ICDAR 2019 Competition on Post-OCR Text Correction

Christophe Rigaud, Antoine Doucet, Mickaël Coustaty, Jean-Philippe Moreux

ICDAR 2019 Competition on Post-OCR Text Correction

Christophe Rigaud*, Antoine Doucet*, Mickaël Coustaty* and Jean-Philippe Moreux†
*Laboratory L3i, University of La Rochelle, 17042 La Rochelle CEDEX 1, France
Emails: {christophe.rigaud, mickael.coustaty, antoine.doucet}@univ-lr.fr
†National Library of France, Quai François Mauriac, 75706 Paris, France
Email: jean-philippe.moreux@bnf.fr

***Abstract*—This paper describes the second round of the ICDAR 2019 competition on post-OCR text correction and presents the different methods submitted by the participants. OCR has been an active research field for over the past 30 years but results are still imperfect, especially for historical documents. The purpose of this competition is to compare and evaluate automatic approaches for correcting (denoising) OCR-ed texts. The present challenge consists of two tasks: 1) error detection and 2) error correction. An original dataset of 22M OCR-ed symbols along with an aligned ground truth was provided to the participants with 80% of the dataset dedicated to training and 20% to evaluation. Different sources were aggregated and contain newspapers, historical printed documents as well as manuscripts and shopping receipts, covering 10 European languages (Bulgarian, Czech, Dutch, English, Finish, French, German, Polish, Spanish and Slovak). Five teams submitted results, the error detection scores vary from 41 to 95% and the best error correction improvement is 44%. This competition, which counted 34 registrations, illustrates the strong interest of the community to improve OCR output, which is a key issue to any digitization process involving textual data.**

I. Introduction

The accuracy of Optical Character Recognition (OCR) technologies considerably impacts the way digital documents are indexed, accessed and exploited [1], [2]. During the last decades, OCR engines have been constantly improving and are today able to return exploitable results on mainstream documents. But in practice, digital libraries have on shelves many transcriptions with a quality below expectation. In fact, ancient documents with challenging layouts and various levels of conservation such as historical newspapers still resist to modern OCRs.

These challenges are addressed by the research project NewsEye[1] which is supporting this competition. It will introduce new concepts, methods and tools for digital humanities by providing enhanced access to historical newspapers for a wide range of users. With the tools and methods created by NewsEye, crucial user groups will be able to investigate views and perspectives on historical events and development and, as consequence, the project will change the way European digital heritage data is (re)searched, accessed, used and analyzed.

[1]https://www.newseye.eu

Moreover, formerly digitized resources processed with outdated OCRs are rarely re-sent through the latest state-of-the-art digitization pipeline, as priority is often given to the ever-growing masses of new arriving documents. In this context, OCR post-correction approaches, either used on former digitized documents or on fresh challenging documents, could strongly benefit digital libraries.

In this context, the competition was open to researchers from several fields (document analysis, natural language processing, data analysis, text data mining, machine learning...) to challenge their method(s) for improving/denoising OCR-ed texts. The benefit is double as it gives a global overview of the methods developed by the community and it sets down a common baseline for further works. An analysis of the state of the art shows that it remains difficult to find benchmarks to assess the performance of OCR correction algorithms.

The first post-OCR text correction competition took place during the ICDAR 2017 conference, it resulted in a publicly available dataset[2], evaluation tool[3] and a paper [3]. The text data consisted in more than 12 million characters from French and English languages and included both noisy OCR-ed texts and the corresponding aligned ground truth (Gold Standard).

This second edition of the competition focused on an almost twice as big and multilingual dataset with 22 million characters in 10 European languages. The dataset was distributed similarly to the first round, using the same evaluation metrics (see Section II-C) and script[2].

II. Competition setup

In this section we describe the two tasks, the dataset, the evaluation and the modalities given to the participants.

A. Task description

Contrary to the first edition of this competition, it has been decided to divide the challenge into two inter-dependent tasks (see Fig. 1), which means the participants have to propose an error detection scheme (task 1) in order to be able to participate on task 2 (error correction). This allowed participant to propose hybrid methods combining the two

[2]https://l3i.univ-larochelle.fr/ICDAR2017PostOCR
[3]https://git.univ-lr.fr/gchiro01/icdar2017/tree/master

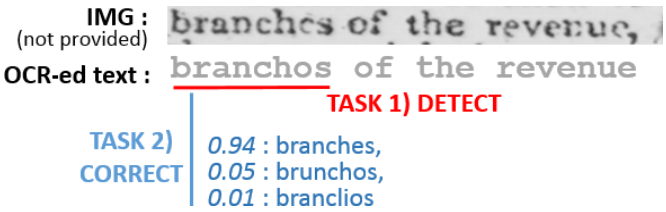

Figure 1. Two tasks: error detection and error correction (from detected errors).

proposed task, even if the dataset is relatively noisy and thus could potentially lead to discouraging scores of error detection (and therefore correction). The two tasks of 1) detection and 2) correction are described below.

1) Task 1 - Detection of OCR errors: Given the raw OCR-ed text (no image provided), the participants were asked to provide the position and also the length of the suspected erroneous tokens. The length information is non-trivial; it is necessary in the case of words that are wrongly split (e.g. OCR-ed separators such as spaces, hyphens or line breaks).

2) Task 2 - Correction of OCR errors: Given the OCR errors in their context (position and length from Task 1), the participants were asked to provide, for each error, a ranked list of replacement candidates (the list may contain only one). The ability to provide multiple candidates enables the evaluation of semi-automated techniques as we will detail later.

B. Dataset

The proposed dataset accounts for 22M OCR-ed characters (754 025 tokens) along with the corresponding ground truth, with an unequally share of 10 European languages (see sources, quantities, mean CER μ and standard deviation CER σ in Table I). The digitized documents come from different collections available, among others, in national libraries or universities. The corresponding GT comes from initiatives such as HIMANIS[4], IMPACT[5], IMPRESSO[6], Open data of National Library of Finland[7], GT4HistOCR [4] and RECEIPT [5].

Degraded documents sometimes result in highly noisy OCR output and thus cannot reasonably be fully aligned with their GT. The unaligned sequences have not been included in the presented statistics (e.g. number of characters and error rates). Error rates vary according to the nature and the state of degradation of the documents. Historical books for example, due to their complex layout and their original fonts have been reported to be especially challenging for OCR engines with up to 50% of wrongly detected characters in some documents.

[4]http://www.himanis.org
[5]https://www.digitisation.eu
[6]https://impresso-project.ch
[7]https://digi.kansalliskirjasto.fi/opendata

Table I
SOURCES, QUANTITIES AND CHARACTER ERROR RATES (μ, σ) INVOLVED IN ALL LANGUAGES OF THE DATASET

Lang.	Source	# file	# character	μ CER	σ CER
BG 1	IMPACT	200	399 636	14.96	12.49
CZ 1	IMPACT	200	274 130	5.79	12.07
DE 1	IMPRESSO	102	575 416	13.54	14.45
DE 2	IMPACT	200	494 328	39.67	16.09
DE 3	Dta19	7 623	10 018 258	24.22	3.26
DE 4	EML	321	509 757	23.95	3.94
DE 5	KA	654	818 711	24.19	3.64
DE 6	ENHG	773	935 014	30.47	3.00
DE 7	RIDGES	415	527 845	24.20	3.63
EN 1	IMPACT	200	243 107	21.28	20.25
ES 1	IMPACT	200	517 723	27.51	17.96
FI 1	NFL open	393	1 960 345	5.67	3.94
FR 1	HIMANIS	1 172	2 792 067	7.14	10.09
FR 2	IMPACT	200	227 039	15.48	13.94
FR 3	RECEIPT	1 968	742 574	9.27	10.91
NL 1	IMPACT	200	764 648	26.84	23.42
PL 1	IMPACT	200	307 144	38.16	18.09
SL 1	IMPACT	200	261 060	10.16	15.83
10	**18**	**15 221**	**22 368 802**	**20.14**	**11.50**

A first part of the dataset (80%) was provided to the participants for training and testing purposes, and the rest (20%) has been used by the organizers for the evaluation. Figure 2 illustrates on a sample file the format provided to the participants. Tokens are simply space-separated sequences, with no restriction on punctuation. Examples of tokens: “i”, “i’am”, “bicycle?”, “qm86-7lk.Qs’g”. Tokens that are considered miss-aligned with the GT are indicated by the “#” signal. The “@” signal is used as padding symbol in the aligned sequences.

C. Evaluation modalities

We proposed two different scenarios to assess the performances of the methods submitted by the participants.

1) Task 1: As it is purely a matter of tokens being truly erroneous or not, this first task is evaluated with usual metrics: recall, precision and F-measure, the latter providing the official ranking of this task. The length information provided by the participants is automatically taken into account by default thanks to the alignment with the GT. For example, the two-token OCR error “we ar” supposed to be “wear” in the GT would penalize (regarding the recall) a solution with only the first token “we”.

2) Task 2: The chosen metric for ranking considers for every token of the text sequence, a weighted sum of the Levenshtein distances between the correction candidates and the corresponding token in the Ground Truth. Consequently, best approaches are those that minimize this distance. Providing multiple candidates enables the evaluation on different modalities reflecting various scenarios. We propose to focus on the two following:

- Fully automated scenario, meant for the comparative evaluation of fully automatic OCR correction tools,

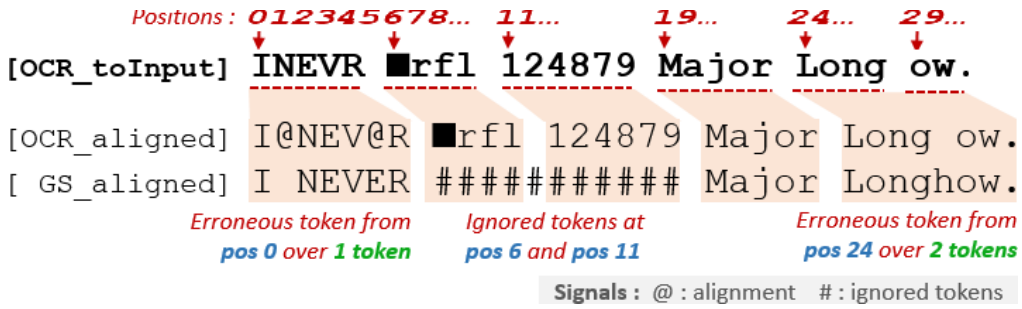

Figure 2. Sample of the training set provided to the participants.

where only the top 1 (highest-weighted word) in each list is taken into account.

- Semi-automated scenario, meant for the comparative evaluation of human-assisted correction tools, where a person typically picks the right correction within a list of system-generated candidate corrections. Thus, it takes into account the list of proposed corrections along with their weights, with an arbitrary limitation to the top 6 candidates.

The evaluation script was available to the participants before and during the competition from the first edition[8]. It computes the metrics presented above over either the training set or the full dataset, with the assumption that the input files are correctly formatted (see Figure 3). The choice of using a structured format *key(pos,lenght) / value(candidates,weights)* rather than asking fully corrected sequences has been motivated by the bias that would have implied any further alignment process between the participants results and the corresponding GT.

Miss-aligned tokens (see "#" signals) are ignored for the computation of the different metrics. Also, given the complexity of dealing with hyphen correction, it has been decided to ignore the tokens containing an hyphen through the evaluation (like for the first edition of this competition). Thus, whether such errors are corrected or not, it does not impact on the final result.

D. Modalities and timeline of the competition

The competition was run in an open mode which mean that the participants must submit their formatted results but not their code. We have relied on the scientific integrity of the authors to follow the rules of the competition. The authors were free to participate, even on sub-parts of the dataset. The training set was made available mid-March 2019. The test set (without the GT), used for evaluating the different methods was made available on the 24^{th} of April with 5 days given for the teams to submit their results.

[8]https://git.univ-lr.fr/gchiro01/icdar2017

III. Submitted methods

In total, 34 teams registered to the competition, for a final number of 5 submissions. Note that we received an extra answer from a Swedish team that applied their word-based method [6] but couldn't get more than 10-20% on the training dataset, so they decided to not submit their results. The following section gives a brief description of the 5 submitted methods. The descriptions were provided by their authors.

RAE1&2 - Team from Centro de Estudios de la RAE[9], Spain

The method implemented using weighted finite-state transducers is an application of the noisy channel model to the optical character recognition error correction of historical texts. The model consists of an error model, a language model and a post-processing module.

Probabilistic character error models are estimated from the training corpus using longest common sub-sequence alignments of tokens which are compiled into weighted finite-state edit transducers. The error models are applied, allowing one edition at most and to address segmentation errors, to tokens, token splits, and concatenations of tokens.

Vocabulary and trigram language models are derived from the Google Books Ngram Corpus for English, French, German and Spanish and n-grams from the Finnish N-gram Corpus, version 1, (FNC1). These n-grams meet the characteristics of quantity and historical amplitude, but contain significant OCR errors. However, they are used under the hypothesis of a good signal to noise ratio.

Using the language model and the lattice of hypotheses generated by the error model, the best path is used to determine the best token sequence. Finally, since historical texts do not follow current standard spellings and typographical conventions, original token's case is restored and some heuristics are applied to punctuation.

[9]Real Academia Española

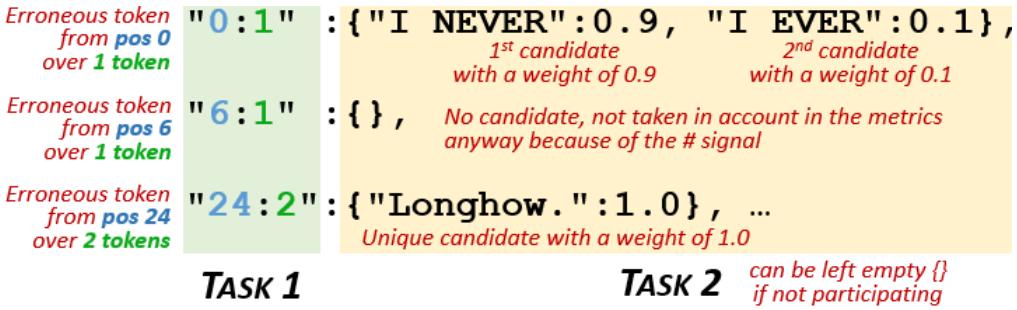

Figure 3. Format expected for submissions to both Task 1 and 2 (JSON).

CCC - Team from Clova AI, NAVER/LINE Corp., South Korea

Our method is Context-based Character Correction (CCC), using the pretrained language model BERT [7] known for its context awareness. Our detection model exploits the pretrained multilingual BERT. The BERT output of each sub-token is then plugged into convolutional layers and fully-connected layers to be classified. The model predictions of sub-tokens are merged into token-level predictions. If more than one sub-token of a token is predicted to be erroneous, then the token is erroneous. Our correction model is a character-level sequence to sequence model with the attention mechanism. The encoder is a bidirectional LSTM and the character embedding is shared between the encoder and the decoder. The encoder input is characters of erroneous tokens and corresponding context information from the BERT, fine-tuned at the detection phase. The decoder generates character-level corrections. The final correction of each erroneous token can be found by using beam search.

CSIITJ - Team from department of CSE[10], IIT[11] Jodhpur, India

We propose a dictionary based solution to the invalid word proposals of an OCR output for English and French Text.

A dictionary based detection scheme is used to mark down OCR tokens not present in the dictionary. We augment the dictionary obtained from training data with external dictionaries[12] [13] to generate our final dictionary of 370 098 unique words.

Given an erroneous OCR token o, we take a dictionary based approach, to generate a set of candidate valid words W, based on edit distance. Ranking amongst the candidate words is based on their likelihood given the error model captured through character *ngram* confusion matrices.

[10]Computer Science & Engineering
[11]Indian Institute of Technology
[12]https://github.com/dwyl/english-words
[13]https://github.com/words/an-array-of-french-words

Formally,

$$w* = arg\max_{w \in W} P(w|o) = arg\max_{w \in W} P(o|w)P(W) \quad (1)$$

where $P(o|w)$ is estimated using n-gram confusion matrices computed from the training set. $P(W)$, the word prior is based on the word frequencies in the training set. Laplace smoothing was used to deal missing values for n-gram confusion matrices and word frequencies.

UvA-seq2seq - Team from Netherlands eScience Center & University of Amsterdam, Netherlands

This approach applied mainly to task 1 which is based on character level seq2seq model (TensorFlow). The model contains multi-layer LSTM as a decoder and just tested on the English partition of the dataset. The fixed-length sequence of characters is fed to the model; we get the same-length characters in return as output. Based on the differences between the input and output, the positions of errors are identified. Furthermore, the model benefits from the pre-trained data of the 2017 edition of this competition to enrich the training dataset. Among various combinations, the model trained on English Monographs was fine-tuned for ICDAR 2019 with the highest score and the result was submitted to this competition.

CLAM - Team from IITB[14] - Monash Research Academy, India

The proposed method relies on the character level attention model with beam search used at decoder's output. We used the open source system OpenNMT[15].

Training: To take care of real word errors as well as non word errors, at network's input we used the characters (with space as char delimiter) from input OCR word o_t along with characters from few words (with $ as word delimiter) on its left: $o_{t-l:t-1}$ and right: $o_{t+1:t+r}$. We also append each input with a language flag and train our model jointly on

[14]Indian Institute of Technology Bombay
[15]http://opennmt.net

all languages. At network's output we used the characters (with space as char delimiter) from Ground Truth word g_t corresponding the input OCR word o_t.

We analyzed the complete dataset and observe that there are at max 10 space related errors where OCR systems introduce fake spaces. To successfully remove such errors during test time we choose $l = 10$, $r = 10$.

Testing: We expect model to jointly learn the language as well as error patterns in OCR output. Since our model majorly abstains from changing the correct words, so for error detection the word that is changed by model is considered as erroneous, else correct. We also use edit distance to find the length of erroneous tokens. The changed word/words is/are considered as the suggestion for the error correction task.

IV. Results and discussion

Table II and III detail the results for each of the 10 languages on task 1 and 2 respectively. The metrics (F-measure and % of improvement) are the average for each of the 10 languages. The percentage of improvement is measured by comparing the weighted sum of the Levenshtein distances between no replacement, top 1 and top 5 replacement (see Section II-C2). The "x" symbol corresponds to no exploitable results (e.g. participation to some languages only). In the context of Task 2, the "=" symbol indicates an equal result for both the automatic and the semi-automatic approaches, which in most cases indicates that participants have provided only one candidate per correction. Note that the UVA method provided overlapping error positions which may have impacted the final score.

The Clova AI team is the best performer on Task 1 with their CCC method achieving the best average F-measure on every corpus language, from 0.67 on French up to 0.95 average F-measure on German, who the latter initially has an important CER (see Table I). Best scores are obtained for German for every team which participated in this language. It is also the language with the highest amount of data (413 703 tokens).

The Clova AI team is the best performer on task 2 as well with their CCC method achieving the best average % of improvement in 8 of 10 languages, from 6% for Spanish to 24% for German. On this task, the IITB team achieved the best performance for Finnish with their CLAM method, 44% improvement in average. Also, the team from the Centro de Estudio de la RAE performed best on French with 26% average improvement with their RAE1 method. For those who provided multiple correction candidates, we sometimes observed slightly better results for the automatic mode than for the semi-automatic mode, but also big drops which shows the limited interest of this latest type of evaluation. The same observation has been done during the 2017 edition of this competition.

Some participants have rightly pointed out some inaccuracies in the GT such as missing or incorrect corrections. The dataset, given its important size and its nature (manually annotated, OCR/GT automatically aligned) is obviously imperfect. Those inaccuracies, although rare, can still trouble both the training (by misleading the final model) and the evaluation phase (by wrongly considering a right correction).

The dataset will be made available on the competition website[16] and submitted to http://tc11.cvc.uab.es after ICDAR 2019 conference.

V. Conclusion

This paper describes the second ICDAR competition on post-OCR text correction. The challenges consisted of two tasks similar to the first edition: 1) error detection and 2) error correction, with the difference of the second task being dependent on the first task. For this edition, we proposed a much larger and multilingual dataset of 22M OCR-ed symbols along with an aligned ground truth. The data came from newspapers, historical books and shopping receipts, covering 10 European languages. This competition demonstrated, through formatted results provided by the participants, the performance of their systems exposed to this specific dataset and to the metrics described in this paper.

Concerning the first task (error detection), the Clova AI team performed the best with a maximum error detection of 95% for German. Its authors proposed a context-based character correction (CCC) method based on the pretrained multi-language model BERT, plugged into convolutional and fully-connected layers for the final classification.

Concerning the second task, (error correction), the Clova AI team also performed the best on 8 of 10 languages using context information from BERT, fine-tuned at the detection phase. IITB team achieved the best performance for Finnish using a character level attention model (CLAM). The team from the Centro de Estudio de la RAE performed best for French.

The submissions that resulted in low scores in the context of this competition could of course work better on different conditions (datasets, languages, formats and metrics).

Comparing to the previous edition of this competition, we see an improvement of the overall error detection performance which has been improved by most of the methods. We also observe that the presented methods generalize well over the proposed multilingual dataset.

In perspective, it would be interesting to test a less complex format for the evaluation with full sequences provided as an input instead of a list of positions/corrections. This would require a posterior automatic alignment phase (e.g. [8]) with its pros (easier for participants), its cons (difficult support of multiple correction candidates) and its risks (miss-alignment). In a nutshell, this competition has illustrated the difficulty of the proposed tasks on a multilingual dataset. It highlights that the amount of data

[16]https://l3i.univ-larochelle.fr/ICDAR2019PostOCR

Table II
SUMMARIZED RESULTS FOR TASK 1 FOR EACH PROPOSED METHOD

	Task 1 (F-measure)									
Language	BG	CZ	DE	EN	ES	FI	FR	NL	PL	SL
Nb tokens	31 164	12 569	413 703	11 443	23 328	54 606	147 432	35 170	13 928	16 682
CCC	**0.77**	**0.70**	**0.95**	**0.67**	**0.69**	**0.84**	**0.67**	**0.71**	**0.82**	**0.69**
CLAM	0.68	0.41	0.93	0.45	0.56	0.51	0.45	0.61	0.72	0.54
CSIITJ	x	x	x	0.45	x	x	0.42	x	x	x
RAE1	x	x	0.90	0.53	0.62	0.44	0.42	x	x	x
RAE2	x	x	0.89	0.57	0.60	0.46	0.45	x	x	x
UVA	x	x	x	0.47	x	x	x	x	x	x

Table III
SUMMARIZED RESULTS FOR TASK 2 FOR EACH PROPOSED METHOD

	Task 2 (% improvement) Auto (top1) / Semi (weighted mean on top5)									
Language	BG	CZ	DE	EN	ES	FI	FR	NL	PL	SL
CCC	**9 / 8**	**6 / =**	**24 / =**	**11 / =**	**11 / 6**	8 / =	5 / =	**12 / 10**	**17 / 16**	**14 / 12**
CLAM	-2 / -3	-1 / =	-7 / =	0.4 / =	-1 / -5	**44 / =**	4 / =	-3 / =	-2 / =	0 / -1
CSIITJ	x	x	x	2 / 1	x	x	x	x	x	x
RAE1	x	x	15 / =	9 / =	7 / =	7 / =	**26 / =**	x	x	x
RAE2	x	x	14 / =	6 / =	7 / =	6 / =	20 / =	x	x	x
UVA	x	x	x	0 / =	x	x	x	x	x	x

seems to be more determinant than the language itself for the tested methods. However, this competition also highlighted the strong interest of the community for this topic, which is of primary interest for enhancing the access to patrimonial content from digital libraries.

ACKNOWLEDGMENT

This work has been supported by the European Union's Horizon 2020 research and innovation programme under grant 770299 (NewsEye). We are grateful to all libraries and institutions that allowed the use of their material for this competition.

REFERENCES

[1] M. C. Traub and al., "Impact analysis of OCR quality on research tasks in digital archives," in *International Conf. on Theory and Practice of Digital Libraries*. Springer, 2015, pp. 252–263.

[2] G. Chiron, A. Doucet, M. Coustaty, J.-P. Moreux, and M. Visani, "Impact of OCR errors on the use of digital libraries - towards a better access to information," in *JCDL'17, ACM/IEEE-CS Joint Conference on Digital Libraries, June 2017, Toronto, Ontario, Canada*, 2017/06 2017.

[3] G. Chiron, A. Doucet, M. Coustaty, and J.-P. Moreux, "ICDAR2017 Competition on Post-OCR Text Correction," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 01, Nov. 2018, pp. 1423–1428. [Online]. Available: doi.ieeecomputersociety.org/10.1109/ICDAR.2017.232

[4] U. Springmann, C. Reul, S. Dipper, and J. Baiter, "GT4HistOCR: Ground Truth for training OCR engines on historical documents in German Fraktur and Early Modern Latin," Aug. 2018. [Online]. Available: https://doi.org/10.5281/zenodo.1344132

[5] C. Artaud, N. Sidère, A. Doucet, J. Ogier, and V. P. D. Yooz, "Find it! fraud detection contest report," in *2018 24th International Conference on Pattern Recognition (ICPR)*, Aug 2018, pp. 13–18.

[6] H. Hammarström, S. M. Virk, and M. 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*, ser. DATeCH2017. New York, NY, USA: ACM, 2017, pp. 71–75. [Online]. Available: http://doi.acm.org/10.1145/3078081.3078107

[7] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[8] I. Z. Yalniz and R. Manmatha, "A fast alignment scheme for automatic ocr evaluation of books."