OCR ERROR DETECTION AND POST-CORRECTION WITH WORD2VEC AND BERTJE ON DUTCH HISTORICAL DATA

Authors:

- Nynke van ‘t Hof (University of Amsterdam)
- Vera Provatorova (University of Amsterdam)
  Twitter: @vera__pro
- Mirjam Cuper (KB, national library of the Netherlands)
  Twitter: @CuperMirjam
  ORCID: 0000-0003-0187-9873
- Evangelos Kanoulas (University of Amsterdam)
  Twitter: @ekanou
  ORCID: 0000-0002-8312-0694

Category: short paper

Plan to attend: not yet decided between virtual or in person.

Keywords: OCR post-correction, Natural Language Processing, Word Embedding Models, historical data, digital heritage

Relevance and introduction:
With a high quality of OCR-output, documents become more accessible to readers, and NLP tasks can thrive on the data. However, the extent to which all this is possible is dependent on the quality of the OCR-output. OCR on historical data often creates a significant amount of errors due to, among others, the poorer condition of the documents, and variances in font size and spelling[1].
This research focuses on post-processing the OCR-generated machine-readable text. Its focus is on post-correcting OCR-output from historical documents with the use of word embedding models (WEMs).

Background:
Two approaches have recently shown to be promising for OCR post-correction: static word embeddings and the more novel context-aware word embeddings[2]. There are two popular techniques that represent them: static word2vec and context-aware BERT. Static methods (like word2vec) generate one and the same embedding for all different senses in which a word can be used. Thus, homonyms for example get the same representation despite having different meanings. Contextualized methods (like BERT) embed different senses of the same word differently, so these methods are aware of the different meanings one word might have.

---

[1] Salimzadeh, Sara. *Improving OCR Quality by Post-Correction*. PhD thesis, Universiteit van Amsterdam, 2019.
[2] Hammarstrom, Harald, and Shafqat Mumtaz Virk, and Markus Forsberg. Poor man’s ocr post-correction: Unsupervised recognition of variant spelling applied to a multilingual document collection. In *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage*, pages 71–75, 2017.;
Nguyen, Thi Tuyet Hai, and 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*, pages 333–336, 2020.

While BERT appears to do really well on OCR post-correction, current studies do not provide an in-depth analysis of its performance on this task in comparison to other models. This study aims to close this research gap by comparing BERT with word2vec on OCR post-correction of historical documents, and identifying the key reasons behind its impressive performance. As it concerns Dutch-language data, the Dutch BERT (BERTje) is used[3]. To perform the comparative analysis, several pitfalls of word2vec and static word embeddings in general were retrieved from related literature: homonyms, historical spelling variations, out-of-vocabulary words, infrequent words and real-word errors[4].

By comparing the performance of contextualized and static WEMs on post-OCR correction on historical data, we can learn from older and newer WEMs to improve the state-of-the-art, and learn what word embedding methods historical documents specifically could profit from. We chose these context-based WEMs as they learn from the context of the historical documents, so that the text can be understood in its own right without having to rely on statistical rules based on modern language.

In this study, we aimed to answer the following research question:
"How does the performance of contextualized word embeddings from BERTje compare to the performance of static word embeddings from word2vec on the task of post-correcting OCR output from Dutch historical documents?"

Methods and data:

The data concerns the OCR-output of 42.000 historical documents all in Dutch[5]. The documents are from the seventeenth up to the twentieth century. These are of different genres: articles from 6.425 newspapers from the 17th century (the "Meertens" set), 2.000 other news articles, 1.567 book pages and 204 typewritten radio bulletins (the "Impact" set) and 219 books from DBNL. The dataset is freely accessible and in possession of the KB, the national library of the Netherlands, which also commissioned the creation of a ground truth that is 99.95 percent accurate[6]. The OCR has been performed by ABBYY. The data come along with multiple metadata sets[7]. The split for this research was 60/20/20 for the train/validation/test sets.

---

[3] Vries, Wietse de, and Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. BERTje: A Dutch BERT Model. arXiv:1912.09582.
[4] Wevers, Melvin, and Marijn Koolen. Digital begriffsgeschichte: Tracing semantic change using word embeddings. *Historical Methods: A Journal of Quantitative and Interdisciplinary History,* 53(4):226–243, 2020.;
Wiedemann, Gregor, and Steffen Remus, Avi Chawla, and Chris Biemann. Does bert make any sense? interpretable word sense disambiguation with contextualized embeddings. arXiv preprint arXiv:1909.10430, 2019.
[5] Colavizza, Giovanni, and Mirjam Cuper. (2021). Is your ocr good enough? a comprehensive assessment of the impact of ocr quality on downstream tasks [data set]. zenodo. http://doi.org/10.5281/zenodo.4498186.
[6] Ground Truth dataset, National Library of the Netherlands, https://lab.kb.nl/dataset/historical-newspapers-ocr-ground-truth
[7] (Metadata 1): DBNL (2016), DBNL OCR data set. KB lab: The Hague. http://doi.org/10.5281/zenodo.3239290;
(Metadata 2): Wilms, I., Nijssen, R., Koster, T. (2020), historical newspaper ocr ground-truth data set. KB lab: The Hague.; (Metadata 3): Impact project, Impact KB ground-truth. KB lab: The Hague. http://lab.kb.nl/dataset/ground-truth-impact-project

The comparative analysis is performed on two different tasks. The two tasks are detection and correction of erroneous sequences. With a detection step, the proportion of errors fed to the correction model makes it less likely to faultily correct the already correct sequences[8]. Another advantage is that it might be easier to detect at what steps mistakes are made. Furthermore, different approaches can be chosen for each step to optimize the procedures separately.

The error detection task is performed on each word in the OCR-output of the dataset. The context of each word is used by both word2vec and BERTje to predict whether the OCR-output is likely to fit in its context. If it does not, it is considered an error. This step is performed by comparing the OCR-output to a list of candidates that would fit given the context. The same candidates list is used for the correction task. A (gold standard) erroneous word is replaced by the candidate that fits the most in context. The performance on both tasks for both methods is evaluated by comparing the detected errors to actual errors found with the ground truth and by comparing the corrections to the ground truth. Both methods were compared to a baseline dictionary-based method.

Results and state of research:

	Error detection task			Error correction task
	Precision	Recall	F1	Accuracy
Baseline	0.784	**0.555**	**0.650**	0.332
Word2vec	**0.799**	0.465	0.588	**0.593**
BERTje	0.225	0.497	0.310	0.485

Table 1: Summary of the results

The first experiments showed that Word2Vec outperformed BERTje on both tasks (see Table 1). The more extensive preliminary results, comparing the performance on specific pitfalls of word2vec, are not included in this abstract due to the length limit. The assumption that BERTje would solve the pitfalls of word2vec could be neither supported nor debunked by the current results. Computational limitations led to only using part of the data. It was hypothesized that BERTje might need more finetuning to achieve its peak performance on historical data. As this is still an ongoing research project, further experiments will test whether using a more extensively fine-tuned BERT model might solve the pitfalls of word2vec when post-correcting OCR output on Dutch historical data. Various training set sizes could be experimented with. Furthermore, the performance of BERTje might be improved by training on the data from different centuries separately, due to historical spelling variations.

[8] Schaefer, Robin, and Clemens Neudecker. A two-step approach for automatic OCR post-correction. In *Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage*, *Social Sciences, Humanities and Literature*, pages 52–57, Online. International Committee on Computational Linguistics.