Language Technologies Institute

Thesis Proposal

Friday, August 6, 2021| 2:00 pm

Digitizing Endangered Language Texts

Shruti Rijhwani

COMMITTEE:

Graham Neubig
(Chair)

Alan Black

Taylor Berg-Kirkpatrick

Antonios Anastasopoulos,
(George Mason University)

Daisy Rosenblum,
(University of British Columbia)

Abstract

Much of the textual data existent in many languages of the world is locked away in non-digitized books and documents. This is particularly true in the case of most endangered languages, where little to no machine-readable text is available, but printed documents such as cultural texts, educational books, and notes from linguistic documentation frequently exist. This thesis addresses the task of digitizing printed materials that contain text in endangered languages. Automatic digitization of these materials is useful for a multitude of reasons. It can aid language preservation and accessibility efforts by archiving the texts and making them searchable for language learners and speakers, as well as enable the development of natural language processing systems for endangered languages.

Typically, optical character recognition (OCR) is used to digitize printed documents. However, state-of-the-art OCR systems are generally trained on large amounts of textual data and transcribed images, which are unavailable for most endangered languages. To overcome this challenge, we propose a suite of OCR post-correction methods that are designed to facilitate learning from small amounts of data and improve the results of existing methods on under-resourced languages.

We first present a benchmark dataset for the task of digitizing endangered language materials, containing transcriptions of printed documents in four critically endangered languages and extensively analyze the shortcomings of existing digitization methods on this dataset, finding that there is considerable room for improvement in the performance of OCR on endangered languages. Then, we present several methods for fixing recognition errors in OCR outputs, targeted to learning models in a data-scarce setting: (1) a neural OCR post-correction method that leverages high-resource translations and structural biases in the model to improve performance, (2) a semi-supervised technique that efficiently uses unlabeled data for post-correction by combining self-training with automatically derived lexica and morphological information, and (3) an unsupervised learning approach proposed in order to improve digitization across all endangered languages, including those for which even a minimal amount of labeled data is unavailable.

https://drive.google.com/file/d/14fTRTflTq_xpZ1fgnW9hQ-dtVkwSplwa/view