Dissertation Defense

Improving OCR Post Processing with Machine Learning Tools

by Jorge Ramón Fonseca Cacho
Tuesday, July 2, 2019
Location: UNLV TBE A-220
Start Time: 12:00 PM Pacific Standard Time

Dr. Kazem Taghva
Advisory Committee Chairperson
Professor
Department of Computer Science
University of Nevada, Las Vegas

Optical Character Recognition (OCR) Post Processing involves data cleaning steps for documents that were digitized, such as a book or a newspaper article. A step in this process is the identification and correction of spelling and grammar errors generated due to the flaws in the OCR system. This work is a report on our efforts to enhance the post processing for large repositories of documents.

The main contributions of this work are:

- Development of tools and methodologies to build both OCR and ground truth text correspondence for training and testing of proposed techniques in our experiments. In particular, we will explain the alignment problem and tackle it with our de novo algorithm that has shown a high success rate.

- Exploration of the Google Web 1T corpus to correct errors using context. We show that over half of the errors in the OCR text can be detected and corrected.

- Applications of machine learning tools to generalize the past ad hoc approaches to OCR error corrections. As an example, we investigate the use of logistic regression to select the correct replacement for misspellings in the OCR text.

- Use of container technology to address the state of the reproducible research in OCR and Computer Science as a whole. Many of the past experiments in the field of OCR are not considered reproducible research questioning whether the original results were outliers or finessed.