

Correction post-OCR en 3 étapes : détection, correction, vérification

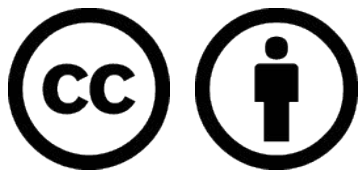
Guillaume Thomas, Joseph Chazalon, Edwin Carlinet
Laboratoire de Recherche de l'EPITA (LRE)

SIFED — 9 juin 2023



(Ré)utiliser cette présentation

Licence **Creative Common BY 4.0**



N'hésitez pas à **réutiliser** ces contenus et nous **citer**.

*Sauf citation ou mention contraire,
nous sommes les auteurs des contenus.*

Consultez le diaporama pendant et après la présentation

<https://bit.ly/sifed23-postocr>



Sondage

“Post-OCR is no longer a topical issue.”

Selon vous, c'est :

- Vrai
- Faux
- Ça dépend...
- C'est quoi ?

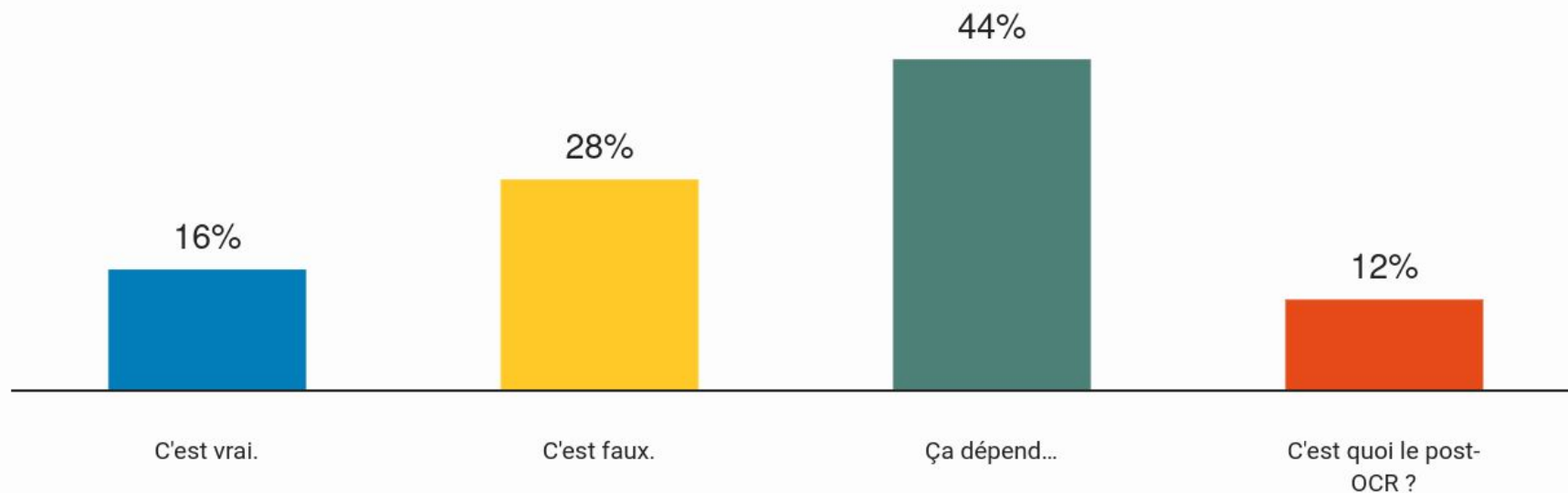
 <https://partici.fi/76203588>





Les 25 réponses

"Post-OCR is no longer a topical issue."





Post-OCR

Ravrio et comp., *fabr. de bronzes et curiosités*, r. Richelieu, 93; la fabrique rue Montmartre, 161.

Some OCR system

Ravrio **et** comp., **fubr.** de bronzes et curiosité**s**, r. Richelieu, 93; la fabrique rue Mo**at**martre, **I**61.

Post-OCR correction

Ravrio **est** comp.**.** **fubr.** de bronzes et curiosité**s**, r. Richelieu, 93; la fabrique rue Mo**nt**martre, **#**161.

- Modèle **optique** $\rightarrow P(\text{texte}|\text{image})$
- Modèle **langage** $\rightarrow P(\text{texte})$

- Insertion / délétion / substitution de fragments de chaîne
- Amélioration CER ou autre métrique
- Spécifique couple (OCR, corpus) ? (modèle optique/bruit, modèle langage)

Motivation



Faible potentiel d'innov. en post-OCR dans le cas grl...

Pour les **documents modernes, majoritairement textuels** (rapports, contrats...)

- ~~Les contenus sont nativement numériques~~
- Les **OCRs modernes** sont **très puissants** (moins pour le **manuscrit**)
- Les **LLM** permettent **d'excellentes corrections** (mais risque de **reformulation**)

Pour les **documents plus anciens**, ou avec **syntaxe/lexique/glyphes spécifiques**

- Arbitrage “**réentraînement OCR**” vs “**entraînement post-OCR**” ?
- Opportunité de capturer facilement un **modèle de langage dur** (?) à intégrer dans un OCR moderne ?
- Application de **niche** encore possible ? **Fine-tuner un OCR** est-il vraiment dur ?

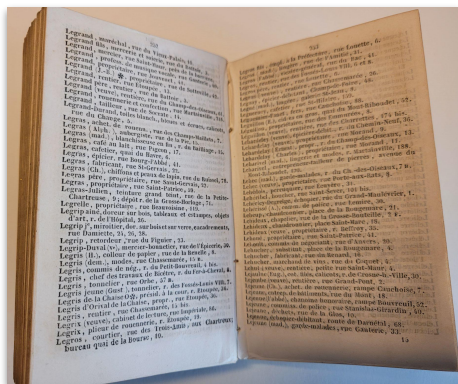


...mais dans notre cas...

Données sources :

Annuaires du commerce du 19e siècle

Une centaine d'annuaires, des milliers de pages par annuaire, **environ 10M entrées au total.**



Données cibles :

Liste des entrées de chaque annuaire avec raison sociale, activité, adresse (voie, no).

Lemonnyer, plomberie, r. de Bondy, 86, et r. Bouchardon, 1.



Lemonnyer, plomberie, r. de Bondy, 86, et r. Bouchardon, 1.



PERSONNE	ACTIVITÉ	RUE	NUMÉRO	RUE	NUMÉRO
Lemonnyer,	plomberie,	r. de Bondy,	86,	et r. Bouchardon,	1.



...cela nous semblait pertinent.

Spécificités syntaxiques :

- Entrées courtes faciles à séparer
- Syntaxe très régulière (quelques variantes)
- Beaucoup de nombres à des positions connues, lexique limité pour les rues...

Et déjà un niveau d'OCR très acceptable avec **PERO OCR** (Univ. Brno, Rép. tchèque).

Données :

"A Dataset of French Trade Directories from the 19th Century"

Accessible librement sur Zenodo [DOI 10.5281/zenodo.6394464](https://doi.org/10.5281/zenodo.6394464)

Bottin 1820

Dufort, bottier, Palais-R., gal. vitrée, 215.
295

Bottin 1827

Baleste, chef aux domaines, S.-Georges, 17.

Bottin 1837

Cattois, pharmac., Bretagne, 46.

Bottin 1854

Fontaine, draperies, Neuve-des-Petits-Champs, 2.

Cambon Almgene 1841

Aron Javal (L.) art. de Paris, r. des Bourdonnais, 17.

Deflandre 1828

DEVILLERS, r. Croix-des-Pet.-Champs, 25.
Cordonn.

Deflandre 1829

Huguenin, épïc., r. de Valois, 8, Pal.-Royal

Didot 1851

Viéville, fab. de boutons, Aumaire, 48, et place
St-Nicolas-des-Champs, 2.

État de l'art



État de l'art (accéléré)

- **Expression régulières** → ✗
- **Lexiques** → ✗
- **Correcteurs d'orthographe** → ✗
- **Modèle de bruit OCR** → ✗

$$\hat{S} = \underset{S}{\operatorname{argmax}} P(S|O) \quad \begin{array}{l} S: \text{chaîne réelle,} \\ O: \text{chaîne OCR} \end{array}$$

$$P(S|O) = \underbrace{P(O|S)}_{\text{Modèle de génération bruit}} \times \underbrace{P(S)}_{\text{Modèle de langage}} / P(O) \quad \leftarrow \text{Probabilité du bruit (négligée ?)}$$

- **Modèle sequence-to-sequence (seq2seq)** → ✓
Estimation directe de $P(S|O)$
(mélange des modèles de bruit et de langage)

Lectures conseillées :

État de l'art :

• T. T. H. Nguyen, A. Jatowt, M. Coustaty, and A. Doucet, "Survey of Post-OCR Processing Approaches," *ACM Comput. Surv.*, vol. 54, no. 6, pp. 1–37, Jul. 2021, doi: 10.1145/3453476.

Modèle de bruit :

• O. Kolak and P. Resnik, "OCR error correction using a noisy channel model," in *Proc. of the second intl conf. on Human Language Technology Research*, San Diego, California: Association for Computational Linguistics, 2002, pp. 257–262. doi: 10.3115/1289189.1289208.

- Estimation sur base d'entraînement.
- Calibration difficile (impossible ?) pour séquences longues ou inconnues.
- Travail au niveau mot inévitable ?



Focus sur les modèles seq2seq

Approches avec les meilleurs résultats récents →

Quelle **unité de travail** ?

- **Mot** ? → délicat, inadapté symboles et nombres
- **Token** ? → et pour les motifs de bruit rares ? replis sur mode caractère ?
- **Caractère** ? → contexte limité, mais approche avec le + de succès

Quelle **architecture** ?

- Présence **décodeur** nécessaire dans tous les cas pour permettre $|\text{input}| \neq |\text{output}|$
- RNN : LSTM...
- Attention ?

ICDAR 2017

• 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), Nov. 2017, pp. 1423–1428. doi: 10.1109/ICDAR.2017.232.

ICDAR 2019

• C. Rigaud, A. Doucet, M. Coustaty, and J.-P. Moreux, "ICDAR 2019 Competition on Post-OCR Text Correction," in 2019 International Conference on Document Analysis and Recognition (ICDAR), Sep. 2019, pp. 1588–1593. doi: 10.1109/ICDAR.2019.00255.

2019 winner: Clova AI, NAVER/LINE Corp., South Korea approach (publication?), reimplemented in

• T. T. H. Nguyen, A. Jatowt, N.-V. Nguyen, M. Coustaty, and A. Doucet, "Neural Machine Translation with BERT for Post-OCR Error Detection and Correction," in Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020, in JCDL '20. New York, NY, USA: Association for Computing Machinery, août 2020, pp. 333–336. doi: 10.1145/3383583.3398605.



Limites des approches seq2seq char-level (1/2)

Problème 1 : **Hyperactivité**

Solution 1 : **Étage de détection**
(niveau token/mot jusqu'à présent)

↳ Problème 2 : Semble imposer une **correction au niveau mot/token qui limite le contexte**

↳ Solution 2: **Injection d'informations supplémentaires au niveau local** –
Approche “CCC” Clova AI, NAVER/LINE Corp., South Korea, non publiée : contexte fourni par le détecteur BERT

- “Our evaluation suggests that future work on NMT for OCR post-correction should focus on improving error detection.”: C. Amrhein and S. Clematide, “Supervised OCR Error Detection and Correction Using Statistical and Neural Machine Translation Methods,” *Journal for Language Technology and Computational Linguistics*, vol. 33, no. 1, Art. no. 1, Jul. 2018, doi: 10.21248/jlcl.33.2018.218.
- R. Schaefer and C. Neudecker, “A **Two-Step Approach** for Automatic OCR Post-Correction,” in *Proc. of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, Online: International Committee on Computational Linguistics, Dec. 2020, pp. 52–57. <https://aclanthology.org/2020.latechclfl-1.6>



Limites des approches seq2seq char-level (2/2)

Solution 1 bis : **Limiter le nombre de modifications autorisées** : seuil sur la distance d'édition, ou sur la différence de longueur avant/après correction.

Word/char-based NMT + control based on absolute edit length

- K. Mokhtar, S. S. Bukhari, and A. Dengel, "OCR Error Correction: State-of-the-Art vs an NMT-based Approach," in 2018 13th IAPR International Workshop on Document Analysis Systems (DAS), Apr. 2018, pp. 429–434. doi: 10.1109/DAS.2018.63.

Length difference

- T. T. H. Nguyen, A. Jatowt, N.-V. Nguyen, M. Coustaty, and A. Doucet, "Neural Machine Translation with BERT for Post-OCR Error Detection and Correction," in Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020, in JCDL '20. New York, NY, USA: Association for Computing Machinery, août 2020, pp. 333–336. doi: 10.1145/3383583.3398605.



Bilan de l'état de l'art

Pipeline à 3 étages simple, extensible et prometteur
⇒ nous l'avons adapté à nos données

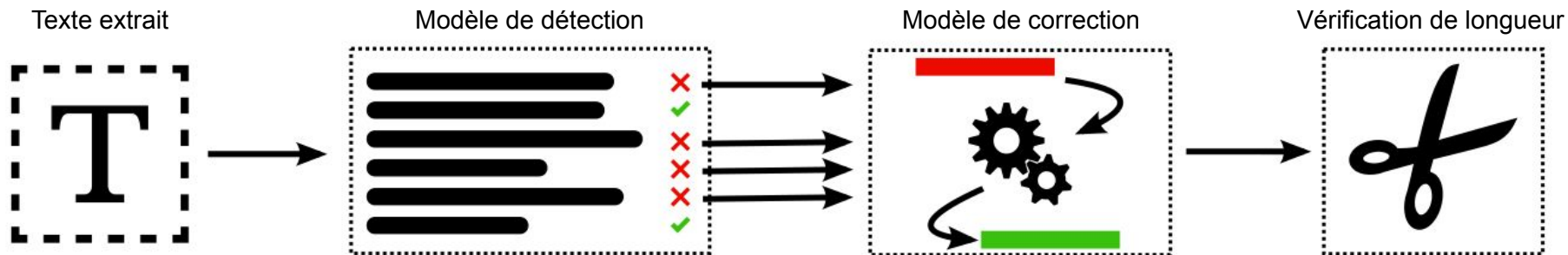
```
input: ocr_str, output: result

result = ocr_str
if detect_errors(ocr_str):
    result = correc_errors(ocr_str)
    if not verify_corrections(ocr_str, result):
        result = ocr_str # restore original
```

Expériences



Notre approche : vue d'ensemble



Différences par rapport aux approches **existantes** :

1. **Détecteur** qui classe des **séquences entières**
2. **Correcteur** qui prend en entrée une **séquence entière**
3. **Vérifieur** qui se base sur des **indicateurs** calculés entre les **séquences entières**

Avantages :

- **Très simple** à implémenter
- **Contexte plus large** pour le correcteur

Inconvénients :

- **Détection moins fine** (mais *chunking* possible)
- Risques d'**hallucinations** et de **résumés** plus importants au niveau du correcteur



Jeux de données, métriques, baseline

Jeu de données FTD

- +8 000 entrées avec transcription manuelle
- train/test split 90/10

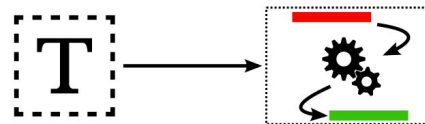
Evaluation CER (micro avg), CER par entité nommée

Performance brute de l'OCR (PERO) : **Baseline CER = 3,76 %**

Aucun post-processing avec “règles du Français”, réduction possible mais inutile pour notre objectif (impact limité sur extraction entités nommées)



Correcteur



Implémentation : encodeur décodeur de base OpenNMT (2×LSTM + attn)

	CER
Baseline	3,76 %
Correcteur	14,31 %

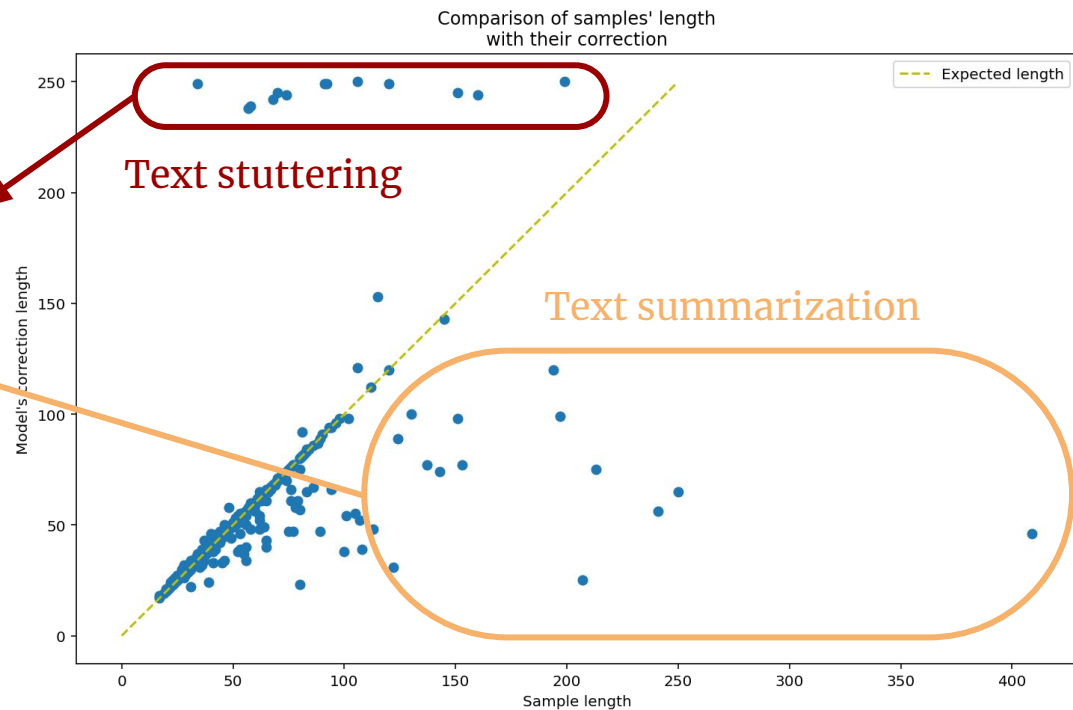


Hallucinations et résumés, besoin d'un **vérifieur**



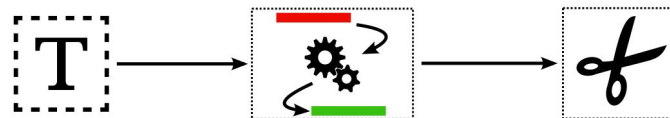
Hallucinations et résumés

Texte OCR	Texte corrigé
Pelletier, V.-Marché-St-Martin, 5.	Petietier, V.-Marché- <u>St-Marché-St-Marché-S</u> <u>t-Marché-St</u> -[...]
Jacquin aîné et Cie, grainiers, fleuristes et pépiniéristes ; commerce en gros et en détail de graines potagères et fourragères, graines de fleurs et d'arbres, oignons à fleurs et plantes et arbres de tous genres ; au Bon Jardinier, quai de la Mégisserie, 14. Cultures expérimentales, au Grand-Charonne, route de Bagnolet, 20, près Paris; médailles d'or et d'argent des diverses sociétés horticoles.	Jacquin aîné et Cie, gramiers, Cleuristes, 14.






Vérifieur



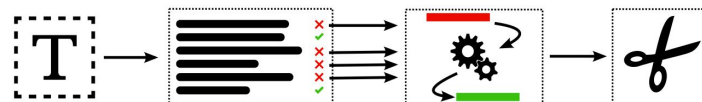
Implémentation : seuils de différence relative de longueur

	CER
Baseline	3,76 %
Correcteur	14,31 %
Correcteur + Vérifieur	3,98 %

 Hyperactivité du modèle (alors que 50 % des entrées sont correctes), ajout d'un détecteur



Détecteur

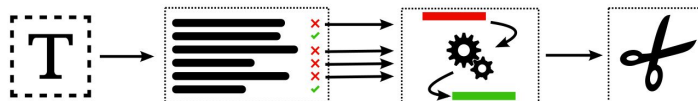


Implémentation : CamemBERTForSequenceClassification 🙌

	CER
Baseline	3,76 %
Correcteur	14,31 %
Correcteur + Vérifieur	3,98 %
Détecteur + Correcteur + Vérifieur	3,51 %



Détecteur



Implémentation : CamemBERTForSequenceClassification 🤖

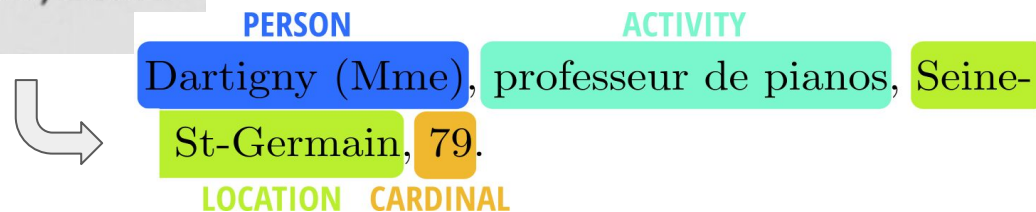
	CER
Baseline	3,76 %
Correcteur	14,31 %
Correcteur + Vérifieur	3,98 %
Détecteur + Correcteur + Vérifieur	3,51 %
Détecteur + Correcteur	7,87 %

⚠ Accuracy détecteur : ~67 %



Evaluation au niveau des entités nommées

Dartigny (Mme), professeur de pianos, Seine-
St-Germain, 79.



CER (micro)

	OCR	Corrigé
<i>Global</i>	3,76 %	3,51 %
Person	2,84 %	2,02 %
Activity	2,52 %	2,53 %
Location	3,47 %	2,40 %
Cardinal	7,23 %	4,33 %



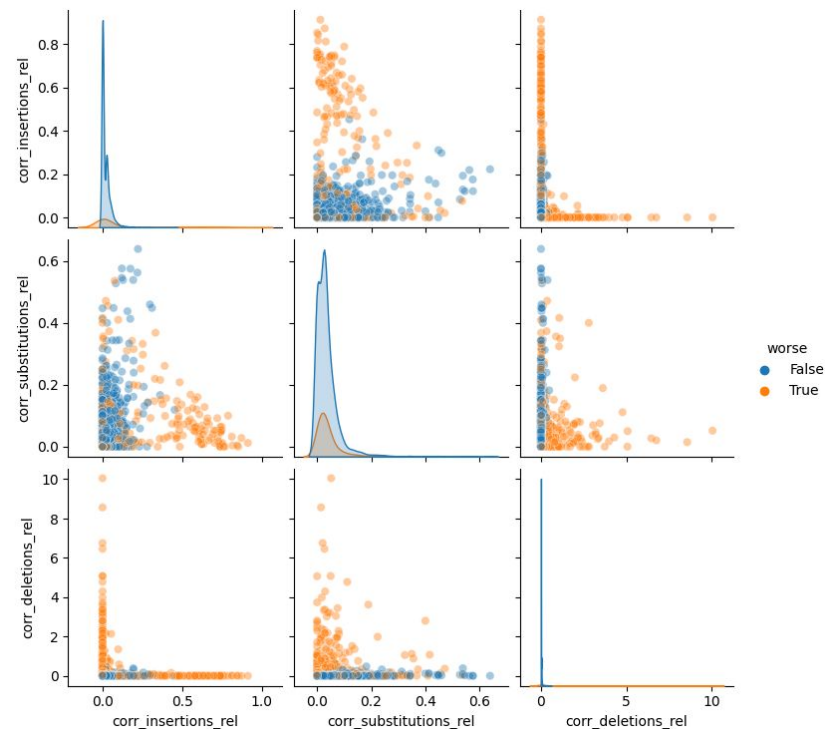
Vérifieur v2

La différence de longueur est trop simpliste :

1 ins + 1 del \Rightarrow même longueur

Mesurer le taux ins/del (et opt. sub)
fourni de bien meilleurs indicateurs.

(en cours d'intégration)





Pipeline complète, API et package Python

👉 <https://github.com/soduco/postocr-3stages>

```
$ pip install postocr-3stages # (ou presque)
```

```
from postocr_3stages import Pipeline

train_dataset, test_dataset= # FIXME load your dataset

postocr = Pipeline()
postocr.fit(train_dataset["OCR"], train_dataset["Gold"])
# ... wait
print("Post-OCR CER:")
print(postocr.score(train_dataset["OCR"], train_dataset["Gold"]))

print("Sample correction")
print(postocr.predict(["Sample string to correct."]))
```

 **Attention ! Travail en cours !** 

En résumé



Post-OCR en 3 étapes

Encore des gains sur certains cas niches

Implémentation simple d'une variante en 3 étapes

Package Python ouvert, facile à utiliser et améliorer

Work in progress, nombreuses pistes d'amélioration :

Comparer avec les quelques solutions ouvertes existantes, tester sur les données ICDAR 2019, tester variantes architecture, tester pre-training self-supervised, comparer avec fine-tuning OCR...

Réserve

Ravrio et comp., *fabr. de bronzes et curiosités*, r. Richelieu, 93; la fabrique
rue Montmartre, 161.

