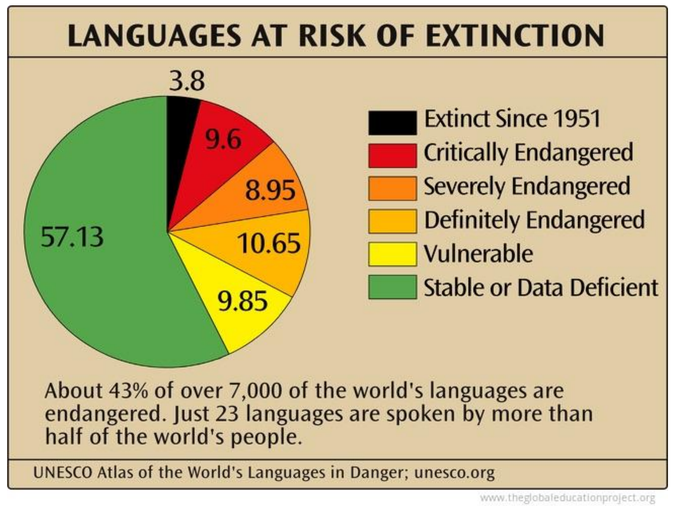
Revitalizing Endangered Languages Using Natural Language Processing

Vrushti Patel, Keith Kreschollek, Vincent Chen
Department of Computer Engineering

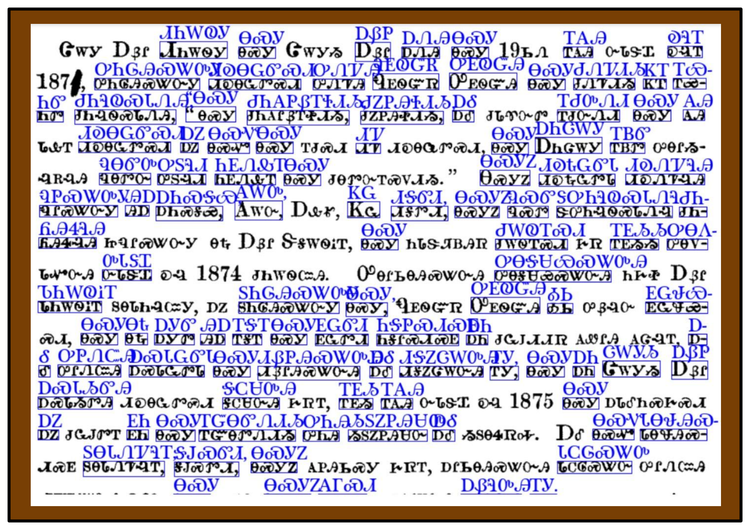
Introduction



- Preserving Culture:
  - There are currently about 7000 languages in the world, 50-90% of which will be dead or endangered by the end of this century
  - Languages are an integral part of our identities and culture
  - Languages also hold a significant amount of knowledge
- Community Priorities:
  - The indigenous communities are highly marginalized community
  - They must to be heavily involved during the process to ensure that the work being done is what they need and want
- Under Representation in NLP:
  - 88% of the languages spoken in the world are underrepresented in NLP
  - Less NLP technology support means less users exposed to the languages
  - Fewer speakers creating language content = scarcity of resources = hindrance in the development of NLP technology

Our Methods

- Neural Machine Translation
  - Train Cherokee-to-English NMT model using OpenNMT-py
- Optical Character Recognition
  - Enhance the pre-existing dataset by retrieving parallel data from sources such as children's books using OCR



Using OCR to extract Cherokee text from PDF

Findings and Discussion

- Machine Translation
  - Cherokee-to-English MT model trained on a Bible dataset yielded a BLEU score of 35.74
  - The BLEU score dropped to 13.695 when the model was trained on a larger dataset since the dataset included both old and new English
  - It is extremely difficult to find more data sources given that Cherokee is a low resource language
- Optical Character Recognition
  - Quality of OCR is highly dependent on the quality of data
  - The available documents are often too noisy

Future Work

- Optimizing the translation model
  - Back translation: produce synthetic data using back translation in order to enlarge the dataset
  - Transfer learning: train the model on a higher resourced but similar language first, such as Inuktitut, and then fine tune the model to Cherokee
- Optical Character Recognition
  - GAN model for image denoising
- Community collaboration
  - It is crucial that we work with the indigenous scholars and community members in order to meet community needs
- Children's books using image recognition and generation

References

- Shiyue Zhang, Ben Frey, and Mohit Bansal. 2022. How can NLP Help Revitalize Endangered Languages? A Case Study and Roadmap for the Cherokee Language.

Acknowledgements

The authors acknowledge the generous support by Lehigh University Department of Computer Engineering and the P.C. Rossin College of Engineering & Applied Science through the Clare Boothe Luce Research Award. The authors also thank the David and Lorraine Freed Undergraduate Research Symposium.

DAVID AND LORRAINE FREED UNDERGRADUATE RESEARCH SYMPOSIUM

LEHIGH UNIVERSITY

Bina Computer Vision and Remote Sensing Lab