Détection des Erreurs d'OCR sur base de Confiance

Arthur Hemmer[1,2], Nicola Bartolo[1], Mickaël Coustaty[2] and Jean-Marc Ogier[2]

[1]*Shift Technology, Paris, France*
[2]*L3i, La Rochelle, France*

Abstract
Malgré les avancées en Reconnaissance Optique de Caractères (OCR), des erreurs affectent les processus en aval. Nous explorons l'utilisation des confiances d'OCR pour améliorer les méthodes de détection d'erreurs type BERT en intégrant les scores de confiance. Nos expériences montrent que les scores de confiance améliorent la détection d'erreurs, variant selon l'étalonnage du système d'OCR.

Keywords
Post-OCR, Détection d'Erreurs, Confiance

1. Introduction

La Reconnaissance Optique de Caractères (OCR) s'est considérablement améliorée sur les documents scannés grâce à l'apprentissage profond et les avancées en vision par ordinateur. Néanmoins, les erreurs d'OCR persistent, affectant diverses tâches de Traitement du Langage Naturel (NLP) [1, 2, 3, 4, 5, 6, 7], y compris la recherche d'information. Dans certains cas, ces erreurs d'OCR sont des erreurs « véritables », signifiant qu'un humain ne serait également pas capable de transcrire à partir de la vision seule. Les méthodes d'OCR, cependant, essaieront toujours de fournir une transcription du texte illisible, ce qui aboutit souvent à du charabia. Pour fournir une transcription correcte, les humains utilisent l'incertitude optique et l'information sémantique contextuelle pour inférer une correction possible.

Pour apporter des réponses à ces limitations des méthodes d'OCR, des méthodes de correction post-OCR ont émergé, offrant une couche de post-traitement pour corriger les erreurs [8, 9, 10, 11, 12]. Ces méthodes utilisent principalement des modèles de langues entraînés sur des corpus pour pouvoir détecter des incohérences sémantiques.

Les scores de confiance des méthodes d'OCR pourrait affiner d'avantage la détection des erreurs. Ce travail investigue l'exploitation de ces scores de confiance de divers méthodes d'OCR pour améliorer la correction des erreurs post-OCR.

2. Données

Nous testons divers systèmes d'OCR, tant open-source que commerciaux, face à des transcriptions de référence issues de trois bases de données publiques et une privée. Les bases de données publiques utilisées pour notre analyse sont CORD [13], FUNSD [14], SROIE [15], qui regroupent

arthur.hemmer@univ-lr.fr (A. Hemmer)

des documents administratifs scannés, en plus d'une base de données privée similaire, et non connue des méthodes d'OCR. Nous avons sélectionné plusieurs méthodes d'OCR tels que Microsoft Document Intelligence, Amazon Webservices (AWS) Textract, Google OCR, DocTR [16], EasyOCR et PaddleOCR [17].

Pour pallier les difficultés d'alignement résultant des sorties hétérogènes émanant de divers méthodes d'OCR, nous utilisons une méthode en deux étapes pour aligner les boîtes englobantes produites par les méthodes d'OCR et celles de la vérité terrain. Dans la première étape, nous établissons des correspondances entre chaque boîte de la vérité terrain et celle de l'OCR qui se superpose le plus, et réciproquement. Par la suite, nous traitons les composantes connexes en tant qu'ensembles de boîtes correctement alignées. Cette méthode permet de calculer le Taux d'Erreur de Caractères (CER), le Taux d'Erreur de Boîtes (BER) et l'Erreur de Calibration Attendue (ECE) [18] pour chaque système d'OCR.

Les résultats montrent que les méthodes d'OCR commerciaux surpassent généralement les options open-source, avec des différences de performances liées au traitement de la ponctuation et des caractères spéciaux. En général, nous constatons un taux d'erreur de caractères (CER) de 1 à 3% sur les OCR commerciaux et de 3 à 30% sur les OCR open-source, DocTR étant celui qui se rapproche le plus des OCR commerciaux.

3. ConfBERT

Pour améliorer la détection d'erreurs post-OCR, nous intégrons aux embeddings de BERT les scores de confiance d'OCR, en nous appuyant sur des travaux antérieurs basés sur BERT [19, 20] tout en apportant des modifications minimales pour limiter la nécessité de réentraîner le modèle. Nous présentons ConfBERT, un modèle qui enrichit les embeddings de BERT avec les scores de confiance d'OCR, de manière similaire au codage positionnel dans les architectures de transformers (voir Eq. 1). Considérant un token t_i, la fonction d'embedding de BERT définie par $\text{Emb} : t \to \mathbb{R}^d$, et la probabilité de confiance de l'OCR p_{ocr} pour ce token, nous définissons l'embedding pondéré par la confiance, e_i^c, selon la formule suivante :

$$e_i^c = (1 - \alpha) \cdot \text{Emb}(t_i) + \alpha \cdot (1 - p_{ocr}(t_i)). \tag{1}$$

Dans notre architecture, les scores de confiance sont fusionnés avec les embeddings, avec un paramètre entraînable, α, qui équilibre l'influence de la confiance d'OCR par rapport à l'embedding du token.

Nos expériences comparent ConfBERT à une baseline utilisant uniquement la confiance OCR, et une autre baseline avec un BERT sans les confiances. Nous observons que l'intégration de la confiance d'OCR maintient ou améliore généralement les scores F_1, avec certaines exceptions expliquées par une Erreur de Calibration Attendue (ECE) élevée ou la granularité grossière des boites de certains méthodes d'OCR. Alors que nos résultats confirment le potentiel d'utilisation de la confiance OCR pour la détection d'erreurs, les améliorations ne sont pas universellement significatives, et dans certains cas, la baseline plus simple basée sur la confiance OCR se montre comparativement meilleure ou équivalente. En moyenne, l'intégration des confiances dans BERT augmente la F_1 de 6 points, avec un écart-type 4.5. L'augmentation sur les OCRs commerciaux est plus élevé (9.3 points en moyenne) comparé aux méthodes d'OCR open-source (2.8).

References

[1] L. L. de Oliveira, D. S. Vargas, A. M. A. Alexandre, F. C. Cordeiro, D. d. S. M. Gomes, M. d. C. Rodrigues, R. K. Romeu, V. P. Moreira, Evaluating and mitigating the impact of ocr errors on information retrieval, International Journal on Digital Libraries 24 (2023) 45–62.

[2] M. Cuper, C. van Dongen, T. Koster, Unraveling confidence: Examining confidence scores as proxy for ocr quality, in: International Conference on Document Analysis and Recognition, Springer, 2023, pp. 104–120.

[3] D. Fleischhacker, W. Goederle, R. Kern, Improving ocr quality in 19th century historical documents using a combined machine learning based approach, arXiv preprint arXiv:2401.07787 (2024).

[4] E. Boros, N. K. Nguyen, G. Lejeune, A. Doucet, Assessing the impact of ocr noise on multilingual event detection over digitised documents, International Journal on Digital Libraries 23 (2022) 241–266.

[5] A. Hamdi, A. Jean-Caurant, N. Sidère, M. Coustaty, A. Doucet, Assessing and minimizing the impact of ocr quality on named entity recognition, in: Digital Libraries for Open Knowledge: 24th International Conference on Theory and Practice of Digital Libraries, TPDL 2020, Lyon, France, August 25–27, 2020, Proceedings 24, Springer, 2020, pp. 87–101.

[6] A. Hamdi, E. L. Pontes, N. Sidere, M. Coustaty, A. Doucet, In-depth analysis of the impact of ocr errors on named entity recognition and linking, Natural Language Engineering 29 (2023) 425–448.

[7] K. Todorov, G. Colavizza, An assessment of the impact of ocr noise on language models, arXiv preprint arXiv:2202.00470 (2022).

[8] A. Jatowt, M. Coustaty, N.-V. Nguyen, A. Doucet, et al., Post-ocr error detection by generating plausible candidates, in: 2019 International Conference on Document Analysis and Recognition (ICDAR), IEEE, 2019, pp. 876–881.

[9] J. A. Ramirez-Orta, E. Xamena, A. Maguitman, E. Milios, A. J. Soto, Post-ocr document correction with large ensembles of character sequence-to-sequence models, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, 2022, pp. 11192–11199.

[10] C. Rigaud, A. Doucet, M. Coustaty, J.-P. Moreux, Icdar 2019 competition on post-ocr text correction, in: 2019 international conference on document analysis and recognition (ICDAR), IEEE, 2019, pp. 1588–1593.

[11] A. İ. Topçu, B. U. Töreyin, Neural machine translation approaches for post-ocr text processing, in: 2022 30th Signal Processing and Communications Applications Conference (SIU), IEEE, 2022, pp. 1–4.

[12] N. Yasin, I. Siddiqi, M. Moetesum, S. A. Rauf, Transformer-based neural machine translation for post-ocr error correction in cursive text, in: International Conference on Document Analysis and Recognition, Springer, 2023, pp. 80–93.

[13] S. Park, S. Shin, B. Lee, J. Lee, J. Surh, M. Seo, H. Lee, Cord: a consolidated receipt dataset for post-ocr parsing, in: Workshop on Document Intelligence at NeurIPS 2019, 2019.

[14] G. Jaume, H. K. Ekenel, J.-P. Thiran, Funsd: A dataset for form understanding in noisy scanned documents, in: 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW), volume 2, IEEE, 2019, pp. 1–6.

[15] Z. Huang, K. Chen, J. He, X. Bai, D. Karatzas, S. Lu, C. Jawahar, Icdar2019 competition

on scanned receipt ocr and information extraction, in: 2019 International Conference on Document Analysis and Recognition (ICDAR), IEEE, 2019, pp. 1516–1520.
[16] Mindee, doctr: Document text recognition, https://github.com/mindee/doctr, 2021.
[17] Y. Du, C. Li, R. Guo, X. Yin, W. Liu, J. Zhou, Y. Bai, Z. Yu, Y. Yang, Q. Dang, et al., Pp-ocr: A practical ultra lightweight ocr system, arXiv preprint arXiv:2009.09941 (2020).
[18] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, On calibration of modern neural networks, in: International conference on machine learning, PMLR, 2017, pp. 1321–1330.
[19] M. Hajiali, J. R. Fonseca Cacho, K. Taghva, Generating correction candidates for ocr errors using bert language model and fasttext subword embeddings, in: Intelligent Computing: Proceedings of the 2021 Computing Conference, Volume 1, Springer, 2022, pp. 1045–1053.
[20] T. T. H. Nguyen, A. Jatowt, N.-V. Nguyen, M. Coustaty, 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, 2020, pp. 333–336.