Information Processing and Management 54 (2018) 874–887

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Statistical learning for OCR error correction

Jie Mei[*,a], Aminul Islam[b], Abidalrahman Moh'd[a], Yajing Wu[a], Evangelos Milios[a]

[a] Faculty of Computer Science, Dalhousie University, Halifax, NS B3H 1W5, Canada
[b] School of Computing and Informatics, University of Louisiana at Lafayette, Lafayette, LA 70503, Canada

ARTICLE INFO

Keywords:
OCR post-processing
OCR error
Error correction
Statistical learning

ABSTRACT

Modern OCR engines incorporate some form of error correction, typically based on dictionaries. However, there are still residual errors that decrease performance of natural language processing algorithms applied to OCR text. In this paper, we present a statistical learning model for post-processing OCR errors, either in a fully automatic manner or followed by minimal user interaction to further reduce error rate. Our model employs web-scale corpora and integrates a rich set of linguistic features. Through an interdependent learning pipeline, our model produces and continuously refines the error detection and suggestion of candidate corrections. Evaluated on a historical biology book with complex error patterns, our model outperforms various baseline methods in the automatic mode and shows an even greater advantage when involving minimal user interaction. Quantitative analysis of each computational step further suggests that our proposed model is well-suited for handling volatile and complex OCR error patterns, which are beyond the capabilities of error correction incorporated in OCR engines.

1. Introduction

Ongoing effort on large-scale digitization has made a massive amount of printed information available to search, access, and analyze. Besides some well-known Internet archives, such as Google Books,[1] Biodiversity Heritage Library,[2] and Project Gutenberg,[3] digitization has been actively organized by public institutions covering a wide range of documentation including historical archives, medical reposts, and government document repositories. However, poor Optical Character Recognition (OCR) accuracy is the primary issue that affects the presentation and analytics of the digitized texts (Alex & Burns, 2014; Lam-Adesina & Jones, 2006; Lopresti, 2009; Reynaert, 2014; Thompson et al., 2016). It is especially problematic for documents that require high reliability, such as medical information (Thompson, McNaught, & Ananiadou, 2015) and government records (Pereda & Taghva, 2011).

However, the scanned image files have different qualities, layouts, and font types which introduce errors that may be hard to avoid in a unified OCR workflow that aims for both efficiency and generality. Also, OCR engines that typically apply dictionary-based validation method has limited correction capability (Smith, 2007).

OCR post-processing is the procedure that aims to fix the residual errors in the OCR-generated text. The general practice of post-processing error correction on the OCR-generated text is agnostic to OCR engines and in the absent of the original scanned document image. An OCR post-processing model is thus able to be applied to digitized texts or a digitization pipeline with any OCR engine. OCR post-processing techniques has been broadly researched for manuscripts of different languages (Al Azawi, Ul Hasan, Liwicki, &

* Corresponding author.
E-mail addresses: jmei@cs.dal.ca (J. Mei), aminul@louisiana.edu (A. Islam), amohd@cs.dal.ca (A. Moh'd), yajing@cs.dal.ca (Y. Wu), eem@cs.dal.ca (E. Milios).
[1] https://books.google.com.
[2] https://www.biodiversitylibrary.org.
[3] https://www.gutenberg.org.

https://doi.org/10.1016/j.ipm.2018.06.001
Received 1 November 2017; Received in revised form 21 April 2018; Accepted 1 June 2018

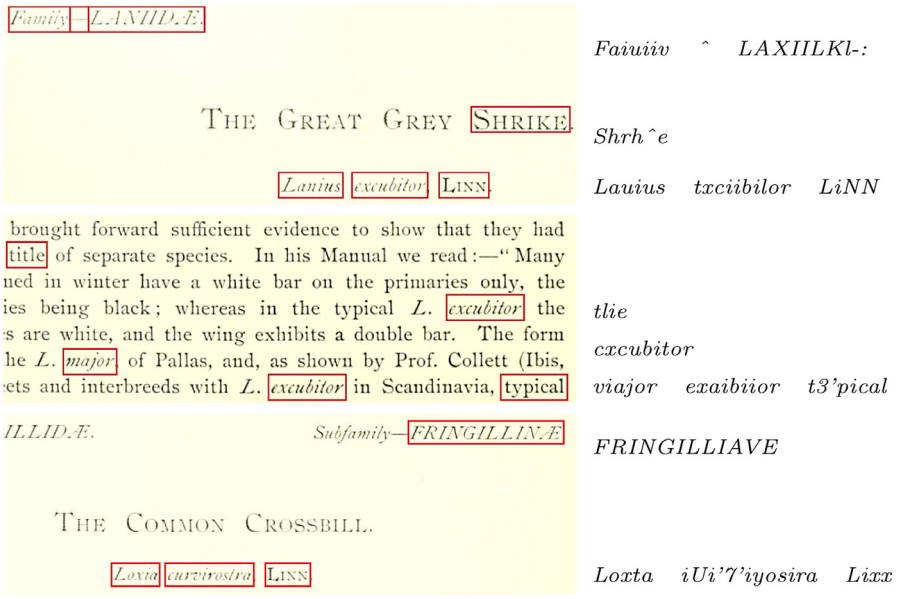

Fig. 1. Sample text image segments and corresponding errors in the OCR-generated text. The original text images corresponding to the OCR-generated errors are highlighted in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Breuel, 2014; Doush & Al-Trad, 2016; Reynaert, 2014).

Given only OCR-generated text as input, OCR error correction is arguably similar to correcting spelling errors. Conversely, OCR errors are inherently more volatile and complex than the misspellings generated by humans. First of all, misrecognized characters in OCR errors are not limited to letters, *e.g.*,⟨*LANIDÆ* → *LAXIILKl* − *:*⟩, which leads to more ambiguous error boundaries and hardly been handled correctly using the standard tokenization rules. Moreover, for handwriting or typing errors, the error patterns are usually being understood by the combination of Damerau–Levenshtein edit operations. However, a large portion of OCR errors violate such one-to-one character modification, *e.g.*,⟨*viajor* → *major*⟩ and ⟨*exaibiior* → *excubitor*⟩ as shown in Fig. 1. Most importantly, the generation of OCR errors results from the combined effects of various factors, including algorithmic defects (*e.g.*,segmentation, classification inaccuracy), limited hardware conditions (*e.g.*,poor scanning equipment), and complex content status (*e.g.*,mixture of text fonts, complicated page layout). It thus leads to volatile error patterns, *e.g.*,⟨*excubitor* → *txciibilor, cxcubitor, exaibiior*⟩, or sophisticated errors that are orthographically far from their corrections, *e.g.*,⟨*curvirostra* → *iUi'7'iyosira*⟩. Hence the causal relation of errors generation is hard to be captured directly using a probabilistic model, especially with limited training instances.

In our OCR post-processing model, we design the entire correction pipeline to handle OCR-specific error patterns. Noisy text tokenization is known to have issues in word boundary detection (Kukich, 1992). Tokenization schemes using space and punctuation inevitably lead to merging (*e.g.*, *ofthe*) and splitting (*e.g.*, *j ust* or *typir,al*) errors. To deal with ambiguous word boundaries in noisy OCR text, we propose a two-stage tokenization scheme, which first utilizes a vocabulary to assert fuzzy boundary detection, and then adopts Peen Treebank conventions to normalize tokens. A comparison experiment shows that our adopted tokenization scheme significantly outperforms other methods on noisy OCR text.

Given the complex nature of OCR errors, we utilize rich linguistic features in lexical, semantics, and context to comprehensively support decision making for error detection and correction. These linguistic features are inferred from web-scale corpora, such as Google Web 1T corpus[4] and English Wikipedia[5] article titles. We use a recall-oriented classification to maximize the error identification probability and adopt ensemble regression to leverage these features for candidate ranking. The quantitative analysis of feature importance shows that *n*-gram context is important for error detection and a combined effort from different feature types is necessary for correction.

To increase the chance of finding corrections for volatile error patterns, we use two methods to explore candidates in Google Web 1T Corpus. We search candidates that are within three reverse Levenshtein distance to the error token from the unigrams of the Google Web 1T corpus. We also collect candidates using similar *n*-gram contexts in bigrams to 5-grams of the Google Web 1T corpus. While these two methods are capable of suggesting a large number of candidates, the computational cost for processing these candidates is overwhelming. We thus use feature-based ranking to prune the candidate set, which considerably reduces its size while maintaining the most appropriate candidates.

[4] https://catalog.ldc.upenn.edu/ldc2006t13.
[5] https://en.wikipedia.org/wiki.