A web platform for collaborative semi-automatic OCR Post-processing

Ana L. Mechaca C.[1], Walter G. Marmanillo[1], Eduardo Xamena[1,2], Juan Ramirez-Orta[3], Ana G. Maguitman[4], and Evangelos E. Milios[3]

[1] Departamento de Informática - Facultad de Ciencias Exactas - UNSa, Salta, Argentina
[2] ICSOH - Instituto de Investigaciones en Ciencias Sociales y Humanidades - CONICET - UNSa, Salta, Argentina eduardoxamena@conicet.gov.ar
[3] Department of Computer Science - Dalhousie University - Halifax, Nova Scotia, Canadá
[4] Departamento de Ciencias e Ingeniería de la Computación - Universidad Nacional del Sur - Bahía Blanca, Argentina

Abstract. Digital Humanities researchers often make use of software that helps them in the task of finding non-trivial relationships among characters in historical text. Usually, the source texts that contain such information come from OCR acquired volumes, carrying high amounts of errors within them. This work explains the development of a web platform for the task of OCR post-processing and ground-truth generation. This platform employs machine learning to predict the correct texts accurately from OCR noisy strings. The method used for this task involves transformers for character-based denoising language models. An active learning workflow is proposed, as the users can feed their corrections to the platform, generating new annotated data for re-training the underlying machine learning correction models.

Keywords: OCR Post-processing · Digital Humanities · Language Models.

1 Introduction

Optical Character Recognition (OCR) is the task of identifying characters and complete texts in images from printed documents. Even though many algorithms and open source software are available for this purpose, the resulting digitized text often exhibits errors. Many approaches have been taken to overcome this issue. Some of them include Machine Learning (ML) procedures to detect and correct errors on the OCR output texts [4]. In particular, the method Context-based Character Correction (CCC), using the pretrained language model BERT achieved the best performance in the OCR Post-correction challenge of the ICDAR Conference [3]. There are also crowdsourcing proposals to generate Ground-Truth (GT) from large OCR-processed corpora [1] in order to compose training data for ML models.

Closely related to the crowdsourcing approach, this work proposes the use of an online web platform for GT production over a popular corpus of Argentinian historical texts, i.e., the volumes of “Güemes Documentado” [5] (GD) were considered as a possible source of noisy OCR text for training language models [2]. The present proposal involves GD as a first use case of a GT production platform for OCR corpora. The corresponding GD texts were processed accordingly to serve as input to the software that handles the GT production process.

[5] http://www.portaldesalta.gov.ar/documentado.html

In order to enhance the efficiency of the GT production task, ML models were designed and trained for suggesting suitable correct texts from the OCR outputs. The objective of the implementation of these models is a reduction of the cognitive effort of human users. State of the art in Natural Language Processing (NLP) tasks is dominated by Transformer architectures due to their transfer learning capabilities and their training in parallel, among other features. By including such ML models in the OCR Post-processing pipeline, users are provided with cleaner text, depending on the performance of the models. Besides, active learning is possible as users generate the final GT for each line and provide that GT for inclusion in the training set, allowing the ML model to improve its accuracy as subsequent text is processed.

2 Web platform for Ground-truth production

The complete design of the developed web platform for OCR post-processing is showed in Fig. 1 and the source code is available online [6]. The current website is part of a larger project that will include Information Extraction and NLP utilities over the same corpora—in this case, GD—and other historical document collections. The GT generation process workflow is performed by users and supported by ML procedures, as shown in Fig. 1. In the first step of the workflow, the GT platform provides the user via the Front end with raw OCR output text strings and initial suggestions for corrections from the corresponding ML prediction models. Next, the user corrects the GT texts and gives the corrected GT texts as feedback to the server and the ML procedures for storing these texts and re-training ML models with the updated information. This active-learning workflow improves both the GT and the ML models.

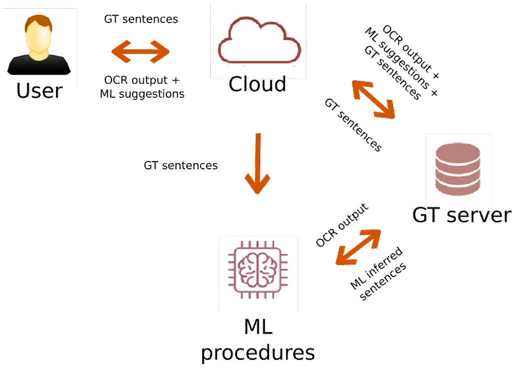


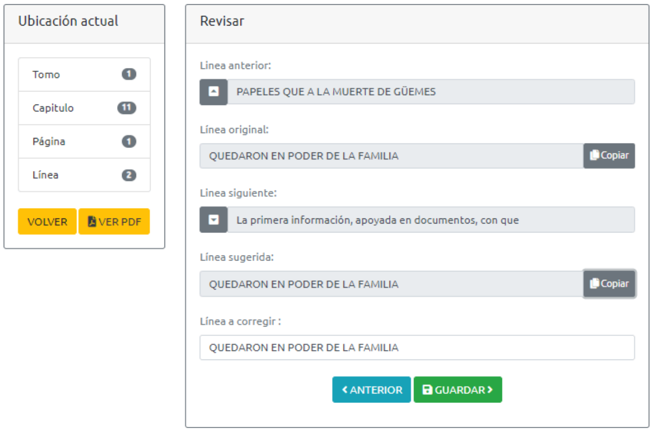
Fig. 1. Visual schema of the data workflow in the web platform.

The front end of the platform consists of a complete website that guides the user through the complete process.[7] The interface after the selection of the current text line is shown in Fig. 2. In the panel named *"Revisar"*, the user can see the three text lines to focus on: the current OCR raw string under-correction (*"Línea original"*), the previous (*"Línea anterior"*), and next (*"Línea siguiente"*) lines to have additional document

[6] https://github.com/GabrielUnsa/gdPage

[7] available at http://nlp.unsa.edu.ar/gtgd/

context. Besides, an additional text field is shown below these three boxes (*"Línea sugerida"*) that has the corresponding ML suggestion according to the associated correction models. The purpose of this text field is to relieve the user's cognitive effort by providing an initial correction suggestion. Such suggestion comes from reliable error-correction ML models. Finally, the fifth text box will contain the GT text line to be stored in the server after clicking the positive button (*"GUARDAR"*). To employ the raw OCR output or the ML suggested text as a base for GT strings, both corresponding text boxes have a button (*"Copiar"*) that copies the text in the box into the actual GT main edition text.



Ubicación actual

Tomo	1
Capitulo	11
Página	1
Línea	2

Revisar

Linea anterior: PAPELES QUE A LA MUERTE DE GÜEMES

Línea original: QUEDARON EN PODER DE LA FAMILIA

Linea siguiente: La primera información, apoyada en documentos, con que

Línea sugerida: QUEDARON EN PODER DE LA FAMILIA

Línea a corregir : QUEDARON EN PODER DE LA FAMILIA

Fig. 2. Main screen of the web platform.

For the navigation between previous and next text lines, the main form contains additional buttons (*ANTERIOR* and *GUARDAR*) that allow the user to go back and forth on the GT production process if necessary, storing the GT text lines. Also a complete guidance on the actual position of the process is provided with the left upper box (*Ubicación actual*), that states the current volume (*Tomo*), chapter (*Capítulo*), page (*Página*) and line (*Línea*) of GD. A pdf visualizer is provided when the button *VER PDF* is clicked to complement the contextual information.

3 Machine Learning models for OCR Post-processing

Several ML models were developed to provide good text suggestions to help users in the correction task. The first approach involves the generation of artificial noise on correct text corpora. The data generation process for this task consists of taking the correct texts as the desired output for ML models and adding noise to the same texts for generating the input texts. This way, an artificial noisy text string represents the text with errors, and the original sequence will be the correct objective string. For the noisy text generation, a threshold is established according to the probability of making changes to every character of each original string. Experimentally, a threshold of 5% (95% probability of keeping the original character as it appeared) was determined as a good value for keeping the semantics required for the strings to be correctable. The possible changes in characters can be adding, removing or turning a character into another one, taking the noise threshold into account. Regarding OCR Post-Correction, the baseline for this task is a model that returns the exact text string provided as input and achieves a Character Error Rate (CER) equivalent to the original noise threshold, in this case about 5%. This baseline, which just returns the input text, is highly competitive for the case of synthetic production of errors in text strings. Encoder-Decoder architectures with LSTM and Attention mechanisms have been evaluated, as well as Transformer models. For the training phase, the method of parameter adjustment employed was back-propagation.The first corpus employed for producing sentences with synthetic noise was the OPUS corpus.[8] This corpus is made up of classical books translated into the Spanish language, with very high-quality texts. As mentioned in [2], high-quality texts are required in this task for achieving adequate language models, avoiding the inclusion of new errors. The best CER value achieved so far was 4.82% for the OPUS corpus.

4 Discussion

The present article describes a software platform for OCR post-processing, combining a web front-end with ML technologies to support users in the text correction process. Several deep learning architectures are currently being developed and evaluated to determine the most suitable mechanism to help users produce GT texts from original OCR volumes. The GT produced with these processes will be of great value in OCR post-processing of similar historical text volumes, as the acquisition algorithms of such texts are similar to the GD, and the language employed shares the same format. The synthetic data approach seems to be a good starting point for initial suggestions, and the corresponding ML models will be improved by re-training on GT while human users produce it.

Acknowledgements: This work was carried out with funds from Universidad Nacional de Salta (Proyectos CIUNSa C 2659 y CIUNSa A 2364), CONICET (Proyecto UE 22920160100056 CO), and Universidad Nacional del Sur (PGI-UNS 24/N051). Besides, this research was enabled in part by Compute Canada (www.computecanada.ca) and the Natural Sciences and Engineering Research Council of Canada. The contribution to this work by Dr. Axel Soto is gratefully appreciated.

References

1. Clematide, S., Furrer, L., and Volk, M. (2018). Crowdsourcing the OCR Ground Truth of a German and French Cultural Heritage Corpus. Journal for Language Technology and Computational Linguistics (JLCL), 33(1), 25-47.
2. Xamena, E., and Maguitman, A. G. (2020). Language modeling tools for massive historical OCR post-processing. In VI Simposio Argentino de Ciencia de Datos y GRANdes DAtos (AGRANDA 2020)-JAIIO 49 (Modalidad virtual).
3. Rigaud, C., Doucet, A., Coustaty, M., and Moreux, J. P. (2019, September). ICDAR 2019 competition on post-OCR text correction. In 2019 International Conference on Document Analysis and Recognition (ICDAR) (pp. 1588-1593). IEEE.
4. Zhang, S., Huang, H., Liu, J., and Li, H. (2020). Spelling error correction with soft-masked BERT. arXiv preprint arXiv:2005.07421.

[8] https://opus.nlpl.eu/