EVALUATION OF TEXT EXTRACTION METHODS ON MULTI-LINGUAL PDFS

A PREPRINT

Erica Sim
Department of Linguistics
University of Washington
Seattle, WA 98105
ejsim@uw.edu

May 10, 2019

ABSTRACT

Abstract goes here.

1 Introduction

2 Related work

Benchmark & eval paper— what else?
see mendeley

Extracting text from a PDF is a difficult task that has been tackled many times before, with varying results.

2.1 About PDFs and text extraction

2.2 Difficulties in text extraction

–why is it difficult?
–previous attempts with different methods
–their problems?

2.3 Difficulties in tool evaluation

see benchmark paper –difficulties inherent in evaluating tools
–their method of groundtruth gathering (are there others?)

3 Motivation

3.1 Working with multilingual documents

–should this be in the main section part?

3.2 Importance of accuracy

Should this be included in the subsection intro? Or in the lower subsections? Or its own thing here?

3.3 Difficulties in text extraction from multi-lingual documents

Ibid

3.3.1 Working with non-Latin text

3.3.2 Working with IPA characters

Should I also have another subsubsection on use in Linguistics papers in general?

3.4 Difficulties in evaluation of accuracy

If they end up being different than with monolingual docs

3.5 Importance for linguists

4 Methodology

In this section I describe the methodology used.

4.1 Data sets used

Bast & Korzen, Language Science books repository
CITE DATA
–should there be a subsection on data problems here? i'd think that many of them are inherent to the problem and the formatting of the data (LaTeX)

4.1.1 Problems with data sets

4.2 Methods of text extraction

Go over the ones that Bast & Korzen use (at least the ones I'm looking at) and also add some more?
–should I have an overview of what/why I'm doing this? or is that better in introduction? (we'll see)

4.3 Replicating previous results

4.4 Considering multilingual documents

4.5 My stuff here

Possible things to add: difficulty with my data (should it be its own subsection? or a sub-subsection?) and whatever the solution ends up actually being

4.5.1 Creating the document set used

4.5.2 Running the benchmark

5 Results

Analysis of results. Should it be called Analysis? Should analysis be its own section? or a subsection?

The conclusion of the paper.

(Bibliography file from Mendeley here)