ASSESSING THE QUALITY OF DIGITAL RE-PUBLISHING OF TEXTUAL DOCUMENTS THROUGH THE FOLLOW-UP OF A CORRECTION PROTOCOL BY CROWDSOURCING

Marthe Lagarrigue, Florence Rossant, Alain Pierrot, Joël Gardes, Christophe Maldivi, Eric Petit

ASSESSING THE QUALITY OF DIGITAL RE-PUBLISHING OF TEXTUAL DOCUMENTS THROUGH THE FOLLOW-UP OF A CORRECTION PROTOCOL BY CROWDSOURCING

Marthe Lagarrigue[(1)], Florence Rossant[(1)], Alain Pierrot[(2)], Joël Gardes[(3)], Christophe Maldivi[(3)], Eric Petit[(3)]

[(1)]ISEP, Institut Supérieur d'Electronique de Paris, France ;
[(2)]i2S, France; [(3)]Orange Labs, France

ABSTRACT

Digitized re-publishing of documents has become nowadays a very important issue. Optical Character Recognition (OCR) has been intensively used to this aim, as it performs the transcription of the text images into electronic files, allowing display functionalities, indexation, enrichment and broadcasting. However, such software still fails in many configurations, so that the transcription does not reach the required editorial quality (99% of recognition are required for an ergonomic reading). In the OZALID project, we propose to rely on crowdsourcing for correcting OCR results. One main issue is then to determine when the crowdsourcing has reached its limits. For that, we present a feasibility study of an original protocol based on indicators that quantify the recognition quality in both semantic and semiotic ways. These indicators are calculated and followed up during the entire crowdsourcing process until stability. Experimental results show that the proposed observables converge after some correction iterations allowing automatically stopping the crowdsourcing process and dealing with huge amount of data.

Index Terms— digital edition, crowdsourcing, quality assessment, OCR, correction protocol, semantics, semiotics

1. INTRODUCTION

The preservation and the availability to the public of documents stored in libraries have become a very important issue. Gallica, a BnF[1] (*Bibliothèque nationale de France*) library of digitized documents, has been created to this aim.
Optical Character Recognition (OCR) software goal is to retrieve the textual information from images of documents, obtained through a digitization process. The output is then stored as ALTO[2] files, an electronic document format with the purpose of analysing the textual content of document images while storing the layout coordinates of textblocks. Yet, even if OCR algorithms reach nowadays high recognition rates for simple and well-printed documents, they still fail in specific cases, where document complexity, variability and quality are out of expectation. Such issues may come from the layout complexity, the quality of the original document, the typewriting itself, the digitization process, and so on. Hence, the OZALID project subscribes to the CIDRE [1] – Cooperative and Interactive Document Reverse Engineering – philosophy which aims at bringing humans in the loop, through a crowdsourcing process, admitting limits of OCR technologies. It is a cybernetics approach [3] [4]. Internet users are invited to spot content errors and interactively correct them through a User Interface (UI). One main and challenging issue is to determine automatically at which correction step a consensus is almost found; then the document can be sent to the chief editor, alleviating his checking task.
User corrections may provide several text versions of a document. The main issue is then to follow up their quality and determine automatically when a satisfying version is obtained. The proposed solution consists in elaborating a protocol which assesses the quality of character identification by analysing the correction process evolution. The document structure recovery is not handled here. In short, the text quality is said "good enough" when a majority of correctors have no additional correction to propose. The role of the editor will be to check if the quality level is actually sufficient. The main contribution of the proposed study is the design of a quality assessment protocol, capable of processing huge amount of data without the need of any ground truth. Here is presented a feasibility study of the protocol with experimental results obtained with two data sets. This paper is organized as follows. First, we describe some related work (section 2). Secondly, we give a general description of the proposed protocol (Section 3). Experimental results are presented in the Section 4 before conclusion and perspective remarks.

2. RELATED WORKS

To the best of our knowledge, no automatic and general process has been proposed in the literature for assessing document reproduction quality, ensuring high-fidelity conformity of the electronic version to the original document.

[1] http://www.bnf.fr
[2] http://www.loc.gov/standards/alto/v3/alto-3-0.xsd, http://www.abbyy-developers.eu/en:tech:features:alto

Crowdsourcing is a very common method for correcting content of digitalized documents. For instance, the work of Garby et al. [5] underlines the great impact of crowdsourcing with Mechanical Turk, and the one of Anh-Hoang et al. proposes a model and a tool to collect users' traces [7]. When it comes to comparing the quality of two text versions of a document, there are two standard methods. The first one compares the number of changes necessary to switch from one version to the other. This method, applied in several works is based on the Levenshtein distance (edit distance) [8] or some of its improved derivative [9]. The Levenshtein distance measures the similarity between strings by computing the minimal cost to align them through basic insertion, deletion and substitution operations. This kind of measure is very common in pattern recognition and error corrections [10] [11]. Unfortunately, such method cannot decide which recognition result is the best one. A reference document, i.e. a ground truth, is needed to make a decision, see for example the work of Gardy et al. [5] or Belaïd et al. [12]. This is a real drawback for systems working with big data sets.

The second method for comparing two text versions aims at analysing the difference between quality indicators computed on each one. A lot of indicators have been proposed to describe a document and the identification of its characters. For example, Ben Salah's layout parameters [13] measure text areas in the document. OCR parameters, such as Word Confidence (WC), which measures the confidence on the character identification and Word Dictionary (WD), which states if a word exists in a dictionary or not, are also straightforward indicators.

In this article, we propose a methodology which analyses the evolution of observables evaluated on every emended version provided by the crowdsourcing. Each observable quantifies the quality of the identification with respect to a specific criterion: based on image analysis features or referring to a dictionary. We assume that the quality level can no more be increased within the system when the proposed observables converge and remain stable, indicating that the correctors have reached a consensus.

3. PROTOCOL FRAMEWORK

The protocol aims at assessing the evolution of the text recognition during correction iterations performed by crowdsourcing. It takes place in two main phases: the document content correction by crowdsourcing and the measure of indicators on each corrected version.

3.1. Document corpus

We tested our protocol on two book extracts. The first one comes from *L'art d'être grand-père* written by Victor Hugo, monograph of 1881, edition C. Levy; the second extract comes from *Le vampire de Dusseldorf*, 1932, edition Le Livre national (Paris). Figure 1 emphasizes strong differences between both documents: *Hugo* is straightforward to process while *Le Vampire* contains thick, close and noisy characters hard to segment. These two documents are thus complementary: the poor OCR result of *Le Vampire* implies much more corrections than the *Hugo* extract.

a) Page of *Hugo* b) Page of *Vampire* c) Zoom (*Hugo*) d) Zoom (*Vampire*)

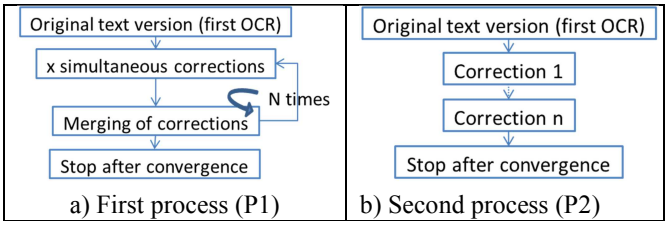
Figure 1: images and zooms extracted from our corpus

Each document extract contains about 40000 characters: about 80 pages for *Hugo* and 12 pages for *Le Vampire*. The crowdsourcing process is performed on a web UI which displays the digitalized images together with the OCR ALTO text output, both provided by the BnF. ALTO files include several confidence indicators as well as character identification and text/word boundaries. Unfortunately, indicators such as word and character confidence scores cannot be trusted since they may be calculated in various ways by different OCR engines.

3.2. Correction protocol: crowdsourcing simulation

We have experimented two crowdsourcing processes. In the first process (P1), the text is simultaneously and independently corrected by several correctors. Their corrections are merged. The merged version is then corrected, in the same way and the process is repeated several times (cf. Figure 2.a). In the second one (P2), the text is successively corrected by several correctors, one by one (cf. Figure 2.b).

Both processes are very realistic, with cases of conflicts between simultaneous corrections (P1) and instabilities in corrections (P1, P2). The conflicts are resolved by majority vote during the merging process (P1). If the alternatives have the same cardinality, the original version is kept. The second process (P2) generally achieves a faster stabilization of the corrections, i.e. accordance between correctors.



a) First process (P1) b) Second process (P2)

Figure 2: Illustration of the two crowdsourcing processes

Volunteers in our study used a web interface developed by Orange Labs[3] for the corrections. Our experimentation involved eight volunteers with various profiles, whose one who sometimes makes intentional errors, and provided five corrections by crowdsourcing process.

3.2. Quality assessment protocol: indicators

We propose two complementary categories of observables. The first one gives priority to semiotics [14], i.e., fidelity to the digitized image; observables of this category are therefore based on image analysis techniques. The second one favours semantics, i.e., conformity of the document; in this case, used tools are dictionary and statistical lexical analysis. Sometimes both are conflicting as in the following example, which can be interpreted either as *pièces* or *pièees*:

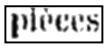


Figure 3: Ambiguous case: semantics vs. semiotics

In this case, we have to favour semantics tools and read *pièces* which exists in French dictionary. On the contrary, in the *Lettre du petit Lily* in *Le château de ma mère* of M. Pagnol, intentionally written with a lot of orthographic errors, the identification should rely mostly on the image.
In our study, the proposed observables come from two software tools. The first one refers to image processing and clustering technics developed by Orange Labs, and the second one is the ABBYY Finereader[5] OCR as used by i2S[4], both labs collaborating to this work.
Observables are measured on every corrected version of the text provided during the crowdsourcing process.

3.2.1. Characters confidence via image analysis techniques

A character confidence indicator, named CCF, has been proposed by Orange Labs in order to estimate the validity of the recognition and also control the users' correction process. Its principle relies on a consistency measure between a set of similar character shapes and their semantic values (i.e. their labels). Its calculus is based on the following steps:

- Segmentation of the bounding box of the words (given in the ALTO file), in order to extract the character images on an entire page.
- Clustering analysis applied on the whole set of character images. For that, an agglomerative clustering algorithm has been implemented, based on a single-linkage criterion.
- For each character shape *s(i)* of label *l(i),* belonging to the cluster *C(i),* counting of the number *n(i)* of shapes inside the cluster *C(i)* that are also labelled *l(i)*.
- The CCF is then given by $ccf(i) = n(i)/N(i)$, where *N(i)* is the total number of elements in the cluster *C(i)*.

Hence, for any given shape, if all the other shapes of the same cluster have the same label than it, its CCF score is $(N(i)-1)/N(i)$. Thus, this score tends to 1 when the cluster size increases, meaning that consistency is maximum considering semiotic aspects. Conversely, if the CCF score is close to zero we are probably faced with either a mis-recognition or a segmentation concern.

3.2.2. The OCR indicators

OCR software usually computes two indicators, namely WC and WD. As we want to analyse the evolution of these scores, we have to guarantee that they are calculated in same conditions. However, the ones given by the initial recognition process cannot be reproduced since the first OCR is not known nor fixed (cf. section 3.2). For this reason, we propose to generate new images from every text version (including the first one), in known and constant conditions, and to apply a known and efficient OCR software to these new images. The OCR chosen in our study is ABBYY's FineReader[5]. WC and WD values provided by ABBYY for every version can then be compared.
The image generator creates a synthetic bitmap from a text document (cf. [15] and [16]) given font, background and noise properties, it has been developed by the LaBRI lab[6], partner of this work. The font is created from the original images of the document, whereas the text comes from the corrected version produced by crowdsourcing, which has to be evaluated.

3.2. Global protocol synthesis

Figure 4 synthesizes the quality assessment protocol that we propose for the evaluation of the text recognition quality, throughout each correction process performed by crowdsourcing. The clusters and CCF are calculated thanks to the original images as well as the current text version. In parallel, the observables WC and WD are calculated by the ABBYY OCR applied on images generated from the current text version.

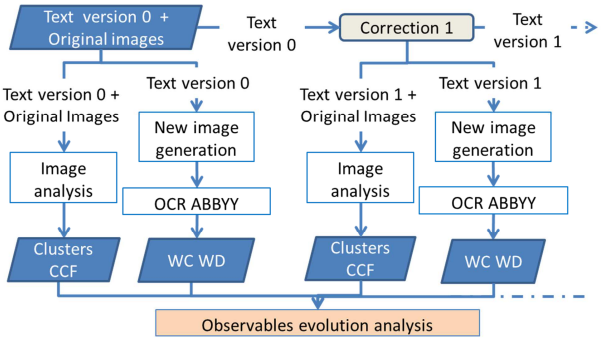


Figure 4 Synthesis of the quality assessment protocol

[3] http://www.orange.com/fr/innovation
[4] http://www.i2s.fr/
[5] http://www.abbyyonline.com/
[6] http://www.labri.fr/

The last step consists in analysing the evolution of these observables, based on statistics calculated on them. The results are presented in the following section.

4. RESULTS

The analysis focuses on the evolution of the indicator values; therefore, the shown graphics represent the evolution of indicators in %. An evolution tending to 0 means that the text identification converges to a stable and satisfying version according to the users.

Our experiments are composed of four different flows since we consider two different crowdsourcing processes applied to two document extracts. The following indicators have been studied, describing:

- global statistics related to identification: the numbers of identified characters and strings;
- semiotics aspect: mean and standard deviation of the CCF, number of element with CCF=1, WC;
- semantics aspect: WD.

We present one graphics by document with the evolution of all the indicators, expressed in %. $\Delta(t_i)$ describes the evolution in % of the score indicators between the i^{th} and the $i+1^{th}$ correction.

Figure 5 emphasizes the evolution of indicators for the *Hugo* extract treated with the crowdsourcing process P2. Here are two important observations. First, there is a big variation between the first and the second correction steps mainly due to the great amount of artefacts considered as characters by the OCR and quickly corrected by users. Second, the variation of all studied indicators tends to 0 in a quite small number of iterations, which indicates, as expected, that correctors reach agreement. This tendency is very similar for the three other experimented flows.

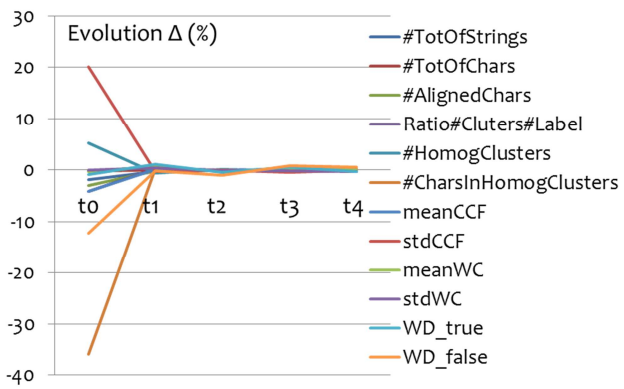


Figure 5 Evolution of the indicators in a Hugo extract

In order to analyse more precisely the indicators evolution, we propose the following two tables.

Table 1 which can be viewed as a very basic fusion rule of all criteria. Table 2 presents the corresponding standard deviation.

	Δ_{1-0}	Δ_{3-1}	Δ_{3-2}	Δ_{4-3}	Δ_{5-4}
Hugo P1	0,72	2,67	0,39	0,40	0,22
Vampire P1	9,70	0,59	0,31	0,19	0,46
Hugo P2	7,36	0,39	0,19	0,23	0,13
VampireP2	3,66	1,89	0,15	0,17	0,00

Table 1: Mean of the absolute value of indicator variation measures, expressed in %

	Δ_{1-0}	Δ_{3-1}	Δ_{3-2}	Δ_{4-3}	Δ_{5-4}
Hugo P1	1,27	3,62	0,92	0,90	0,56
Vampire P1	3,97	1,09	0,11	0,10	0,19
Hugo P2	10,76	0,30	0,31	0,25	0,19
VampireP2	4,45	3,37	0,18	0,21	0,00

Table 2: Standard deviation of the absolute value of indicator variation measures, expressed in %

These two tables show that all indicators tend to 0 for the four experimented flows. This means that for this set of correctors, we are near to obtain a stable state of the identification, i.e., the correction activity can stop. It is worth noting an increase of the global variation measure for *Le Vampire* processed with P1, between the 4th and the 5th corrections. Indeed, one corrector introduced intentional errors which were corrected during the 5th iteration.

5. CONCLUSION AND PERSPECTIVES

This work is a preliminary study for assessing the quality of digital re-publishing of textual documents through the follow-up of a correction protocol by crowdsourcing. It is based on statistics, semantics and semiotics indicators. The experimental results presented above are very positive, since they show a convergence of the indicator variation measurements towards 0, meaning that a consensus has been reached between correctors. The protocol is being integrated on the OZALID platform and it will be tested on large databases during the next experiment.

Further investigations will focus on two important points. The first one will address the issue of indicators interpretation by implementing advanced fusion rules, especially when semantics and semiotics diverge. The second one will assess the establishment of a decision criterion notifying the end of the correction process, meaning that the current document version can be sent back to the editor.

Acknowledgements: We would like to thank Nicolas Journet, Vincent Labreux and Kieu Van Cuong from the LaBRI lab for their kind and efficient collaboration to this work. We also thank Isabelle Josse and Sébastien Roux for their participation to the document correction and all the interesting discussions when elaborating this protocol. Finally, we are grateful to Jean-Luc Bloechle for his help brought to this article.

This study was accredited by Cap Digital, and has been financed by BPI-France, Ile-de-France Region, General Council of Seine-Saint-Denis.

6. REFERENCES

[1] F. Bapst, R. Brugger, A. Zramdini and R. Ingold, "Integrated Multi-Agent Architecture for Assisted Document Recognition," in *DAS'96*, 1996.

[2] K. Hadjar, O. Hitz, L. Robadey and R. Ingold, "Configuration REcognition Model for Complex Reverse Engineering Methods: 2(CREM)," *Document Analysis Systems V Lecture Notes in Computer Science,* vol. 2423, pp. 469-479, 2002.

[3] G. von Bonin, "Cybernetics or control and communication in the animal and the machine," *The bulletin of mathematical biophysics,* vol. 2, no. 11, pp. 145-147, 1949.

[4] N. Wiener, "Speech, langage and learning," *The Journal of the Acoustical Society of America,* no. 22, 1950.

[5] C. Grady and M. Lease, "Crowdsourcing document relevance assessment with Mechanical Turk," in *Proceedings of the NAACL HLT, Workshop on creating speech and laguage data with Amazon's Mechanical Turk*, Los Angeles, 2010.

[6] H. Gelas, S. T. Abate, L. Besacier and F. Pellegrino, "Quality assessment of crowdsourcing transcriptions for African languages," *Proceedings of interspeech,* 2011.

[7] L. Anh-Hoang, M. Lefevre et A. Cordie, «Collecting Interaction Traces in Distributed Semantic Wikis,» *3rd International Conference on Web Intelligence, Mining and Semantics,* vol. 21, pp. 1-11, 2013.

[8] V. Levenshtein, "Binary codes capable of correcting deletion, insertions, reversals," *Soviet Physics-Doklady,* pp. 707-710, 1966.

[9] L. Yujian and L. Bo, "A normalized Levenshtein distance metric," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 29, no. 6, p. 1091–1095, 2007.

[10] S. J. F. n. T. Rice, "The tird annual test of OCR accuracy," *Annual raport of ISRI, University of Nevada, Las vegas,* pp. 11-40, 1994.

[11] N. Munyaradzi and H. Suleman, "Quality assessment in crowdsourced indigenous language transcription," in *International Conference on Theory and Practice of Digital Libraries*, Valletta, Malta, 2013.

[12] A. Belaïd and L. Pierron, "A generic approach for OCR performance evaluation," *Electronic Imaging,* 2002.

[13] A. Ben-Salah, N. Ragot and T. Paquet, "Adaptative detection of missed text areas in OCR outputs: application to the automatic assessment of OCR quality in mass digitalization projects," in *Proceedings of SPIE*, 2013.

[14] J. Gardes, Le document numérique : la complexité des formes et les formes de la complexité Systèmes et interfaces utilisant des descripteurs sémiotiques, Institut National des Sciences Appliquées de Lyon, 2009.

[15] V. Rabeux, N. Journet, A. Vialard and J.-P. Domenger, "Document perceptual quality ground-truth creation," in *10th IAPR International Workshop on Document Analysis Systems*, Gold Coast, Queensland, Australia, 2012.

[16] V. Kieu, N. Journet, M. Visani, R. Mullot and J. Domenger, "Semi-synthetic Document Image Generation Using Texture Mapping on Scanned 3D Document Shapes," in *ICDAR*, Washington, DC, USA, 2013.