Overview - Measuring the quality of texts digitized by Optical Character Recognition

Mike Weltevrede

March 9, 2020

Contents

1 Introduction

The goal of this project is to investigate methods to measure the quality of OCRed documents. As such, this project does not consider preprocessing or postprocessing of the documents or the modelling of OCR software. It is known that there is a large impact of poor initial quality of the documents, especially for historical documents. This poor quality may be due to a distortion in brightness and contrast, as well as some random noise such as spots on the image versions of the documents. It is a prime field of research, for example on contrast adaptive binarization (Feng & Tan, 2004). However, we will not look further into these methods and will only focus on given OCR outputs, possibly with (hand-corrected) ground truths.

Actually, almost no research has been done to quantify the quality of OCRed texts without resorting to traditional quality measures that use the ground truth data, such as accuracy. Closest to this is research into post-processing, for example to clean words based on closeness to dictionary words. However, the measure we will develop should be independent of this so that it gives an indication of the quality even if there is no time and/or resources to do post-processing of the OCR output.

2 Materials and datasets

Historical documents (mostly newspapers and books) undergo OCR by ABBYY FineReader. The National Library of the Netherlands (KB - de Nationale Bibliotheek) firstly has ALTO XML files available as received as output from ABBYY. These contain the OCRed text divided up in several ways, most importantly on a word level with a given level of confidence in the word $WC \in [0, 1]$ and a confidence metric per character CC, which gives a score of 0 to 9 for the prediction of each character in that word where 0 is best and 9 is worst. Even though we do not know how these confidences are computed, the word confidence WC is adjusted towards 1 if the word that they recognized is found in a dictionary. So, even a word that has only 9s for the character confidences CC can have a high word confidence WC if only that word actually exists. The ALTO XML files also contain a page confidence measure $PC \in [0, 1]$.

In addition, the KB also has quote-on-quote ground truth data available for a lot of these texts. These are created by human correction of the OCR results. They promise that these are nearly perfect, though of course not free from error since humans can also overlook errors. In that sense, they recommend to take these as the actual ground truth of the OCRed text. Clearly, we will not have (digitised) ground truth data for future texts; it is the purpose of this project to indicate quality without the ground truth.

Because the documents of interest are often historical, regular dictionaries will not be suitable for use, if needed. We gained access to some dictionary-like websites and APIs, namely a word list by Instituut voor de Nederlandse Taal (INT - Institute for the Dutch Language)[1] and the Delpher API.[2] Lastly, the KB has provided metadata information that can be accessed via API as well.

3 Methods

Our current goal for the end of the project is:

> To develop a dynamic quality measure (having an adjustable parameter - *see weighted harmonic mean* - to indicate importance of aspects, such as searchability) with the ability to 1. discern the quality with respect to the entire corpus and "similar" documents (such as documents from the same time period or from the same source) and 2. break up the measure into the importance of individual components driving it.

With this, I mean to develop some sort of quality measure that is driven by an ensemble of methods to quantify quality. For instance, the length of the text or the words therein, the quality of significant words, or the distance from words to the dictionary. Ideally, I would like to make the measure dynamic in the sense that the user (in this case the KB) can set a level of importance for one or more of the parameters in determining the quality level. For example, it is clear to me that the text should be easily searchable on their database or website, so they will likely find this more important.

[1] `https://ivdnt.org/taalmaterialen/102-taalmaterialen/2126-tstc-int-historische-woordenlijst`
[2] Example for the word zwommen: `https://www.delpher.nl/nl/api/lexicon?wordform=zwommen`

3.1 Desired properties

Some aspects of the measure that are currently likely to be desired:

- The measure will be in between 0 and 1. This is up to arbitrary scaling so it should not be a problem. For the ground truth data, the measure should, therefore, evaluate to (nearly) 1.
- The measure can be converted to a categorical scale, such as *bad quality*, *okay quality*, and *good quality*. This is for the purpose of displaying the information on, for example, the website. Also, because the measure will (likely) not indicate a percentage of qualitativeness, it is not clear how a certain value should be interpreted: how much more different is 0.762 from 0.815? And what about 0.104 to 0.163?
- We can show how we determined the measure, i.e. it will not be a black box. It is most interesting to know what was the reason to classify a certain text as *bad quality*, for instance that many words were not found in a dictionary or that the text is extremely long or short.
- The measure can be contextualised with regards to the entire corpus and with regards to "similar" documents. We should explore which documents exactly we can classify as "similar". For instance, should we look at the time period? And do we then also subdivide this into articles and advertisements? Or do we consider the sources (e.g. which newspaper it was) separately?

3.2 Explored options

Firstly, do note that we have no (reliable) indication of quality yet on any of the texts. The only metrics that we have access to are the confidence levels as reported by ABBYY FineReader. Therefore, it is not possible to apply a supervised machine learning technique without manually labelling (some of) the texts ourselves first. We will elaborate on this in Section 4: Future research and limitations, where we also discuss the limitations of these techniques and if we can address them.

1. **Language classifier**

   We discovered the `langid` repository by `saffsd`.[3] This is a model that predicts the language of the text that is given as input. We tested whether this model classifies our texts as Dutch. As a conclusion, we stated that, if it does not, the text **may** be gibberish. However, it turns out there are also bilingual texts and documents in other languages in the corpus of the KB, so this may not be reliable. However, there is some merit in utilising this as a first check. For instance, we have seen an extremely bad quality OCRed text that was classified as Quechua because the original text of 564 characters was reduced to only 100 characters.

2. **Text lengths**

   Building onto the results from the previous points, the length of the text may also be an indication of the quality, as you would expect that newspapers and books generally are not that short. Looking at the tail text lengths of the texts in the ground truth corpus revealed that the lowest 5 text lengths out of 2,000 documents were, in order, 23, 56, 110, 190, and 229 words. Though not confirmed, I suspect that these short texts constitute advertisements

[3] `https://github.com/saffsd/langid.py`

and similar articles. All in all, we do see that the articles are relatively long and that a method using text lengths may have merit.

I tested the text lengths of the ground truth corpus and they fit a $Gam(\alpha = 41.68, \beta = 222.09)$ distribution ($p = 0.745$). See Figure 1. One could use this to determine a critical value below which we classify a text of a low (or high) length as having bad quality.

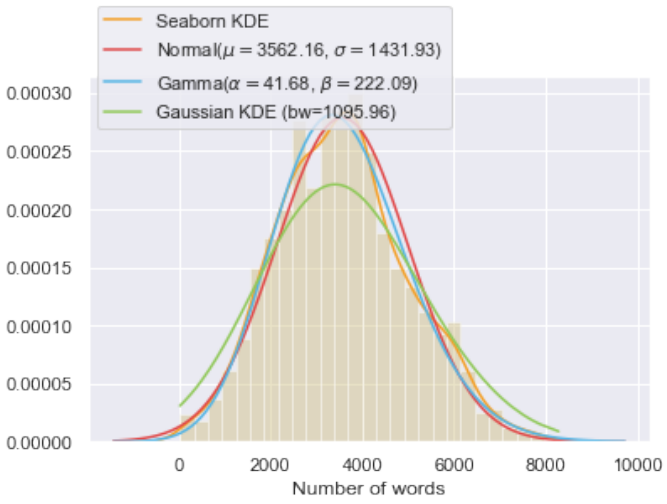

Figure 1: Lengths of texts in the ground truth corpus with fitted distributions.

3. Garbage words

From (Kulp & Kontostathis, 2007) and (Wudtke, Ringlstetter, & Schulz, 2011) we have learned about a method that tries to recognize whether a word in OCR output is "garbage" or not. Where (Kulp & Kontostathis, 2007) uses a rule-based approach, (Wudtke et al., 2011) build a garbage detection model using features based on (Kulp & Kontostathis, 2007) as input to a Support Vector Machine (SVM). Since SVMs are supervised learning models, they need labelled data. In this case, that meant whether a word was "garbage" or not. They achieved this with the IMPACT project. We are currently looking into whether we can access that data as well and, if so, whether it is available for Dutch texts too.

In the meantime, I implemented the rule-based algorithm by (Kulp & Kontostathis, 2007). In addition to these rules, I added one restriction and one rule of my own. The restriction is that I exclude words that are only one character long. The rule that I add is that words that consist for more than half of special symbols are garbage. Unfortunately, this algorithm was not successful. This seems obvious since I have not yet tested the actual numbers that they use. Likely, these are dependent on the language. For instance, German may have 4 consecutive consonants less often than Dutch (example: *schrikken* - to be startled). The following results were obtained:

Number of garbage words	0	1	2	3	4	5
Number of documents	1849	122	22	4	2	1

Table 1: Number of garbage words along number of documents

If labels can be obtained, it is highly likely that a good method can be constructed, whether it be SVM or another method.

4 Future research and limitations

Firstly, let us consider the main questions to be answered, after which we discuss other possible methods to determine the quality as well as limitations of the currently addressed methods.

4.1 Research questions

Getting to the measure

1. **Which methods to measure quality can and do we want to use?**

   We need to find several methods to determine the quality of a text. However, this does not mean that every one of these methods performs well. Tests should be performed to see if a method is useful or not. Starting slightly on the next question: if a model is used for which we can easily determine p-values and/or a confidence interval for the parameters, this can be used to test whether the parameter of a model is equal to zero. For less interpretable models, Monte Carlo simulation can be used to determine the above.

2. **How do we determine the weighting for several measures of quality?**

   If we do have labelled data, we could train a machine learning model for this purpose. As it is important to the KB to also know how the model achieved this weighting, we recommend a more interpretable model (for example, a neural network would be too difficult - regardless of the likely lack in available (training) data). Since the outcome variable would take clearly ordered discrete values like *bad quality*, *medium quality*, and *good quality* (or more granular), we recommend an ordered logit/probit model. The choice between the two is a matter of personal taste and should not matter much.

   If we cannot obtain labelled data, we need to manually determine some weights based on perceived importance and, perhaps more crucially, reliability (do note that this means that the weights will inherently be subjective). For instance, we have measures A, B, and C. We find measure A to be twice as important as measure B, while we think that measure B is thrice as important as measure C. Then, we would allocate the weights $(6, 3, 1)$ to measures A, B, and C, respectively. Once we have these weights, we can determine how to use them. There are several methods, for example a simple weighted average. However, there is some merit in utilising the harmonic mean[4], as also the F1 score uses. Most notably, it is less prone to outliers than a weighted mean. Do note that the harmonic mean cannot be used if one of the scores has value 0. Moreover, it is less interpretable than the weighted mean.

[4] https://en.wikipedia.org/wiki/Harmonic_mean

After getting the measure

1. **How can we best test the measure in a good statistical way?**

   That is, how well does it perform in determining the quality of the text? Moreover, it may be interesting to consider not just whether the model predicted correctly but also the distance to the correct label or whether it did so with a certain margin (Rennie & Srebro, 2005).

2. **How sensitive is the measure to noise?**

   It would be good if we could introduce some sort of noise ourselves. This would be done to the ground truth text, as the measure should be (close to) 1 for these texts so it is easily checkable. We could, for example, introduce many gibberish words or controllably change significant words (if these play a large role in determining the score) and see how this changes the score. For this, we would need to first have an indication of how much we expect the score to change.

4.2 Methods

1. **Language classifier**

   We currently can test for the language but we do not know how to deal with this yet. One could check for texts classified as non-Dutch whether a significant amount of words have matches in the dictionary of the language that is was assigned (e.g. French). However, this means that one should have a continuous access to dictionaries in many languages (also see below for possible limitations, even when these are accessible). Moreover, a threshold needs to be determined. Note that not many documents in the ground truth corpus have been classified as non-Dutch and only 5 documents had a different language than as classified in their ground truth as well (using the metadata as provided by the KB), meaning that it will not be possible to determine a (reliable) statistical bound.

2. **Text lengths**

   It was proposed that shorter texts might be indicative of wrongly OCRed texts. Nonetheless, one should be cautious as some sources simply utilise shorter texts and length in itself is not a measure of quality. One can combine this with some sort of unsupervised classification technique to determine if an article is expected to be short. However, as with the language classifier, we do not have much data on this which poses an issue.

3. **Garbage words**

   If we want to implement this method, we have to ask ourselves one question first: do we have access to labelled data and, if not are we willing to manually label data? In any case, we need to consider the rules for classifying a string as garbage. As mentioned before, the rules used by (Kulp & Kontostathis, 2007) and (Wudtke et al., 2011) are likely based on German words so they need to be tweaked to match Dutch data.

If we are able to achieve labelled data, we can apply a method simimlar to the one that (Wudtke et al., 2011) apply. They apply a Support Vector Machine approach to classify garbage words. We can either apply a similar method to classify an entire text or we apply the same method to (a sample of) a text and possibly set some sort of bound on which we determine that the entire text is garbage. In the former case, we only need labels on a text-wide level (in our sample $n = 2,000$) but the features are more complicated and need to be determined from scratch (in (Wudtke et al., 2011), they consider features on a word level) - for instance, instead of using the number of special characters in a string, we denote the number of words that have more special characters than some threshold (or quotient to the word length) λ in that text.

If we cannot obtain labelled data but we do want to apply garbage classification, we either apply the rule-based algorithm (with the tweaked rules) or we can try some sort of unsupervised classification method. The limitation of the latter is that such a method does not give us an indication of which clusters, if any clusters are even clear, constitute garbage words/texts.

4. **Dictionary distance**

   We mentioned before that we have access to the Delpher API and a word list by the INT. No analysis into the word list of INT has been conducted as the file format in combination with its memory size was tricky to work with. One should look into efficient methods to be able to look up words in this list, or to find another good dictionary source that also contains words in older spelling variations.

5. **Document similarity**

   A promising area is to control for document similarity. However, it is yet unclear how to define this. From the metadata of the KB, we for example have the publishing date available. It seems reasonable to expect that older documents are more difficult to recognise, for example because of wear and tear, but also because of different spellings and printing styles. Perhaps we could also find the font (e.g. Arial or Times New Roman) or typeface (e.g. Gothic) that the text was in, as we can expect that the ABBYY FineReader is not trained for being able to read all fonts.

6. **Quality of significant words AKA searchability**

   An important aspect for the KB is that people will be able to search for the text on their website. Consider the newspaper article header *Quarantainemaatregelen voor kwart van Italiaanse bevolking vanwege coronavirus*. People won't search for this article with simple words such as *voor* or *kwart* but rather more rare words that identify the text, such as *quarantainemaatregelen* and *coronavirus*. Given that we have 2,000 documents in our ground truth corpus, one could apply a method such as Term Frequency-Inverse Document Frequency (TFIDF) to try to identify significant words in texts. The argument usually is that words that appear the least are more indicative of the text. As such, we want lesser used words to be classified correctly. Do note that a surge in coverage on a certain topic, e.g. the *coronavirus*, may cause these terms that *are* often used as search terms to appear more than one could expect.

7. **Accuracy with respect to the ground truth**

   Note that this is only useful if we have access to the ground truth of the text which we generally do not have (otherwise we need not apply OCR in the first place). However, perhaps one can find a way in which this quick-fix is useful in combination with other methods as well.

   Two current issues are that ground truth and OCRed texts are not one-to-one comparable since there are spacing issues (e.g. "€7.50" in the OCRed text versus "€ 7.50" in the ground truth text) as well as hyphens appearing non-consistently (mostly hyphens to break a word to the new line). A quick fix to this is to simply remove all spacing and hyphens and then checking the similarity of the OCRed text with its ground truth text, for example with a distance metric like the Levenshtein distance.

8. **Regression on count of various errors**

   This method is applicable if we have labelled data and is a simple approach. One should first construct several errors, such as character error rate, word error rate, etcetera. Then, we can perform a simple regression (ordered logit/probit) or classification algorithm (e.g. K-means).

References

Feng, M.-L., & Tan, Y.-P. (2004). Contrast adaptive binarization of low quality document images. *IEICE Electronics Express*, *1*(16), 501–506.

Kulp, S., & Kontostathis, A. (2007). On retrieving legal files: Shortening documents and weeding out garbage. In *Trec.*

Rennie, J. D., & Srebro, N. (2005). Loss functions for preference levels: Regression with discrete ordered labels. In *Proceedings of the ijcai multidisciplinary workshop on advances in preference handling* (Vol. 1).

Wudtke, R., Ringlstetter, C., & Schulz, K. U. (2011). Recognizing garbage in ocr output on historical documents. In *Proceedings of the 2011 joint workshop on multilingual ocr and analytics for noisy unstructured text data* (pp. 1–6).