OCR quality assessment: Beyond ground truth

KB, National Library of the Netherlands

About KB, National Library of the Netherlands

The KB, national Library of the Netherlands collects everything that is published in the Netherlands and about the Netherlands. The KB collections include books, newspapers, and periodicals, but also websites and e-books. With 450 fte the KB is one of the largest cultural heritage institutions in the Netherlands. The Research department (20 fte) focuses on topics ranging from providing insights into the development of the public library sector to applying AI techniques to make the collections accessible to a broad international audience. The KB is an active player in the field of Artificial Intelligence: it is co-founder of the Cultural AI Lab and the working group Culture&Media of the Netherlands AI Coalition.

Our challenge: OCR quality assessment beyond ground truth

Heritage organizations such as libraries and archives host a wealth of written resources of great cultural relevance to researchers and the public. Thanks to considerable investments in the digitization of historical resources, digital images of an unprecedented number of documents are now available. At the current moment, the KB collection already includes more than one hundred million digitized books, newspapers, and periodicals.

While this is a step in the right direction when it comes to making historical resources more accessible, it falls short where searchability is concerned. Only the availability of digital texts can dramatically improve access to digitized resources. This is why Optical Character Recognition (OCR) and Handwritten Text Recognition (HTR) are so central to the strategy of libraries and archives. Recent advances in machine learning have indeed allowed to automatically extract texts from historical resources at scale, yet a question always remains open: **is the OCR quality good enough?** This question is critical as it has been shown that the quality of OCR impacts the usability of the resulting texts [4,6].

The systematic assessment of OCR ideally requires high-quality ground truth, manually checked by human experts [2]. This is a laborious and costly endeavor, which has so far slowed down digitization efforts that, invariably, include OCR as a key component of their workflow. **We propose a challenge centered on devising, developing, testing and comparing approaches to OCR quality assessment not requiring ground truth.**

This timely topic has only recently started to be explored in the literature, with promising results. Possible approaches include using lexicon-based methods such as dictionary lookup and pre- trained language models [5], combining OCR assessment metrics [1], or applying metrics devised in domains of research, such as reading comprehension [3]. Time is ripe for making a step towards a systematic comparison of such automated OCR quality assessment methods not requiring ground truth. This task will be based on the use of ground truth provided by the KB to measure how well non-ground truth methods compare against established, ground truth ones. The data proposed for the challenge will also allow to compare results over language, source typology and time, and more. This challenge is centered around ICT competences, and will be of interest to computer scientists, linguists, and digital humanists.

Bibliography

1. Cuper, Mirjam. “Examining a Multi Layered Approach for Classification of OCR Quality without Ground Truth.” DHBenelux Journal 4 (2022).
2. Neudecker, Clemens, Konstantin Baierer, Mike Gerber, Clausner Christian, Antonacopoulos Apostolos, and Pletschacher Stefan. “A Survey of OCR Evaluation Tools and Metrics.” In The 6th International Workshop on Historical Document Imaging and Processing, 13–18. Lausanne Switzerland: ACM, 2021. https://doi.org/10.1145/3476887.3476888.
3. Nguyen, Hai Thi Tuyet, Adam Jatowt, Mickaël Coustaty, and Antoine Doucet. “ReadOCR: A Novel Dataset and Readability Assessment of OCRed Texts.” In Document Analysis Systems, edited by Seiichi Uchida, Elisa Barney, and Véronique Eglin, 13237:479–91. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2022. https://doi.org/10.1007/978-3-031-06555-2_32.
4. Strien, Daniel van, Kaspar Beelen, Mariona Ardanuy, Kasra Hosseini, Barbara McGillivray, and Giovanni Colavizza. “Assessing the Impact of OCR Quality on Downstream NLP Tasks.” In Proceedings of the 12th International Conference on Agents and Artificial Intelligence, 484–96. Valletta, Malta: SCITEPRESS - Science and Technology Publications, 2020. https://doi.org/10.5220/0009169004840496.

5. Ströbel, Phillip Benjamin, Simon Clematide, Martin Volk, Raphael Schwitter, Tobias Hodel, and David Schoch. “Evaluation of HTR Models without Ground Truth Material.” arXiv, April 29, 2022. http://arxiv.org/abs/2201.06170.
6. Todorov, Konstantin, and Giovanni Colavizza. “An Assessment of the Impact of OCR Noise on Language Models.” In Proceedings of the 14th International Conference on

Team Leaders

Academic Team Leaders

- Dr. Giovanni Colavizza, University of Amsterdam

KB, national Library of the Netherlands

- Mirjam Cuper, Data Scientist

Explainable Fall Risk Prediction for Elderly People

VeiligheidNL

About VeiligheidNL

VeiligheidNL is the Dutch knowledge center for injury prevention. For more than 35 years, we have been committed to making the lives of millions of people safer by encouraging safe behaviour in a safe environment. VeiligheidNL is a Public Benefit Organisation (ANBI, Algemeen nut beogende instelling). Around 60 employees work for VeiligheidNL. The yearly budget is around 7.5 million euros per year. Most of our funding is provided by the Ministry of Health, Welfare and Sport (VWS). Other sources of funding include the Ministry of Infrastructure and Water Management, ZonMw and local authorities. Core activities are research, professional development, consulting and education.

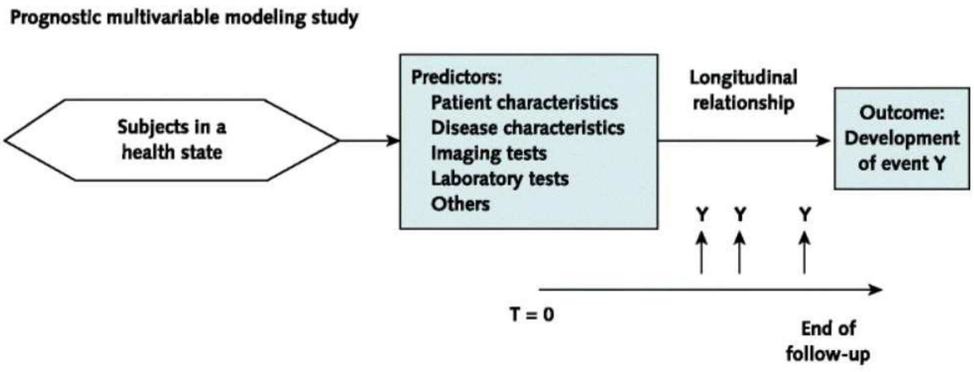
Our challenge:

Elderly fall prevention is one of the six core programs of VeiligheidNL. Every 5 minutes, a person over 65 ends up in a Dutch Emergency Room after a fall accident. And this is only 10% of the total number of falls of the elderly. The consequences of a fall are enormous, both in terms of personal suffering and social care costs. A fall accident has a major impact on the self-reliance of the elderly, the ability to live at home longer and the quality of life. The direct medical care costs of fall accidents have already exceeded 1 billion euros in 2020. One of the goals of VeiligheidNL is to reduce the number of fall injuries in the elderly by developing effective interventions and disseminating knowledge and expertise (https://www.veiligheid.nl/kennisaanbod/valpreventie-bij-ouderen).
Fall prevention has been included in the 2021-2025 coalition agreement. This makes substantial resources available that contribute to the structural implementation of effective fall prevention for the elderly. Municipalities are given the task of providing fall prevention programs for their residents aged 65 and over. This requires investments in detection of elderly people at a high risk of falling, and the development and execution of fall prevention programs.

In this case study we would like to investigate if we can predict the fall risk of elderly people based on information that is collected in the primary health process, such as emergency room transcripts and electronic health records, and studies such as the Longitudinal Aging Study Amsterdam (LASA, https://lasa-vu.nl/).

Screening tools have been developed to analyse the fall risk of an individual elderly person such as the Fall risk 65+ (https://interventies.loketgezondleven.nl/leefstijlinterventies/interventieszoeken/1401588). However, these tools require substantial effort from medical professionals to fill the detailed analysis out, while there is already a lack of available medical professionals. Therefore we would like to use machine learning, and in particular text mining to make predictions about fall risk.

In medicine, usually risk predictions are developed using multivariable prediction models for individual prognosis as described in Collins et al [2015] (Collins, G.S., Reitsma, J.B., Altman, D.G. et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. BMC Med 13, 1 (2015). https://doi.org/10.1186/s12916-014-0241-z)


Prognostic multivariable modeling study
Subjects in a health state
Predictors:
Patient characteristics
Disease characteristics
Imaging tests
Laboratory tests
Others
Longitudinal relationship
Outcome: Development of event Y
Y
Y
Y
T = 0
End of follow-up

One of the major activities in such a study is to select the predictors, only predictors that are likely to contribute to the prediction, and preferably predictors that have proven to be a risk factor in earlier research will be taken into account. In this case study we will not select the predictors manually, but take a data driven approach and use machine learning to train an optimal model to predict fall risk, taking into account all the available information, i.e. open text (answers) as well as categorical data.
So by using machine learning we aim to have a more efficient way to develop and execute fall risk prediction tools. The feasibility of this approach is also confirmed in a case study by Ye et al [2020] (https://doi.org/10.1016/j.ijmedinf.2020.104105) on English EHR Data, but they did not use text mining on the notes in the record.
An important requirement of the models to be developed is that they are able to explain why the model has come to the conclusion of a high fall risk, i.e. which aspects are the risk factors. These risk factors can then be used to determine the most efficient intervention. A part of the case study could also consist of developing an interface in which the results of the explainable models are presented to the medical professionals.

Team Leaders

Academic Team Leaders

- Dr. Maya Sappelli, HAN University of Applied Sciences

- Dr. Rianne Kaptein

Gender Bias Detection in Job Ads

Textmetrics

About Textmetrics

Textmetrics offers an augmented writing assistant that gives writers support and feedback during their writing process. Most of our users are recruiters writing job applications. Our editor works with a number of modules, addressing different contextual aspects of text. This includes various discrimination checks but also feedback to the tone of voice and compliance to brand identity.

Our challenge:

One of our discrimination modules checks for gender bias in job advertisements. Fields which are dominated by a certain gender, job advertisements contain gendered wording. Especially for women, this can be a reason to not apply to a position they are perfectly qualified for. For this task, we would like you to come up with a way to predict gender bias in job advertisement.
Our current approach involves word lists that have shown to be either female or male connoted. However, at this moment we are only taking single words into consideration, because there are no scientific resources with lists of short expressions etc. Furthermore, the word list is only disambiguated by POS tagging. We hope that the gender bias which was shown to be encoded in the vector space of large language models can be utilized to detect gender biased words and expressions. Translating social psychological concepts into algorithms is challenging. There is no ground truth and no labeled data. There are pointers as to which concepts and words contain gender bias, you can use them as anchor points in your approach.
Additionally, the final product should be explainable in a way that the bias can be “reduced” to sets of words, phrases or sentences in the job advertisement. On the one hand this allows to give actionable advice. On the other hand, this adds the necessary transparency to the model, which is especially needed when it comes to discrimination topics.

Ideally we would like to be able to determine gendered words / phrases and suggest changes to the users of our platform. Given the challenging and open nature of this project, we do not want to define the final project too rigidly but give you the freedom to decide what fits. Your final product could be an extended list of words and phrases that you found while exploring the vector space or an approach to classify/quantify the gender bias in a text while keeping the explainability in mind.

Team Leaders

Academic Team Leaders

- TBD
- Kyrill Poelmans

Automatic Trailer Storyline Generation

RTL Nederland

About RTL Nederland

RTL Nederland is the largest commercial broadcaster in the Netherlands, with a yearly turnover of around €350m. RTL has the mission to tell unmissable stories that touch heart and mind. With a team of 10 data scientists, RTL works on challenges such as personalization, forecasting and automatic content generation.

Our challenge:

Every day, RTL produces many hours of video content that aims to touch the heart and mind of viewers. With advanced artificial intelligence technologies, we believe that we can support our creative people, by supporting the creative process of promotional content creation.
For an earlier case at ICT with Industry 2022, we picked up the challenge of automatically selecting scenes that could be used in trailers for TV programs. Our goal was to generate from a full length video a short teaser that increases interest in viewing the content, without spoiling the experience. While we were able to successfully rank scenes by their “trailerness”, we noticed we were missing key storytelling capabilities for making a compelling trailer. This is what we would like to study in our case for ICT with Industry 2023.
For our trailerness model, we created a dataset of episodes from our daily soap opera GTST with shot-level annotations of recurring video content. Each GTST episode starts with about a minute of recaps from previous episodes and ends with a short preview of the next episode. This recurring content makes for ideal training material in a weakly-supervised learning task for what content is interesting to include in promotional material.
We know from directors that good storytelling in trailers is key, as it compels attention. A better understanding of the narrative in video content would unlock an important aspect of the automatic generation of promotional video content. We found in our previous case that just selecting “trailerworthy” scenes is not enough, because a compelling trailer tells a story and understands the storyline.
We are seeking to close the gap between a good human-made trailer and automatically selected content. Our goal is to understand storylines better and to generate trailers that summarize future content more coherently. Combining the applied expertise of RTL and the Hogeschool Utrecht, will make for exactly the right atmosphere for tackling this challenge.

Team Leaders

Academic Team Leaders

- Dr. Stefan Leijnen, Hogeschool Utrecht
- Floor Schukking, Hogeschool Utrecht

- Dr. Daan Odijk
- Iskaj Janssen
- Prajakta Shouche