Language modeling tools for massive historical OCR post-processing*

Eduardo Xamena[1,2] and Ana Gabriela Maguitman[3]

[1] Instituto de Investigaciones en Ciencias Sociales y Humanidades (ICSOH) - CONICET - Universidad Nacional de Salta (UNSa)
[2] Departamento de Informática - Fc. Exactas - UNSa, Salta, Argentina
examena@di.unsa.edu.ar
[3] Instituto de Ciencias e Ingeniería de la Computación (ICIC) - CONICET - Universidad Nacional del Sur agm@cs.uns.edu.ar

Abstract. Upon these days, there is a large number of available historical documentary collections that have not been exploited to extract information. Many efforts are being made to digitize these volumes and make them available for digital platforms. However, various obstacles appear in the task of processing their content. Due to the deterioration of documents and other factors such as the different dialects and language variants, the quality of the digitizations is usually low. By means of NLP tools it is possible to increase the quality of texts. The current proposal consists in the employment of NLP tools, particularly neural language models, for processing the output of different OCR mechanisms. Important improvements in the quality of the texts are expected, as this has been the case in many related tasks. The ultimate purpose of this work is the use of the resulting digitized texts in information retrieval (IR) and information extraction (IE) platforms.

Keywords: OCR post-processing · Neural Language Models · Information Retrieval.

1 Introduction

The entire history of a country can be read through documents that tell the daily activities of its inhabitants. There are documentary collections in which these records are kept, in various conditions and several formats. We can find printed books or manuscripts with reports of typical events related to people or companies. The information contained in these volumes can be very useful for different findings with relative weight in the history that we know. That is why the extraction of all that information for analyzing through text mining (TM) and natural language processing (NLP) techniques is a very important task. Before applying any task related to NLP to printed books or manuscripts

* Supported by Universidad Nacional de Salta (Proyectos CIUNSa C 2659 y A 2364), Universidad Nacional del Sur (PGI 24/N051) and CONICET (PUE 22920160100056CO).

it is necessary to obtain quality digital texts, in terms of lexicon and syntax, often coming from OCR digitization. Although there are tools with a high level of precision in character recognition, this process does not yield sufficiently adequate digital representations. Therefore, a fundamental step is the detection and correction of errors in the process of OCR acquisition [4].

In this paper we focus on a particular corpus of the history of Salta and Argentina, the "Güemes Documentado" (GD), available online.[4] The purpose of the GD collection is to vindicate the image of General Güemes in different aspects of his life, through the contribution of letters, reports and other types of documents that support the history. The project that hosts this task consists of the generation of platforms for searching, recovering and extracting historical information about the period that comprises his life. In a previous work [5], different topic models were elaborated based on the contents of some volumes of the GD. Despite the results obtained, the development of an automated correction mechanism for the texts is necessary in order to improve the performance of topic modeling and other TM and NLP tools.

The next section describes other research proposals related to the OCR post-processing task. Then, our proposal is explained in terms of phases, models and corpora employed. After that explanation, some preliminary results are depicted, and the final section enumerates conclusions.

2 Related work

The OCR post-processing task has been addressed by many authors so far, going from more or less simple Graphical User Interfaces (GUI) for assisting users in the manual correction of single documents [1] to semi- or fully-automatic frameworks for aggregate processing of massive corpora [3]. According to the level of automation of the task, on the one hand there are works showing exclusively manual or semi-supervised processing, providing documents to a community of volunteers for transcription, perhaps enabling some GUI –for example, the "Transcribe Bentham" [1] or the "Transcriptorium" [3] projects–. On the other hand, fully automated processing has been studied in more recent works, using combinations of deep learning Sequence to Sequence (Seq2seq) models with large n-gram corpus to achieve corrections that are more in line with real users' criteria [2].

3 Neural OCR Post-processing platform

The proposal of this work is made up by a combination of different processes on the output of OCR software. The first step is a process called "Initial Correction", that employs neural models of language at the character and word level. Several spelling and word confusion errors are expected to be corrected at the output of this stage. Then, in order to avoid "real word" errors –where a word

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

seems to be correct but it is not–, n-gram models are run on the processed text. This phase is called "Final Correction". The purpose of applying these n-gram models on the text is to detect rare word combinations, which could be non-detected errors or errors produced in the initial correction stage. The flow of data describing these processes can be seen in Fig. 1.

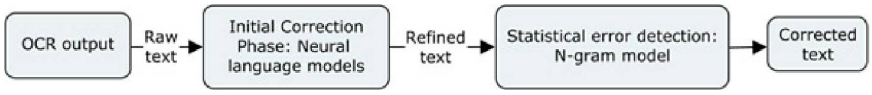

Fig. 1. Text flow over processing steps.

The model used for the initial correction phase should acquire the basic rules of language correction through a training process. A simple way to obtain a synthetic error correction dataset is the generation of data instances with "random noise", upon the base of a corpus of correct sentences. Such noise will consist of the addition of random errors in well-formed text strings. Each data instance in the final dataset will be made up of a pair of character sequences, one with errors with a given probability distribution and the other without errors. As with any machine learning mechanism, it is important to use as much data as possible for the training phase. In order to have a large volume of well-formed texts for building the correction models, known corpora previously employed in other projects will be used as source texts. Such is the case of the Spanish Billion Word Corpus (SBWC), which consists of a compendium of Spanish texts from different origins.[5] In addition, the Spanish Wikipedia will be also used to have texts edited by specialists.[6] A subset of well-formed sentences and another set of corrected sentences from the GD will be used, including original OCR errors. The proposed seq2seq model is composed of an Encoder-Decoder architecture with bi-directional LSTM layers for the Encoder, and an attention mechanism for the decoding phase. This scheme plays the role of a language model, which consists of a mechanism capable of predicting the next word or character, given a text string of t tokens –words or characters–. In other words, given a sequence of tokens $(x_1, x_2, ..., x_{t-1})$, the model must return the most suitable element x_t as a result. In probabilistic terms, the following equation details the formulation of such a model for its implementation through machine learning:

$$f(x_1, x_2, ..., x_{t-1}) = argmax(P(x_t|x_1, x_2, ..., x_{t-1}))$$

The final correction phase requires a thorough mapping of n-grams with associated frequencies for arbitrary n values. Hence, the instances of each word sequence in the original corpora should be counted to determine the probability of finding this sequence in a well-formed text. Besides, each of the sequences extracted from the reference corpus will be considered as a well-formed text.

[5] https://crscardellino.github.io/SBWCE/
[6] https://es.wikipedia.org/wiki/Wikipedia:Portada

4 Preliminary results

The first task of this project is a preliminary test of deep learning architectures for the initial correction stage, involving Encoder-Decoder structures with Attention mechanisms and bidirectional LSTM layers, as explained in the corresponding section. An early set of experiments has been carried on, employing a small fraction of SBWC. As a starting point, the model was trained and tested over the extracted samples of SBWC. Loss values of approximately 0.1 were obtained, constituting a very encouraging first step. The training phase of this work will include the original texts pointed as sources for building SBWC instead of its raw version, due to the excessive preprocessing of the texts.

5 Conclusions

The first steps on the construction of an OCR post-processing platform for historical Spanish documents have been described. The corresponding process consists of two phases, namely an initial correction involving word and character seq2seq language models, and a final stage that refines the work by means of statistical n-gram look-up over the analyzed sentences. As a partial result of this project, an Attentional Encoder-Decoder was implemented and trained for the first phase, with interesting and encouraging results. This first stage of the project only employed a character-based model, and the usefulness of SBWC was analyzed in terms of the present task. The immediate future work will consist of thorough tests on the seq2seq models with different corpora for the generation of datasets, including the Wikipedia and GD texts.

References

1. Causer, T., Tonra, J., Wallace, V.: Transcription maximized; expense minimized? crowdsourcing and editing the collected works of jeremy bentham. Literary and linguistic computing **27**(2), 119–137 (2012)
2. Dong, R., Smith, D.A.: Multi-input attention for unsupervised ocr correction. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2363–2372 (2018)
3. Gatos, B., Louloudis, G., Causer, T., Grint, K., Romero, V., Sánchez, J.A., Toselli, A.H., Vidal, E.: Ground-truth production in the transcriptorium project. In: 2014 11th IAPR International Workshop on Document Analysis Systems. pp. 237–241. IEEE (2014)
4. Mei, J., Islam, A., Moh'd, A., Wu, Y., Milios, E.: Statistical learning for ocr error correction. Information Processing & Management **54**(6), 874–887 (2018)
5. Xamena, E., Marmanillo, W.G., Mechaca, A.L.: Rebuilding the story of a hero: Information extraction in ancient argentinian texts. In: V Simposio Argentino de Ciencia de Datos y GRANdes DAtos (AGRANDA 2019)-JAIIO 48 (Salta) (2019)