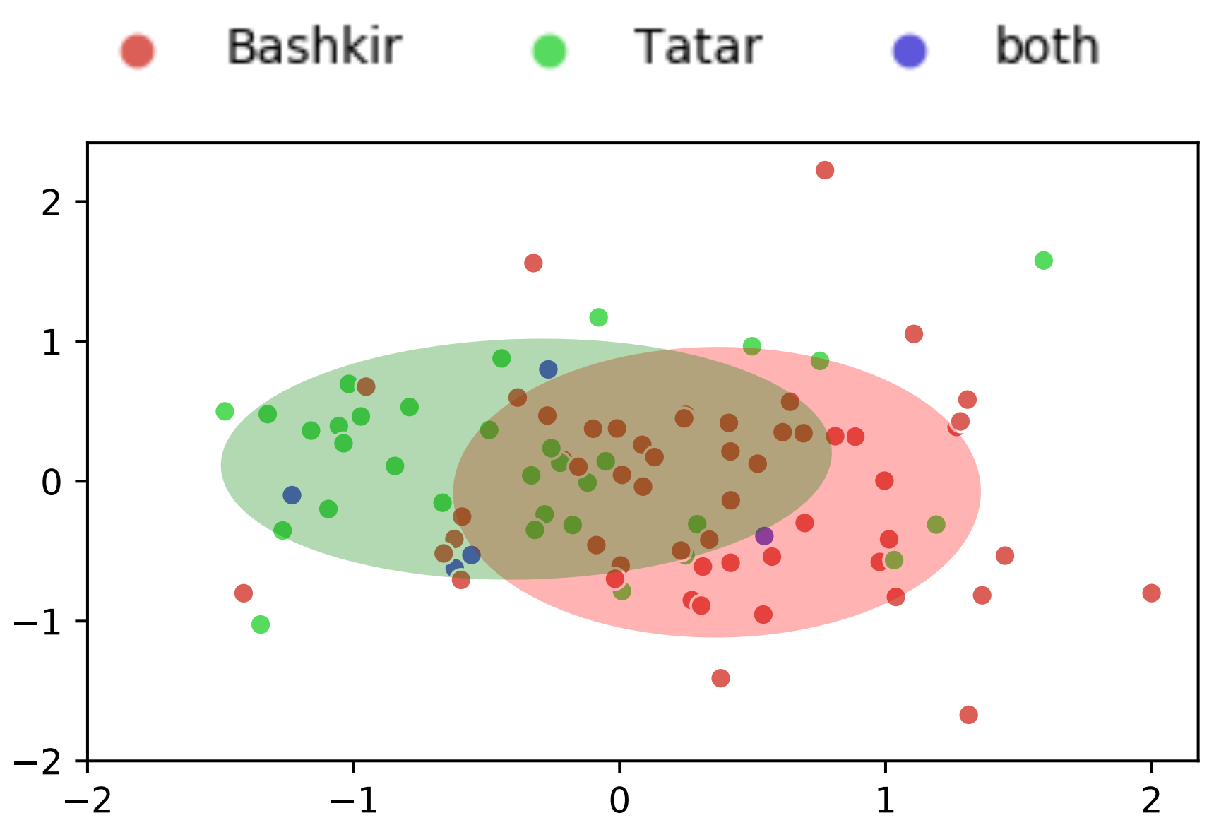
Graham Neubig
gneubig@cs.cmu.edu

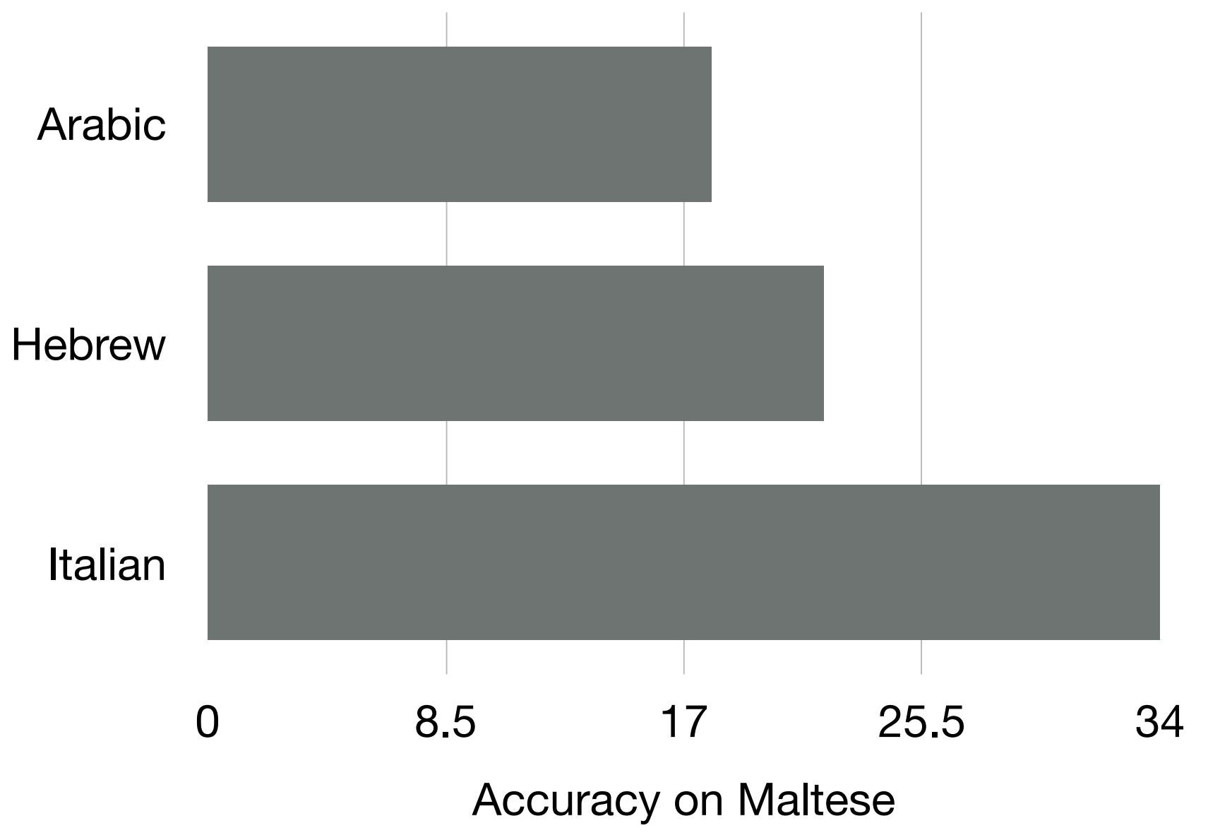
Overview

- **Morphological Inflection** is the task where, given a lemma and a set of morphological tags, one has to generate the correctly inflected form, e.g.,

`aguar + V;PRS;2;PL;IND;` ⟶ `aguà`

- Cross-lingual transfer between typologically related languages has been successful for morphological inflection.
- But if the languages do not share the same script, current methods yield more modest improvements.





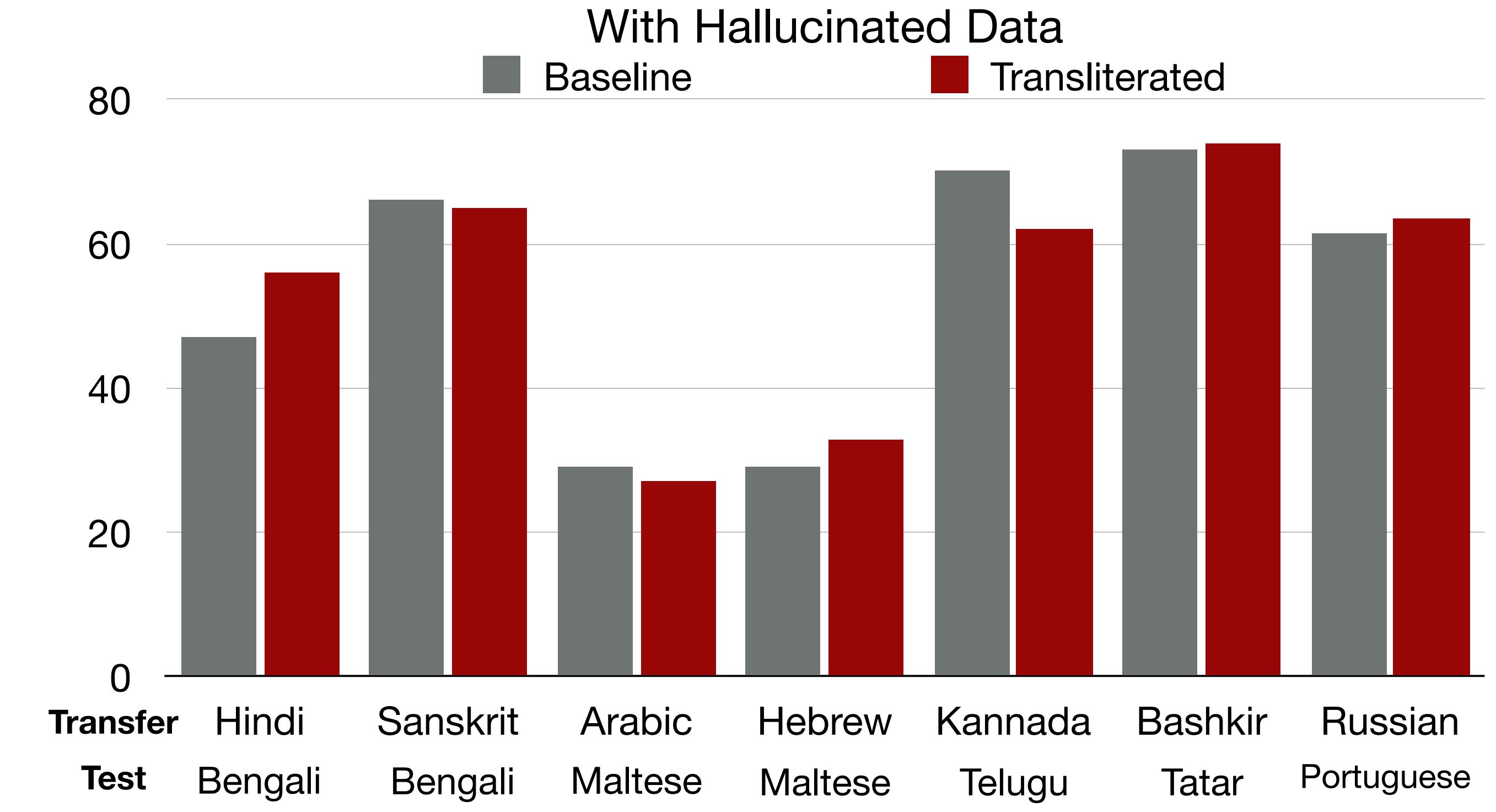
Exemplifying the Problem

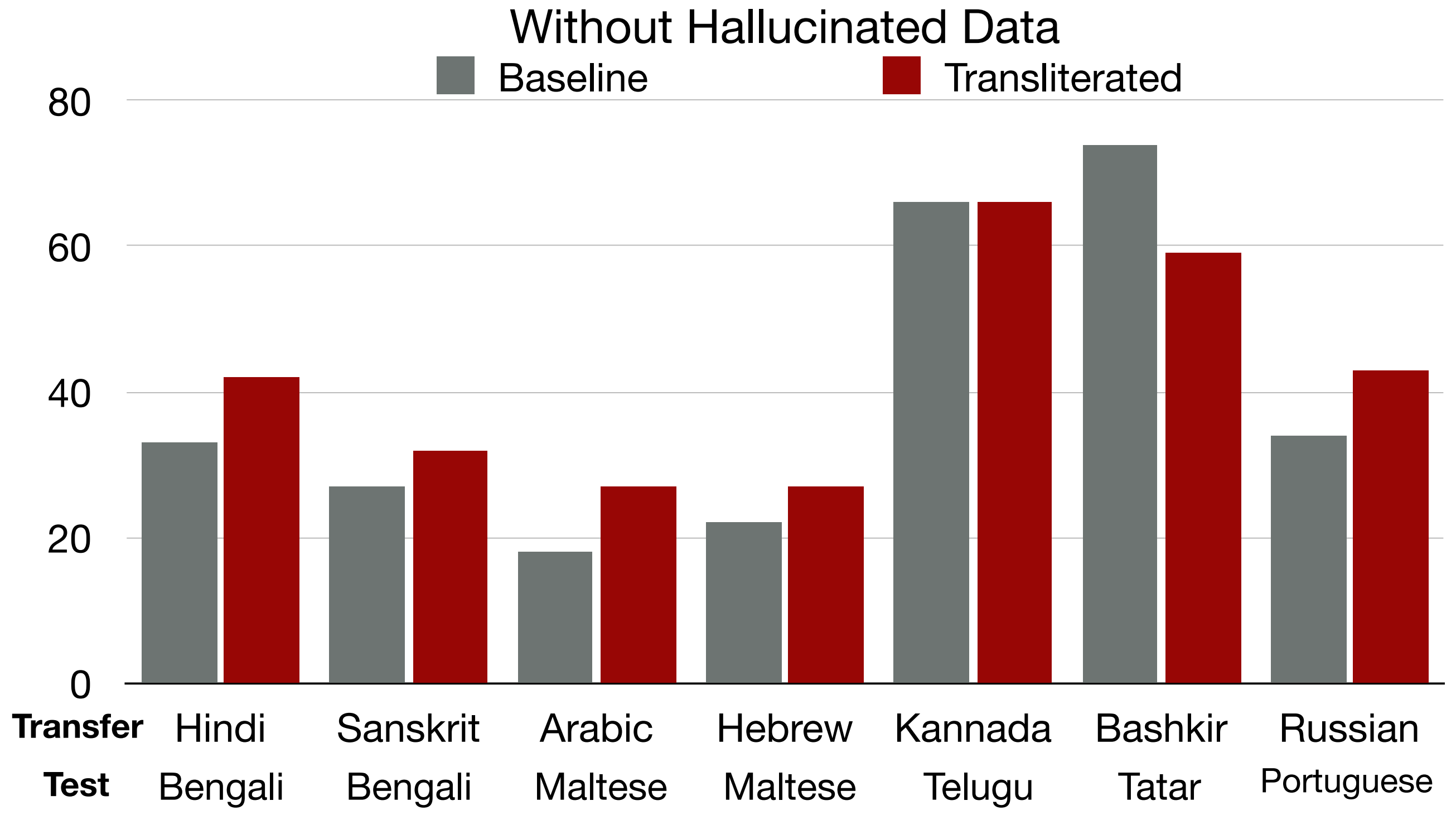
- A visualization of the character embeddings learned after cross-lingual training for Bashkir and Tatar, which have different scripts, shows that the two languages are to an extent separable. Thus, a shared representation is needed to properly cross-lingually train.
- When training a system on Maltese data, Maltese is typologically closer to Arabic and Hebrew than Italian. However, accuracy is higher when transferring from the same-script language of Italian.

Methodology

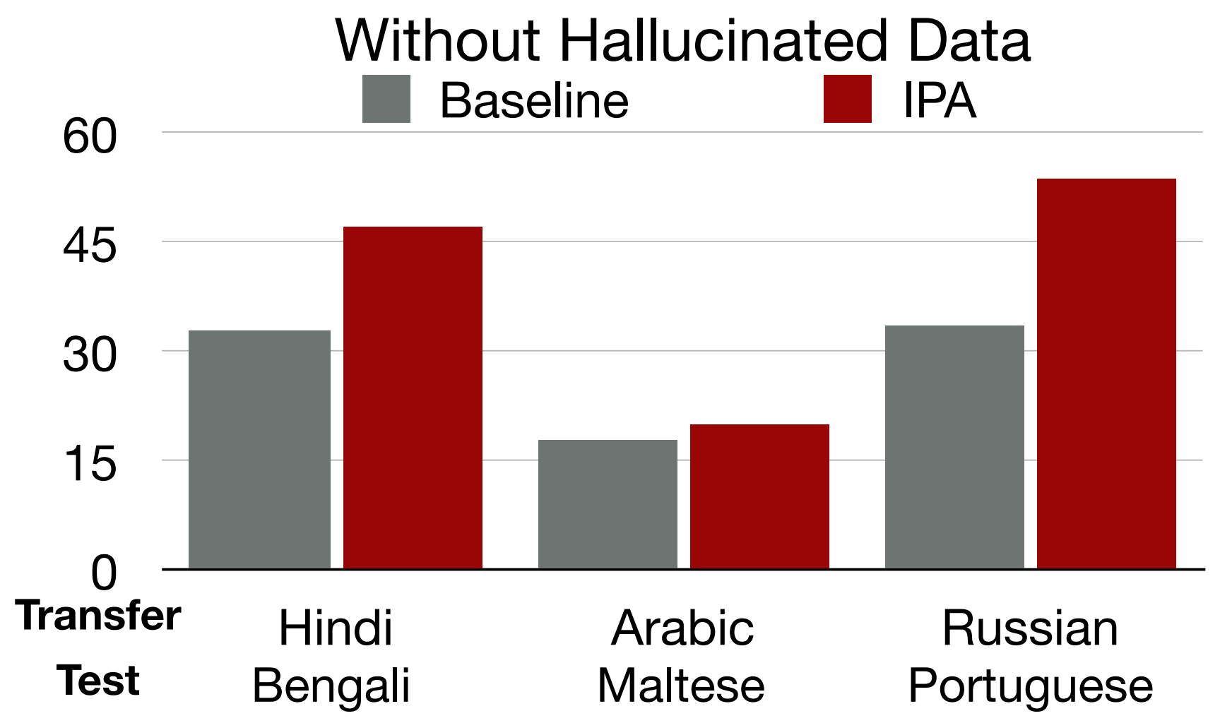
- We first transliterate the transfer language data into the script of the test language, and then use the data to train the inflection model. Additional experiments were run with converting to a phonemic transcription using the International Phonetic Alphabet (IPA) and as well as romanization. In all experiments, the model was run with and without augmented hallucinated data.
- As our baseline, we used the exact same data, model, and process, only removing the transliteration pre-processing step.
- The data was drawn from the SIGMORPHON 2019 Shared Task on Morphological Inflection[1], and for transliteration various libraries such as IndicNLP[2], URoman[3], and Epitran[4] were utilized. The morphological inflection model used was from Anastasopoulos and Neubig[5].

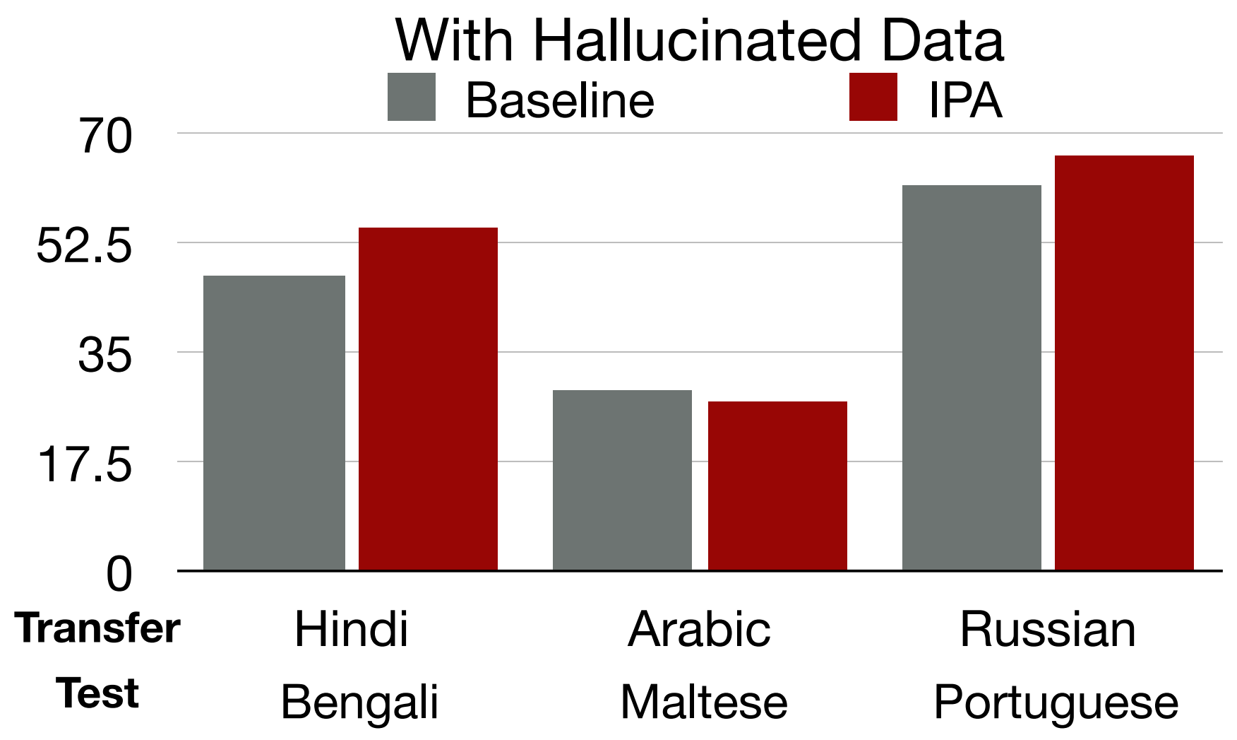
Results



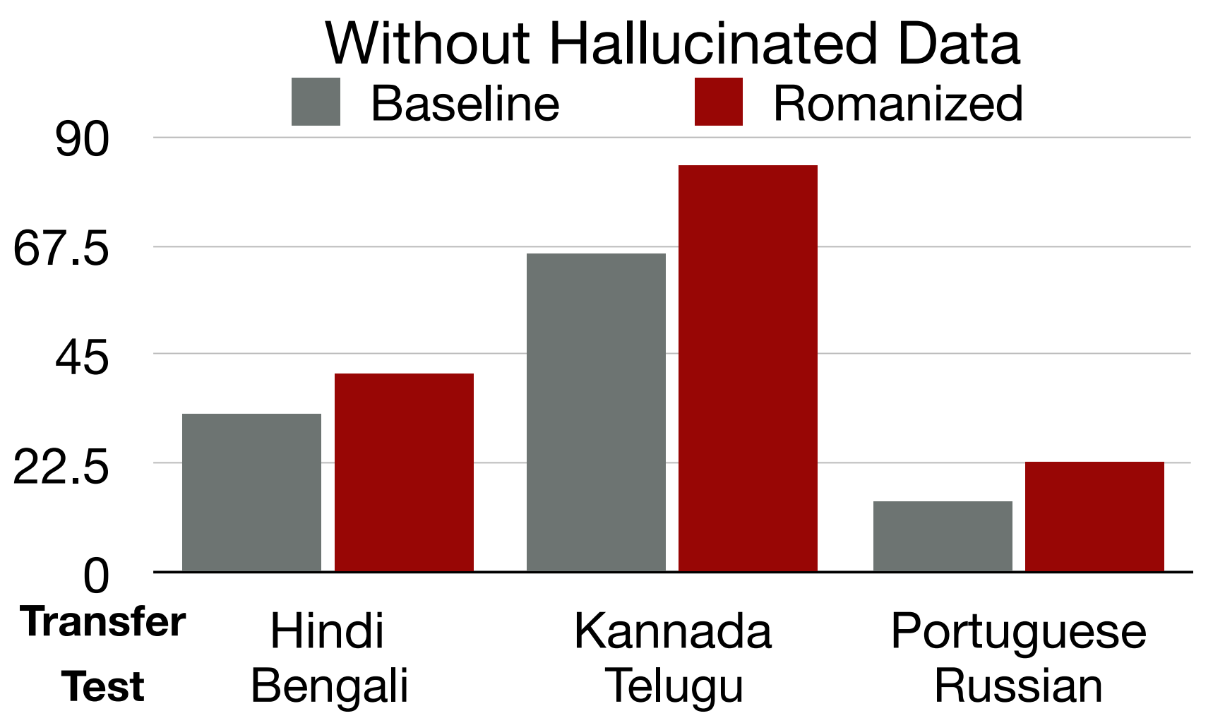


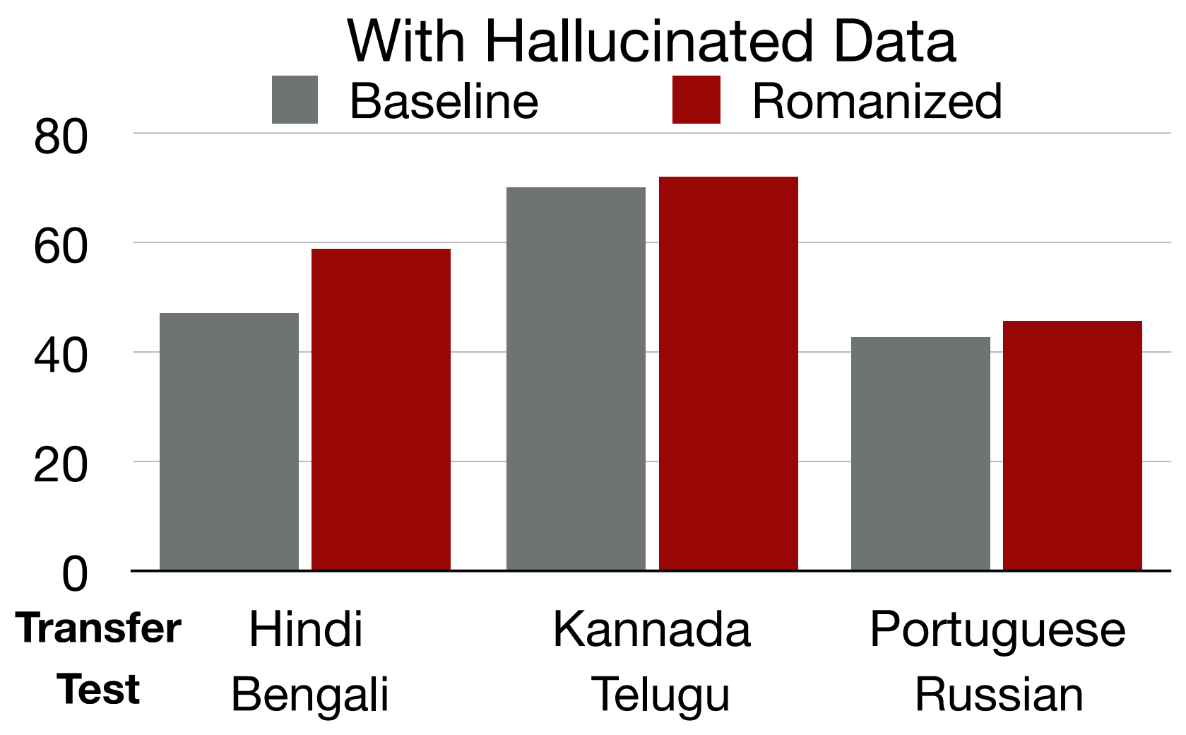
(above) Transliteration of the transfer language into the script of the test language improves accuracy in some cases, with and without hallucinated data. In some language pairs it can be harmful. We report the exact match accuracy on the test set.





- (left) Phonemic transcription with IPA improves accuracy in most cases, with and without hallucinated data. We report the exact match accuracy on the test set.
- (right) Romanization of the transfer language (if it is not in the roman script already) and the test language improved accuracy in all cases, with and without hallucinated data. We report exact match accuracy on the test set.





Conclusions

- In most cases, transliteration results in accuracy improvements, some being statistically significant. Of the three language pairs run for the IPA and roman transliteration, the improvements were similar to those of the transliteration of the transfer language into the test language's script, though some cases did mark a larger improvement.
- We noted that the improvements are orthogonal to those obtained by data augmentation through hallucination, even in typologically distant languages.
- However, the experiments were restricted by the lack of reliable transliteration tools for most scripts. Additionally, some of the models do not account for phenomena such as vowelization for Abjad scripts like Arabic.

Citations

1. Arya D. McCarthy, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garett Nicolai, Miikka Silfverberg, Sebastian J. Mielke, Jeffrey Heinz, Ryan Cotterell, and Mans Hulden. 2019. The SIGMORPHON 2109 shared task: Morphological analysis in context and cross-lingual transfer for inflection. In *Proceedings of the 16th Workshop on Computational research in Phonetics, Phonology, and Morphology*, pages 229-244, Florence, Italy.
2. Anoop Kunchukuttan. 2020. The indicnlp library.
3. Ulf Hermjakob, Jonathan May, and Kevin Knight. 2018. Out-of-the-box universal romanization tool uroman. In *Proceedings of the 56th Annual Meeting of Association for Computational Linguistics*, Demo Track. ACL-2018 Best Demo Paper Award.
4. David R. Mortensen, Siddharth Dalmia, and Patrick Littell. 2018. Epitran: Precision G2P for many languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).
5. Antonios Anastasopoulos and Graham Neubig. 2019. Pushing the limits of low-resource morphological inflection. In *Proc. EMNLP*, Hong Kong