Преглед НЦД 22 (2013), 67–74

Geneviève Cron
Bibliothèque Nationale de France, Paris, France

OCR RATE COMPUTATION IN MASS DIGITIZATION PROGRAMS

Abstract: When digitization for libraries began about 20 years ago, the main issue was the scanning quality in order to obtain the best images both for conservation and dissemination. Since 2005, the Biblithèque Nationale de France (French National Library) has been launching tenders including conversion from image to text. This conversion can be done either by using software (Optical Character Recognition, OCR), or manually, or a combination of both. Irrespective of the course of proceedings the library expects the quality of the transcribed text. After description of the context, we will present the academic way of computing the OCR accuracy of an OCR output. Then, we will expose all parameters that any content holder needs to take into account for the definition of the OCR accuracy computation. This paper does not give any global formula for evaluation computation, but it raises questions for defining this formula. This paper will show that estimating the quality of the text depends on many of factors, including the use of the transcribed text. We also show that the reliability of this assessment is probably far from meeting the expectations of the libraries.

Keywords: digitization, OCR, Biblithèque Nationale de France, IMPACT project

1. Introduction: OCR and digital libraries

In order to offer a digital library that would display its valuable and fragile documents, the French national Library, Bibliothèque Nationale de France (BnF) began its digitization programs in the early nineties. Only scans, mainly black and white were required from service providers. After 15 years, as full text indexation began to make major improvements, tender were launched with some text conversion requirements. OCR output is stored in XML ALTO[1] format, which is maintained by the Library of Congress (LOC). Now, BnF's digital library Gallica[2] contains about 2 million documents, including books and newspapers with textual transcription, images, music scores, manuscripts, etc.

In order to convert an image containing text into a machine-encoded text, Optical Character Recognition engines were created in the early 1950's. Techniques [4] have been developed mainly for recent archives conversion. With the development of computer sciences, they have been applied to checks, postal sorting machines and in the era of digital libraries, to cultural heritage material, too. Algorithms conceived for handwritten material are very different from the ones for printed documents, mainly because the concept of 'character' is hard to isolate on a handwritten document, and one has to work on word level rather than on character level. Machine-printed OCR engines are all composed in the same way: some pre-processing is achieved (deskewing, dewarping, page splitting, binarisation); then comes the segmentation step which is composed of various sub steps (connex components computation (CC, see [7]), CC filtering, CC gathering into lines, and blocks).

The most difficult part of the OCR process [10] is segmentation. This is also the part where researchers are mainly focusing on the results improvements [3]. Segmentation can be processed at word level, based on white space between CCs. Several segmentation hypotheses at character level can be kept at this stage in order to choose the best one after the recognition step. Recognition is then applied on CC that has characteristics that allow interpreting them as characters. Each character model

[1] http://www.loc.gov/standards/alto/
[2] http://gallica.bnf.fr/

('a', 'b', ... 'A' ... 'Z' ... '1' ... ';' ...) gets a likelihood level of the character to be recognized. Here, each word hypothesis is submitted to the most likely entries in the dictionary, mainly using Levenshtein distance [9], to compute the global distance between the recognized word and the word in the dictionary. The best match can be either chosen, or, if distance is too great, rejected. At each step, an OCR has to handle probabilities, confidence levels, and decisions. Most OCRs are now able to attach a confidence level to each word. Details on the computation of the level are often inaccessible to the content holder. At this step, it is important for the content holder such as the BnF to have information about the global quality of the output text: is each paragraph, line, word described in the output file? Is there over information -e.g. noise? Is each word correctly segmented? Are they well recognized? What should be counted? These are the questions we are going to focus on in this paper.

1.1. OCR for what use? Evaluation for what use?

For most questions that arise in the evaluation issue, it is important to come back to the context of this computation, i.e. for what need transcribed text is produced. The main purpose of OCerizing a digitized book is indexation. But many other uses [6] can also justify the need for a transcribed text:

- **Indexation** was the very first purpose of transcription: content providers can no longer limit the query answer to the bibliographic record data, but want the output of queries to content information from the full text content of each document. This is very important for newspapers with consistent bibliographic records from issue to issue, and in which readers are looking for information at the article level, comprised in article or in article titles.
- **Indexation usefulness** computation appeared at the BnF after months of the full exploitation of OCR's outputs. The BnF carried out a study [8] which concluded that when estimated OCR accuracy is below 60%, the indexation process is polluted by so many misrecognized words that it is not efficient any more.
- **Quotations** are very useful for extraction of full paragraphs or pages.
- **E-books production** is a very interesting outlet for transcribed texts. This production has very high quality requirements, both in segmentation and recognition, but also in the structural encoding of the page.
- **Phonemisation**[3] is provided by the BnF to its users since 2008. This function helps blind people to access their collections.
- **Collaborative correction or ground truth (GT) production**, and evaluation of collaborative correction. Lots of projects are working on the correction of the OCR output to compensate for OCR defects. These programs allow users to correct misrecognized or doubtful words. But how to detect a) if a page needs to be corrected b) which words should be submitted to manual correction?
- **Word highlighting** allows the user to find the required word in a page. For newspapers, it is highly necessary because of the number of words on a page (about 40 000). This function is efficient when both recognition and localization of the word are good enough.

Why did quality evaluation of an OCR output become mandatory for a content provider? To answer this question we have to stress that most transcriptions (manual or via OCR or both) are done by service providers which are paid depending on the quality of their production. Thus, the first requirement is to control and evaluate service providers' work. If a threshold of the quality has been defined in the contact, quality insurance must be checked. If the quality is not what was expected, the

[3] Example: http://gallica.bnf.fr/ark:/12148/bpt6k54789860/f18.vocal.langFR

OCR output has to be rejected and re-processed. Second, the "OCR-ed" collection becomes a part of the library collection, thus it is the librarians' duty to know their collection as much as possible. Thirdly, in order to optimize the OCR-ing workflow, it is very important that the digital department is acquainted with expectation of an OCR, a collection or certain type of documents. Correlations and statistics could be done for further projects or tenders. In the same idea, selection in the future can be influenced by the knowledge of certain information about "what OCR quality can be expected on what kind of document".

1.2. OCR rate computing in research

OCR can be evaluated from various points of view: is each word correctly transcribed (precision), is each word present (non detection, recall), are there words that are not in the ground truth (false alarms)? Should the evaluation be done on character or on word level? Because most of the time OCR is done for indexation, in order to estimate the quality of a transcribed text, research scientists use a perfect transcription. They then compare it to the OCR output [11]. This requires an alignment, i.e., finding which word in the ground truth most likely corresponds to the word in the OCR output. Character level accuracy is more useful from a scientific or statistical point of view. At word level, once GT and OCR output are aligned, one has to compute:

- N: number of words in the GT;
- R: number of correct words in the OCR output;
- I: number of words present in the OCR output but not in the GT (inserted);
- M: number of words present in the GT but not in the OCR output (missing);
- E: number of misrecognized words (errors).

Where

$$N = R + M + E$$

Then a raw recognition rate at word level can be defined as

$$E_0 = \frac{R}{N}$$

One may, or not, take into account inserted, missing, and errors, with various weights on each error

$$E_1 = \frac{\alpha.R - \beta.M - \gamma.E - \delta.I}{N}$$

Character level accuracy gives higher recognition rates. Take for example a list of words containing only 10-character-long words. Suppose then that character recognition rate is 90%. This means that, on average, each tenth character is misrecognised, and that, in average, no word is correct. Here character error rate is 10% and word error rate is not far from 100%!

However the theoretical error rate is computed, it requires ground truth which is very hard and expensive to produce, so it is often not available in the context of mass digitization programs.

2. OCR rate computing in mass digitization context

2.1. Working without ground truth

This paper examines the situation of content holders who have to estimate the quality produced by their service providers. The produced data is, most of the time, a XML file describing each word which has been treated by the OCR engine (or the typist).

Contrary to evaluation in research field, no ground truth is available. Indeed, if GT were available, one wouldn't need to use OCR. GT can be produced on a very little number of pages, as was done in the EU founded Impact project. This is very helpful to have a precise idea of the quality of an OCR result on very specific material. Because no more than 0.01% of the corpus can be ground-truthed, one usually wants to find baselines there for estimating the OCR quality on a given page without GT. To bypass the GT problem, some researchers use texts that have been already transcribed (the Bible, e-books [12]). Nevertheless, a method that gives a quality level for each page cannot be defined using GT in mass digitization context.

Here, we focus on the estimation of OCR accuracy at word level. Theoretically, this is easy to compute. One should count the total number of words to be recognized (N), the total number of words which are well recognized (R) and the ratio R/N is the recognition rate. But one has to face a major issue: none of these values are available. To know the real value of both R and N, one would have to have the ground truth (GT), i.e. the perfect text (all the words, and all the words correct). Both these values have then to be estimated.

2.2. Estimation of number of words

The number of words (N) is known when the segmentation at all levels (block, line, word, character) is correct, which can never be assured. This number is usually estimated by the software itself, which supposes that the number of found words IS the number of real words. If one tries to estimate the precision (proportion of correct words among transcribed words), this number is not needed. But most of the time, one wants to compute the recall, i.e. the number of correct words among words to be transcribed. In a production process, it is almost never possible. In [2], a full algorithm is developed to face this estimation issue. It also gives the location of the textual elements that weren't detected by the OCR engine.

2.3. Estimation of number of correct words

The number R is the number of well recognized words. How to estimate whether a word is correct or not? A clue could be its presence / absence in the dictionary. This supposes that all words in the text are in the dictionary, which is quite unlikely (names, places for example). Moreover, some very close words (e.g. "word" / "world") can be confused. This type of error would not be counted. If the transcription was made automatically, some "word confidence" can be produced by the system. This value is not easy to understand because it's an internal value that is a combination of segmentation quality, character recognition quality and word decision quality, at least. This value can, on average, produce a page-level OCR rate. Alternately, one can fix or compute a threshold and R would then be the number of words that have a confidence level higher than this threshold.

3. Advanced Parameters for OCR rate computation

3.1. Segmentation quality

The segmentation process is the part of the OCR which computes the location of the text area. In XML ALTO, zones are delimited by a rectangle. Three levels have major importance: blocks, lines, words.

3.1.1. Segmentation quality at block level

The quality of the segmentation at block level is defined by the match between the real text block and the text block found by the OCR:

- Are all textual blocks detected by the OCR? This is related to the recall. Any missing block means missing line and words.
- Are all detected block really textual blocks? This is related to the precision. This determines if there is any noise in the detected text.
- Is location (coordinates) precise enough? Is the bounding area precise enough, does it contain any other information which is not textual? Are there missing words?
- Are the blocks merged or split? For some uses (indexation, quality measure) it is not so important. But if the user wants to understand the global structure of the text, it is important: for the phonemisation, the end of the paragraphs or sentences are the key points for deciding the pronunciation of the last word, and for determining the time for breathing between words. For e-book conversion, if one block is split into two, the layout will be incorrect.

These topics are discussed in detail in [1].

3.1.2. Segmentation quality at line level

The segmentation of the lines is the process where one decides where the lines are located. Except in very complicated layouts, and because line segmentation is often the very first algorithm in the segmentation process, it is often efficient. The line level is often less studied, probably because word and blocks are the basic concepts in the use of OCR. This is highly correlated with both,
word and block segmentation.

3.1.3. Segmentation quality at word level

The segmentation of words is the process where a line is split into words. The quality of the segmentation at word level is measured by:

- Are words split or merged? This is important for all uses of the transcribed text, and, very basically, for the indexation: a split or a merged word cannot be indexed (in the indexation process). Even if each a character is well recognized, nothing can be done with the missegmented word. This is a major limitation of the OCR engines.
- Are words precisely located? This is important only for word highlighting.

3.2. Reading Order and article level description

In newspapers, text blocks are numerous, and the reading order and the break up into article may be very complicated. They may also show some complex layouts. For some uses like e-book production, the quality of both, reading order and the article level description, has to be taken into account inthe quality evaluation of the OCR process.

3.3. Validity of words and blocks

Let us suppose we want to count the accuracy at word level. Should every word be counted? What if it is illegible?

3.3.1. Stop words and numbers

Because the very first outlet of OCR was indexation, the quality measure was computed against the quality of the outputs of the queries. As most queries are based on meaningful words, the most frequent short words like articles or auxiliary verbs would have biased the computation of the OCR rate. Thus, most OCR accuracy rates are based on non-stop words (non empty words). If the intended use is a full text (for quotations, e-books production, phonemisation), it is necessary that the stop-word list should be empty and the evaluation should be done on all the words in the document. Numbers have been excluded from the computation of OCR accuracy for the same reasons. They should be completely included in order to calculate accuracy in the new uses as well.

3.3.2. Diacritics

In French, there are diacritics (accents) in about 20% of the words. Diacritics errors are very frequent because they are very small components that can disappear in the binarisation or segmentation process. Two words can have exactly the same spelling but for the diacritics: "tache" and "tâche", "jeune" and "jeûne". The very first evaluation system at BnF didn't count diacritics errors. It was related to the fact that the indexation system did not use the diacritic information in its internal system. For new uses diacritics errors have to be taken into account.

3.4. Zones

For convenience, each word, line and block can be considered as important as any other. But should one give the same importance to main articles, titles, running titles, footnotes, advertisements or tables (with numbers)? A complete scope of this question is proposed in [5]. At the BnF now there is no weighting system in place.

3.5. Illegibility

And what if a block, a word, a line or a character is illegible? Illegibility should then be defined but according to what? Software? An operator fluent or not in the written language? These notions have been much discussed recently at the BnF. One can consider that if a book is globally readable for the OCR (known language and script, not-too-bad background, not-too-strange font family), a block with a high percentage of:

- hard to segment lines / words
- non recognized words
- words with very low word confidence
- words out of the dictionary

is probably illegible. This issue can prove delicate in a contractual situation where the concept of readability has to be very precisely described, at word and paragraph level.

3.6. Word confidence precision

Global word accuracy on a page can be estimated using various piece of information at word level, mainly using a word confidence level. For the content holders, it is very hard to obtain much information on the way the word confidence has been computed. Then, in case the global accuracy estimation is based on any function (mean, median, number of word above a threshold) related to this value, it is hard to master the global precision of the page.

Table 1: Table of **segmentation** parameters for OCR Evaluation according to the use of the transcribed text

	block segmentation contol	Word segmentation contol	Layout	Reading Order
Indexation		+		
Quotation	+	+	+	+
Ebooks production	+	+	+	+
Phonemisation	+	+	+	+
Correction		+		
Word highlighting		+		

Table 2: Table of **recognition** parameters for OCR Evaluation according to the use of the transcribed text

	Diacritics	Empty words	Illegibility	Word confidence
Indexation	depending on system requirements	basic uses: yes else: no	+	
Quotation	+	no		
Ebooks roduction	+	no empty words	no zone should be illegible	
Phonemisation	+	no		
Correction		depending on use after correction		+
Word highlighting		yes		

4. Tables' parameters vs. use of transcribed text

In section 3, we described major points of interest for the computation of the OCR accuracy. We saw that the parameters to be taken into accounts depend very much on the use of the transcribed text or OCR output. Tables 1 and 2, show the complete and synthetic information about each duet {use, parameter}.

5. Conclusion

To conclude, it appears that the computation of an OCR rate on page level is highly dependent on the intended use of the OCR output and the terms of the conversion operations. If the aim is to index the document, one may not take into account the empty words and punctuation. If the indexation is done without diacritics, they do not have to be taken into account in the error rate. Conversely, if the aim is to produce full text for quotation or e-book creation, all these errors should be counted. Anyway, the rate would be a poor estimation since both, the numbers of words and the number of correct words are estimated by the system producing the output. It is important to retain as much information as possible from the production process to be able to re-compute the estimated OCR rate in case of a reorientation of the use of the text.

References

[1] C. Papadopoulos A. Antonacopoulos, C. Clausner and S. Pletschacher. Historical document layout analysis competition. In *Proceedings of the 11th International Conference on Document Analysis and Recognition (IC-DAR2011), Beijing, China, September 2011, pp. 1516-1520.*

[2] N. Ragot A. Ben Salah, G. Cron and T. Paquet. Automatic ocr quality control without ground truth : Segmentation verification. *ICPR, International Conference on Pattern Recognition,* Nov 2012.

[3] R.G. Casey and E. Lecolinet. A survey of methods and strategies in character segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 18(7):690-706, jul 1996.

[4] Mohammed Cheriet, Nawwaf Kharma, Cheng-lin Liu, and Ching Suen. *Character Recognition Systems: A Guide for Students and Practitioners*. Wiley-Interscience, 2007.

[5] C. Clausner, S. Pletschacher, and A. Antonacopoulos. Scenario driven in-depth performance evaluation of document layout analysis methods. In *2011 International Conference on Document Analysis and Recognition, IC-DAR 2011, Beijing, China, September 18-21, 2011,* pages 1404-1408. IEEE, 2011.

[6] G. Cron. Use of digitised and ocred text collections by end users. *In Research results and practical experience from the IMPACT project*, may 2009.

[7] Michael B. Dillencourt, Hannan Samet, and Markku Tamminen. *A general approach to connected-component labeling for arbitrary image representations*. *J. ACM,* 39(2):253-280, April 1992.

[8] D. Stutmann. Affichage du texte issu de la reconnaissance optique de caractères (ocr) : propositions du groupe sinum / visualisation. 2008.

[9] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*., 10(8):707-710, February 1966.

[10] Cheng-Lin Liu and Hiromichi Fujisawa. Classification and learning methods for character recognition: Advances and remaining problems. In Simone Marinai and Hiromichi Fujisawa, editors, *Machine Learning in Document Analysis and Recognition*, volume 90 of Studies in Computational Intelligence, pages 139-161. Springer Berlin / Heidelberg, 2008.

[11] S.V. Rice, J. Kanai, and T.A. Nartker. An evaluation of ocr accuracy. In *SDAIR93,* page XX, 1993.

[12] Ismet Z. Yalniz and R. Manmatha. A Fast Alignment Scheme for Automatic OCR Evaluation of Books. pages 754-758, September 2011.

gcron@free.fr