OCR Correction for Corpus-assisted Discourse Studies: a Case Study of Old Newspapers

Dario Del Fante[1], Giorgio Maria Di Nunzio[2]
[1] Università di Padova – dario.delfante [at] phd.unipd.it
[2] Università di Padova – giorgiomaria.dinunzio [at] unipd.it

ABSTRACT

The use of OCR software to convert printed characters to digital text is a fundamental tool within diachronic approaches to Corpus-assisted discourse Studies. However, OCR software is not totally accurate, and the resulting error rate may compromise the qualitative analysis of the studies. This paper proposes a mixed qualitative-quantitative approach to OCR error detection and correction in order to develop a methodology for compiling historical corpora. We present a case study on newspapers of the beginning of the 20th century for the linguistic analysis of the metaphors representing immigrants.

KEYWORDS

corpus-assisted discourse studies, OCR detection, OCR correction

TALK

1 INTRODUCTION

In Corpus-assisted Discourse Studies [9], the processes of corpus design and corpus compilation have a marked impact on the entire research and, depending on it, the results may shift dramatically. Especially for diachronic studies, there is a scarcity of digitized version of paper documents, and it is often necessary to manually transcribe the texts under analysis or to use Optical Character Recognition (OCR) software which have a fundamental role in the study of digitizes manuscripts [6]. However, OCR errors may significantly affect the compilation of a corpus in CADS [2]. There are procedures which are adopted to correct OCR errors [1] may not work properly in those cases where the quality of the scan is poor. In this paper, we propose a replicable semi-automatic method for detection and correction of OCR errors. The outcome of this project consists of a set of rules which are, eventually, valid for a different context and applicable to different corpora and which can be reproduced and reused.

The proposed procedure, in terms of computational readability, is aimed at making more readable and searchable the vast array of historical text corpora which are, at the moment, only partially usable given the high error rate introduced by an OCR software.

2 SEARCHING FOR METAPHORS TO REPRESENT IMMIGRANTS IN 1900

Our case study is the analysis of the metaphors used in the newspapers to represent migration to/from the United States of America and Italy from a diachronic perspective since the beginning of the XX century. Given the limited space available in this paper, we will refer to only one particular moment in history which have a significant value in relation to migratory movement: from 1900 to 1914, just before World War I, because this time period, as in [4], is a particularly significant moment for migratory movements, specifically from Europe to the U.S. The availability of data, the newspaper political leaning, and the registration fees were additional constraints that narrowed the selection of the newspapers to the New York Herald[1] for the U.S., and La Stampa[2] for Italy.

Afterwards, we needed to select the keywords to filter the articles useful for our study. The starting point for English was the set of words identified by [5] named under the acronym RASIM: *refugee**,[3] *asylum seeker**, *immigrant**, and *migrant**. We added a fifth word to this list: *emigrant**. As for Italian, we needed to select a set of comparable search terms between English and Italian. We consulted the diachronic Diacoris Corpus,[4] a 15 million words collection of Italian texts produced between 1861 and 1945. The best candidate translation for migrant, immigrant and emigrant were *migrant**, *immigrat**, *immi- grant**, *emigrant**, *emigrat**; for refugee and asylum seeker the candidate Italian terms were *rifu-giat**, *profug**, *clandestin** and *richiedent* asil**.

In the first two rows of Table 1, we show a summary of the statistics for each compiled corpus. The tokens and types of values represent the total number of occurrences versus the number of unique words, respectively. We report both the type/token ratio (TTR) and the standardized type/token ratio (STTR).

3 OCR ERROR DETECTION AND CORRECTION

In this section, we propose a semi-automatic mixed approach to OCR detection, which brings together the dictionary-based and the context-based approaches. A careful analysis of a sample of texts showed that there were a lot of misspellings or non-meaningful words in both corpora caused by the OCR software. The first problem in our case study concerns the fact that we did not have the corresponding ground truth version of the corpora. Therefore, we decided to compile two contemporary newspaper corpora whose texts were digitalized since the beginning: The New York Times for the U.S., La Stampa for Italy (see last two rows of Table 1).

---

1 http://chroniclingamerica.loc.gov

2 http://www.archiviolastampa.it

3 We use the symbol '*' to indicate the possibility of plural, or feminine/masculine for the Italian words.

4 http://corpora.dslo.unibo.it/DiaCORIS/

The error detection correction task consisted of a three-step procedure:

1. Detection of errors by comparing the list of words of the old corpus with the new corpus. The words that do not appear in the latter, or that have a statistically significant difference in frequency compose a list of plausible error candidates.
2. Analysis and categorization of the error in the list of candidates: i) an error containing the same number of characters than the respective correct form.; ii) an error containing a higher/ smaller number of characters than the correct form; iii) a word interpreted by the OCR as two distinct words (i.e., 'department' vs 'depart' and 'ment').

Define the error correction rule as a regular expression.
The implementation of these procedure follows the principles described by [10] where the idea is to mine textual information from large text collections in an efficient and effective by means of pipelines allowing for a sequential process of text analysis. For our experiments, we used the R programming language, which has a set of packages, named 'tidyverse'[5], that implements this idea[6]. A total of 2,313 errors for English and 269 errors for Italian have been individuated and, respectively, as many correcting rules have been written for each language.

Corpus	Years	Documents	Tokens	Types	TTR	STTR
New York Herald	1900-1914	8,540	55,796,968	2,326,897	4.17%	50.24%
La Stampa	1900-1914	3,092	18,773,664	817,865	4.36%	56.63%
New York Times	2000-2014	125	58,915,060	308,251	0.52%	48.39%
La Stampa	2000-2014	62	15,332,063	275,103	1.79%	62.78%

Table 1: statistics

In general, it is not easy to predict in what way OCR correction will work (see Table 2). On one side, the Italian corpus dimensions have been increased in relation to the number of tokens. The increase of tokens might be because many errors were not previously recognized as valid tokens. On the other side, the English corpus dimension has been decreased in relation to the number of tokens. The decrease might be due to the correction of split errors, such as *depart* and *ment*, corrected in *department*.

Corpus	Tokens (B)	Types (B)	Tokens (A)	Types (A)	Δ Tokens	Δ Types
New York Herald 1900-14	55,796,968	2.326.897	55.555.708	2.323.790	-0.43 %	-0.13 %
La Stampa 1900-14	18.773.664	817.865	18.778.210	817.858	0.02%	-0.001%

Table 2: Statistics about errors before (B) and after (A) OCR corrections

5 https://www.tidyverse.org

6 https://github.com/gmdn

4 FINAL REMARKS AND FUTURE WORK

In this paper, we presented a semi-automatic method for detection and correction of OCR errors for the discourse analysis of old newspaper documents. The outcome of this project consists in a set of rules which are, eventually, valid for a different context and applicable to different corpora and which can be reproduced and reused. There are still open questions that we will investigate in this line of work: how many documents have we missed during the compilation of the corpus given that a search keyword may be subject to OCR correction as well? How these types of keyword search error can affect a CADS analysis? For this reason, we intend to use error models to predict the relative risk that queried terms mismatch targeted resources due to OCR errors, as suggested by [3]. We also want to compare our analysis with other approaches that make use of BERT pre-trained neural networks to post-hoc error correction [8], especially in those cases where the context is not clear given multiple OCR errors in the same paragraph, or that take advantage of multiple OCR engines by aligning and comparing their different outputs in order to reduce the error rate [7].

REFERENCES

1. Bassil, Youssef, and Mohammad Alwani. 'Ocr Post-Processing Error Correction Algorithm Using Google Online Spelling Suggestion'. *ArXiv Preprint ArXiv:1204.0191*, 2012.
2. Bazzo, Guilherme Torresan, Gustavo Acauan Lorentz, Danny Suarez Vargas, and Viviane P. Moreira. 'Assessing the Impact of OCR Errors in Information Retrieval'. In *European Conference on Information Retrieval*, 102–109. Springer, 2020.
3. Chiron, Guillaume, Antoine Doucet, Mickaël Coustaty, Muriel Visani, and Jean-Philippe Moreux. 'Impact of OCR Errors on the Use of Digital Libraries: Towards a Better Access to Information'. In *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 1–4. IEEE, 2017.
4. Cohen, Robin. *The Cambridge Survey of World Migration*. Cambridge University Press, 1995.
5. Gabrielatos, Costas. 'Selecting Query Terms to Build a Specialised Corpus from a Restricted-Access Database.' *ICAME Journal* 31 (2007): 5–44.
6. Kettunen, Kimmo, Eetu Mäkelä, Teemu Ruokolainen, Juha Kuokkala, and Laura Löfberg. 'Old Content and Modern Tools-Searching Named Entities in a Finnish OCRed Historical Newspaper Collection 1771-1910'. *ArXiv Preprint ArXiv:1611.02839*, 2016.
7. Lund, William B., and Eric K. Ringger. 'Improving Optical Character Recognition through Efficient Multiple System Alignment'. In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries*, 231–240, 2009.
8. Nguyen, Thi Tuyet Hai, Adam Jatowt, Nhu-Van Nguyen, Mickael Coustaty, and Antoine Doucet. 'Neural Machine Translation with BERT for Post-OCR Error Detection and Correction'. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, 333–336, 2020.
9. Partington, Alan, Alison Duguid, and Charlotte Taylor. *Patterns and Meanings in Discourse: Theory and Practice in Corpus-Assisted Discourse Studies (CADS)*. Vol. 55. John Benjamins Publishing, 2013.
10. Wachsmuth, Henning. *Text Analysis Pipelines: Towards Ad-Hoc Large-Scale Text Mining*. Vol. 9383. Springer, 2015.