Repenser l'évaluation de l'OCR pour l'extraction d'informations dans les documents commerciaux

Ngoc Nhi Nguyen[1] Ahmed Hamdi[2] Antoine Doucet[3,4] Adam Jatowt[5]
Mickaël Coustaty[3]
(1) Université des science et technologies de Hanoi, Vietnam
(2) IRIT, Université de Toulouse, France
(3) L3i, La Rochelle université, France
(4) University de Ljubljana, Slovénie
(5) Université d'Innsbruck, Autriche
nhinn.22bi13351@usth.edu.vn, ahmed.hamdi@irit.fr,
antoine.doucet@univ-lr.fr, mickael.coustaty@univ-lr.fr,
adam.jatowt@uibk.ac.at

RÉSUMÉ

L'usage croissant de l'OCR pour l'analyse de documents commerciaux numérisés favorise l'automatisation, mais introduit de nouveaux défis pour l'extraction d'informations. Malgré de bonnes performances en conditions contrôlées, les systèmes OCR restent imparfaits, et les métriques classiques (CER/WER) ne mesurent pas pleinement l'impact des erreurs sur les tâches en aval, notamment lorsque des tokens critiques pour les tâches concernées sont affectés. Dans cet article, nous étudions la relation entre qualité OCR et performance d'extraction à travers un cadre d'évaluation reposant sur une injection sélective d'erreurs réalistes. Les résultats montrent que les performances d'extraction sont très sensibles aux erreurs ciblant les tokens pertinents, même à faible taux de bruit, et que les métriques standards reflètent mal cet impact. Cela souligne le besoin de protocoles adaptés et de chaînes de traitement robustes pour les conditions réelles.

ABSTRACT

Rethinking OCR Evaluation for Information Extraction in Business Documents.

The increasing reliance on OCR technologies to analyze digitized business documents has enabled large-scale automation but also introduced new challenges for information extraction systems. While OCR engines perform well under ideal conditions, they remain prone to errors. Traditional OCR metrics like character and word error rates fail to capture the impact of such errors on downstream tasks, particularly when only semantically critical words are affected. In this paper, we systematically investigate the relationship between OCR quality and extraction accuracy in business documents. We introduce a controlled evaluation framework that simulates realistic OCR noise scenarios by selectively injecting errors into clean datasets. Our experiments show that standard OCR metrics poorly reflect the impact of noise on information extraction performance, especially under heterogeneous noise. These findings and highlight the need for task-specific OCR evaluation protocols and more resilient pipelines tailored to real-world settings.

MOTS-CLÉS : extraction d'information, analyse de documents, numérisation, reconnaissance optique de caractères, documents commerciaux.

KEYWORDS: information extraction, document analysis, digitization, optical character recognition,

1 Introduction

La numérisation des flux documentaires est un élément essentiel de la transformation numérique, tant dans le secteur public que privé. Les organisations convertissent massivement leurs documents commerciaux (factures, contrats, formulaires) dans des formats numériques exploitables pour automatiser les traitements et réduire les coûts opérationnels. Les technologies d'OCR (Reconnaissance Optique de Caractères) jouent un rôle clé dans ce processus en transformant les images numérisées ou les PDF en texte éditable et interrogeable.

Cependant, malgré les progrès en analyse de mise en page et en modèles vision-langage (Xu *et al.*, 2020; Appalaraju *et al.*, 2021), les systèmes d'OCR actuels restent vulnérables aux erreurs, en particulier sur les documents semi-structurés avec des mises en page complexes ou inhabituelles (Li *et al.*, 2020; Šimsa *et al.*, 2023b; Ehrmann *et al.*, 2020). Les erreurs vont de simples coquilles à des distorsions majeures affectant considérablement les tâches en aval. La qualité de l'OCR est mesurée par le taux d'erreur de caractères (CER) et le taux d'erreur de mots (WER), qui quantifient la proportion de caractères ou mots mal reconnus (Neudecker *et al.*, 2021). Bien qu'utiles pour la transcription, ces métriques ne reflètent pas l'importance de certaines zones textuelles pour l'extraction d'informations.

Cette limite est particulièrement marquée dans l'analyse de documents commerciaux, où l'objectif n'est pas de restituer fidèlement l'intégralité du texte, mais d'extraire des champs clés (ex : numéros de facture, dates, montants). Dans ces scénarios, même des erreurs d'OCR mineures sur des tokens critiques peuvent dégrader fortement les performances du système. À l'inverse, un bruit important sur du texte non pertinent peut avoir un impact marginal sur l'extraction. La plupart des études négligent ces cas, et il n'existe pas de benchmarks ni de protocoles standards qui prennent en compte l'importance relative des erreurs selon les zones du document (Todorov & Colavizza, 2022; Jiang *et al.*, 2021; Hamdi *et al.*, 2023).

Pour combler ce manque, nous proposons une étude systématique du bruit OCR dans les documents commerciaux, ciblant spécifiquement l'extraction de champs clés et la reconnaissance des lignes tabulaires. La première tâche, consiste à identifier et à extraire automatiquement des informations structurées présentes dans des zones spécifiques du document, telles que le numéro de facture, la date, le nom du fournisseur, le montant total ou encore les informations relatives au client. Cette tâche relève généralement de l'extraction d'information et repose sur des méthodes de détection d'entités et de compréhension de la structure du document. La seconde tâche, la reconnaissance des lignes tabulaires, vise à détecter et interpréter les tableaux présents dans les documents, en particulier les lignes décrivant les articles ou services facturés. Cela implique non seulement l'identification de la structure tabulaire (lignes et colonnes), mais également l'extraction correcte des différentes cellules associées à chaque ligne, telles que la description du produit, la quantité, le prix unitaire ou le montant. Les deux tâches impliquent de détecter des éléments informationnels spécifiques en réponse à des critères explicites, qui s'apparente à un processus de recherche ciblée à l'intérieur du document. Bien que leurs objectifs opérationnels diffèrent, les deux tâches reposent sur l'identification, la sélection et l'interprétation d'informations pertinentes dans un contexte textuel bruité.

Dans un contexte OCR, les erreurs de reconnaissance de caractères peuvent altérer ces champs critiques, entraînant des erreurs de localisation, de classification ou de normalisation des valeurs

extraites. Ces tâches qui manipulent des documents structurés sont particulièrement sensibles au bruit OCR, car les erreurs de segmentation, de reconnaissance ou d'alignement peuvent perturber la reconstruction de la structure des documents. Nous montrons que les métriques standards ne reflètent pas l'impact réel des erreurs OCR. Notre contribution principale est un cadre d'évaluation qui simule un bruit réaliste, à la fois dans les zones pertinentes et non pertinentes, pour différents taux et types d'erreur. Les jeux de données existants ne fournissant ni scores de confiance OCR ni données d'alignement, nous injectons du bruit de façon contrôlée, dans des documents sans erreurs, afin de créer plusieurs versions de qualité OCR pour une analyse fine. Toutes les données utilisées dans cette étude sont publiquement accessibles [1].

Nos expériences sur un jeu de données de documents commerciaux mettent en évidence la nécessité d'approches plus nuancées dans l'évaluation de la qualité de l'OCR et motivent la conception de méthodes de post-traitement en tenant compte, ou de méthodes d'extraction pondérées par la confiance, adaptées aux contextes applicatifs réels.

La suite de cet article est organisée comme suit. La section 2 passe en revue les travaux connexes sur l'évaluation de l'OCR et l'extraction d'information dans les documents numérisés. La section 3 décrit le jeu de données et la simulation de bruit OCR. La section 4 présente ensuite les résultats expérimentaux et leurs analyses. Enfin, la section 5 conclut l'article et propose des perspectives.

2 Travaux connexes

L'impact du bruit OCR sur les tâches d'extraction d'information est un domaine de recherche en pleine expansion dans le traitement de documents numériques. Plusieurs études ont mené des évaluations systématiques montrant que la dégradation due à l'OCR nuit aux performances des tâches de traitement automatique des langues (TAL) (Lopresti, 2008; Van Strien *et al.*, 2020) et de recherche d'information (Chiron *et al.*, 2017; Giamphy *et al.*, 2023). Van Strien et al. (Van Strien *et al.*, 2020) ont conduit une évaluation systématique démontrant que la dégradation de l'OCR dégrade systématiquement les performances sur six tâches de TAL. La reconnaissance et la désambiguïsation d'entités nommées souffrent également de manière considérable (Hamdi *et al.*, 2020, 2019; Linhares Pontes *et al.*, 2019). On estime que la perte de performance en reconnaissance d'entités nommées peut atteindre 30 points de pourcentage en score F1 lorsque le CER passe de 7% à 20% (Hamdi *et al.*, 2023). Dans le même contexte, la campagne CLEF-HIPE 2020 (Ehrmann *et al.*, 2020) a systématisé l'évaluation de l'impact du bruit sur les mentions d'entités pour la reconnaissance, en utilisant la distance de Levenshtein pour quantifier la distorsion des entités induite par l'OCR. Les résultats montrent qu'un bruit textuel même minime peut réduire de moitié les performances des modèles.

Des recherches plus récentes continuent d'explorer cette question à l'aide de modèles transformers robustes. Par exemple, Todorov et al. (Todorov & Colavizza, 2022) et Jiang et al. (Jiang *et al.*, 2021) ont évalué des modèles de langue pré-entraînés comme BERT et RoBERTa sur des corpus dégradés par l'OCR, révélant que bien que ces modèles soient plus résilients que les architectures traditionnelles, ils subissent encore des baisses de performance significatives en présence d'erreurs d'OCR, en particulier pour les tâches centrées sur les entités. Des résultats récents montrent également que, malgré la robustesse des grands modèles de langue (LLM), ceux-ci peinent encore face à des entrées bruitées (González-Gallardo *et al.*, 2024, 2023).

1. https://doi.org/10.5281/zenodo.16638897

Bien que les études antérieures aient mis en lumière l'impact du bruit OCR sur les tâches d'extraction d'information, en particulier la reconnaissance d'entités nommées, elles se concentrent souvent sur la dégradation globale des performances sans distinguer comment différents types de contenu textuel sont affectés, ni comment des schémas de bruit spécifiques influencent la qualité de l'extraction. Notre étude propose ainsi une analyse plus fine en évaluant systématiquement l'impact de la qualité de l'OCR sur l'extraction d'information dans des scénarios multiples, incluant des cas où les mots pertinents et/ou non pertinents sont affectés différemment. Nous simulons une gamme de taux et de types d'erreurs, ce qui nous permet d'évaluer non seulement l'ampleur de la perte de performance, mais aussi la sensibilité des systèmes d'extraction d'information à la nature des erreurs d'OCR. À notre connaissance, il s'agit de la première étude à proposer une dégradation contrôlée de l'OCR pour mesurer l'impact sur différents types de termes, offrant ainsi de nouvelles perspectives quant à la robustesse et aux cas d'échecs des chaînes d'extraction dans des contextes réels de traitement de documents, tels que les documents commerciaux et administratifs (Hamdi *et al.*, 2021).

3 Modélisation du bruit OCR

Nous nous appuyons sur le jeu de données DocILE (Šimsa *et al.*, 2023a), et ses 6680 documents commerciaux annotés (factures, reçus, etc.). DocILE fournit des annotations clé-valeur, mais il ne propose pas de variantes dégradées par des erreurs d'OCR. Nous enrichissons alors DocILE avec du bruit OCR synthétique via un processus en deux étapes. Premièrement, nous appliquons des distorsions visuelles (ex : encre délavée, artefacts d'arrière-plan) à l'aide de la bibliothèque AuGraphy[2], puis nous extrayons le texte à l'aide de plusieurs systèmes d'OCR. Pour un contrôle plus fin, nous introduisons également du bruit au niveau des caractères directement dans le texte tokenisé, en corrompant sélectivement les champs clés, le contenu non pertinent, ou l'intégralité des documents. Les schémas de bruit reflètent les erreurs OCR courantes, incluant des confusions typiques comme "0"/"O" ou "m"/"rn", et préservent les annotations originales.

Cette configuration permet une évaluation systématique et reproductible de l'impact des erreurs d'OCR, en distinguant notamment les zones critiques des zones secondaires. Elle ouvre ainsi la voie à des protocoles d'évaluation mieux adaptés aux tâches considérées et contribue à orienter la conception de systèmes plus robustes.

3.1 Détection des mots pertinents

Pour mettre en œuvre une stratégie d'injection de bruit OCR tenant compte de la pertinence des termes, nous avons développé deux méthodes complémentaires pour identifier les termes pertinents dans un document :

— **Appariement spatio-textuel** : Un segment de texte est considéré comme pertinent si le centre de sa boîte englobante se situe à l'intérieur de la boîte englobante d'un champ annoté et si son contenu textuel correspond, totalement ou partiellement, au texte de référence associé à ce champ. Ce critère combine ainsi une contrainte spatiale, assurant que le segment est correctement localisé dans la mise en page du document, et une contrainte sémantique, vérifiant la correspondance entre le texte reconnu par l'OCR et l'annotation de référence.

2. https://github.com/sparkfish/augraphy

Cette approche se justifie par l'hétérogénéité des sorties d'OCR, qui varient selon les outils ou les configurations. À contenu textuel identique, les boîtes englobantes peuvent varier en granularité (mot, ligne, phrase), en taille ou en position, notamment en raison de l'analyse de mise en page. Cette variabilité est particulièrement problématique pour les structures tabulaires, où un champ annoté peut couvrir plusieurs éléments textuels. Notre méthode assure ainsi une association fiable entre segments OCR et champs annotés, malgré ces incohérences géométriques.

— **Appariement basé sur le type de champ** : dans cet appariement, la pertinence repose uniquement sur le contenu textuel. Tous les tokens annotés étant associés à l'un des 55 types de champs, ils sont considérés comme pertinents par construction. Plus généralement, tout terme dont le contenu correspond, totalement ou partiellement, à un texte annoté est considéré pertinent, indépendamment de l'alignement de sa boîte englobante dans le document. Cette seconde approche ne retient donc qu'un critère sémantique, sans contrainte spatiale. Cette approche élargit l'inclusion et réduit le risque de négliger un contenu important en raison de légers décalages spatiaux ou de la tendance de l'OCR à déplacer les positions du texte. Le principe sous-jacent est que dans les documents commerciaux, les types de champs (ex : numéro de facture, montant total, date, nom du fournisseur) ont une importance spécifique au domaine, indépendante de leur disposition spatiale. En priorisant les mots liés à ces types de champs, nous garantissons que même si l'OCR déplace ou déforme légèrement les boîtes englobantes, le contenu lui-même est toujours capturé comme pertinent. Conceptuellement, cette méthode traite le fichier d'annotation comme un dictionnaire de valeurs "à capturer", et le fichier OCR comme l'ensemble candidat dans lequel nous recherchons ces valeurs.

Le tableau 1 illustre la différence entre les deux méthodes. Tout terme qui n'est identifié comme pertinent par aucune de ces deux méthodes est considéré comme non pertinent pour les tâches d'extraction de champs clefs et la reconnaissance des données tabulaires. Cette distinction nous permet de simuler et d'évaluer les performances des systèmes d'extraction dans un large éventail de scénarios OCR réalistes : des erreurs affectant uniquement des termes pertinents, uniquement des termes non pertinents, ou une combinaison des deux. De telles configurations contrôlées sont essentielles pour analyser la robustesse des pipelines d'extraction face au bruit et pour mieux comprendre l'impact des erreurs introduites par l'OCR sur les performances globales du système.

TABLE 1 – Exemple d'entrée et de sortie pour la détection des mots pertinents

OCR :

```
"words":
"value": "Main", "snapped_geometry": [[400, 300], [450, 320]],
"value": "Street", "snapped_geometry": [[120, 205], [160, 220]]
```

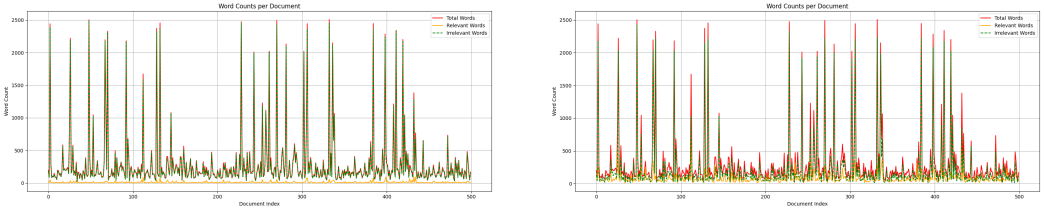
Annotation :

```
"text": "123 Main Street"
"bbox": [100, 200, 300, 220]
```

Résultats :

— **Appariement spatio-textuel :** "Main" → irrelevant (pas de chevauchement), "Street" → relevant

— **Appariement basé sur le type de champ :** "Main" and "Street" → both relevant



(a) Appariement spatio-textuel

(b) Appariement basé sur le type de champ

FIGURE 1 – Distribution des mots pertinents et non pertinents selon les deux méthodes d'étiquetage

Comme le montre la figure 1, la distribution des mots révèle que la méthode d'appariement spatio-textuel considère plus de 80% des termes comme non pertinents, tandis que la méthode basée sur le type de champ en identifie environ 60% comme non pertinents. Ces disparités dans l'étiquetage de la pertinence suggèrent que le bruit OCR peut affecter les performances du modèle différemment selon l'endroit où il se produit dans le document, soulignant l'importance d'évaluer les effets du bruit de manière contextuelle.

3.2 Scénarios d'injection de bruit OCR

Pour étudier les effets des erreurs OCR sur l'extraction d'information, nous avons conçu un cadre contrôlé d'injection de bruit simulant différents types et niveaux de sévérité d'erreurs. Nous avons défini trois scénarios : (1) un bruit injecté exclusivement dans les termes pertinents, représentant le cas présumé le plus critique où les informations essentielles sont altérées ; (2) un bruit appliqué aux termes non pertinents, permettant de tester l'interférence contextuelle ; (3) un bruit réparti sur l'ensemble du document, simulant une dégradation générale de l'OCR.

La simulation du bruit s'appuie sur un ensemble de confusions de caractères de 126 substitutions inspirées des erreurs OCR courantes, incluant des confusions chiffres-lettres (ex : `0` ↔ `O`, `1` ↔ `l`, `5` ↔ `S`) et des confusions lettres-lettres (ex : `B` ↔ `R`, `c` ↔ `e`, `m` ↔ `rn`). Les substitutions sont bidirectionnelles, limitées aux caractères alphanumériques, et préservent la structure des champs.

La sévérité du bruit est contrôlée par deux paramètres : le WER (proportion de mots corrompus) et le nombre de caractères modifiés par mot. Par exemple, un WER de 20% et un nombre d'erreurs de caractère modifiés par mot de 2 indiquent que 20% des mots sont altérés, chacun avec deux caractères substitués. Lorsque la longueur d'un mot est inférieure au nombre d'erreurs de caractère modifiés par mot spécifié, les substitutions sont appliquées jusqu'à concurrence de sa longueur, afin qu'aucun mot sélectionné ne soit exclu.

4 Analyse empirique et résultats

Dans cette section, nous présentons le dispositif expérimental et les résultats d'évaluation pour deux tâches : la Localisation et Extraction d'Informations Clés (KILE) et la Reconnaissance de Lignes Tabulaires (LIR), sous différents niveaux de bruit OCR simulé.

4.1 Mise en œuvre des modèles de référence

Nous avons évalué quatre modèles de référence : RoBERTa (Liu *et al.*, 2019) et LayoutLMv3 (Huang *et al.*, 2022), chacun disponible en deux variantes : une pré-entraînée uniquement sur des documents réels, et une deuxième sur un mélange de documents réels et synthétiques. Tous les modèles ont été affinés sur les ensembles d'entraînement et de validation de DocILE en suivant le protocole officiel du benchmark pour KILE et LIR (Šimsa *et al.*, 2023a). L'ensemble de test n'a pas été utilisé, car il n'est pas publiquement accessible.

L'évaluation des performances a utilisé les métriques standard d'extraction d'information, notamment la Précision Moyenne (AP), le score F1, la précision et le rappel. Selon le benchmark DocILE, l'AP a été utilisée comme métrique principale pour l'évaluation de la tâche KILE, tandis que le score F1 était la métrique principale pour la tâche LIR. Nos résultats d'évaluation sont cohérents avec ceux rapportés dans l'article original sur DocILE (Šimsa *et al.*, 2023a).

4.2 Dispositif expérimental

Pour étudier l'impact des erreurs d'OCR sur l'extraction d'information, nous avons mené des expériences systématiques dans des conditions de bruit contrôlées. Le bruit a été injecté selon des niveaux prédéfinis de WER et nombre d'erreurs de caractère modifiés par mot.

Chaque scénario est défini par quatre valeurs de WER :

- $\text{WER}_{\text{rel\_stm}}$ pour les mots pertinents selon les méthodes d'appariement spatio-textuel
- $\text{WER}_{\text{rel\_ftm}}$ selon les méthodes d'appariement basé sur le type de champ
- WER_{irr} pour les mots non pertinents
- WER désignant le taux d'erreur global au niveau mot.

Il est important de noter que ces valeurs de WER ne correspondent pas aux paramètres prédéfinis, mais aux taux d'erreur réels mesurés dans chaque document après l'injection de bruit. En considérant toutes les combinaisons de 10 WER cibles (de 5% à 50%, par pas de 5%), 6 niveaux de CER (de 1 à 6 erreurs de caractères par mot), 3 options d'injection (pertinent, non pertinent et aléatoire) et 2 méthodes (appariement spatio-textuel (STM) et appariement basé sur les champs-clefs (FTM)), nous obtenons 300 versions bruitées par document.

4.3 Résultats et analyse

Nous avons évalué RoBERTa et LayoutLMv3 sur les tâches KILE et LIR sous différents niveaux de bruit OCR. Les performances diminuent régulièrement avec l'augmentation du CER et du WER, sans preuve de stabilisation ou d'effet de seuil, ce qui indique que les erreurs d'OCR dégradent la précision de l'extraction d'informations de manière progressive et cumulative. Le tableau 2 montre l'évolution des performances selon les différents scénarios et niveaux d'erreur.

Pour la tâche KILE avec RoBERTa, l'injection de seulement 2,5% de WER sur les mots pertinents (STM) fait chuter l'AP à environ 42%. Un niveau similaire n'est atteint qu'avec des niveaux de bruit bien plus élevés, soit 16% de WER sur les mots non pertinents ou 13% de WER global, ce qui met en évidence l'impact disproportionné des erreurs affectant les mots pertinents. Avec la définition basée sur le type de champ, un WER de 2,5% produit une AP plus élevée (49%), proche de celle obtenue avec 5% de WER dans la configuration spatio-textuelle. À 18,3% de $\text{WER}_{\text{rel-ftm}}$, l'AP chute à 21%,

TABLE 2 – Impact de l'OCR sur les résultats d'extraction d'information selon les quatre scénarios (%)

		mots non pertinents			mots pertinents STM			mots pertinents FTM			mots aléatoires		
		WER	AP	F1	WER	AP	F1	WER	AP	F1	WER	AP	F1
KILE	RoBERTa	4.41	52.7	66.2	0.71	53.7	67.5	1.58	53.5	67.0	5	53.0	66.4
		9.18	50.1	63.4	0.77	53.3	67.4	3.38	50.8	64.3	10	50.7	63.7
		13.88	47.0	60.1	0.88	52.1	66.5	5.22	48.7	62.5	15	44.9	58.5
		18.71	41.6	55.4	1.05	50.5	65.6	7.13	43.1	57.2	20	37.8	52.7
		23.45	36.5	50.2	1.25	48.4	63.9	8.97	39.0	53.9	25	31.0	45.6
		28.13	32.5	46.2	1.45	46.3	62.4	10.80	33.2	48.4	30	25.1	40.0
		32.87	29.0	42.5	1.65	45.9	61.9	12.65	30.6	44.8	35	20.4	34.9
		37.67	24.6	38.7	1.95	44.1	60.7	14.54	27.0	41.4	40	16.5	31.0
		42.36	22.1	35.5	2.14	43.6	60.4	16.33	24.4	39.4	45	12.9	26.1
		47.25	19.3	32.6	2.49	41.6	58.8	18.33	21.0	35.4	50	9.8	22.1
	LayoutLMv3	4.41	52.4	66.5	0.71	52.5	67.0	1.58	52.5	66.7	5	51.7	66.3
		9.18	50.2	64.8	0.77	51.5	66.8	3.38	51.5	66.4	10	49.6	66.4
		13.88	47.2	62.5	0.88	50.0	66.7	5.22	50.5	64.7	15	46.0	61.0
		18.71	43.8	58.6	1.05	46.9	66.2	7.13	46.9	62.4	20	41.7	57.6
		23.45	39.9	55.0	1.25	44.4	64.5	8.97	44.4	59.6	25	36.3	52.0
		28.13	36.5	51.3	1.45	39.7	63.8	10.80	39.7	55.8	30	31.2	47.2
		32.87	34.2	49.3	1.65	37.3	63.1	12.65	37.3	53.1	35	26.8	43.0
		37.67	29.8	45.3	1.95	34.3	62.1	14.54	34.3	50.3	40	23.3	38.8
		42.36	26.9	41.9	2.14	32.1	61.5	16.33	32.1	48.1	45	19.2	34.7
		47.25	23.7	38.6	2.49	28.9	60.9	18.33	28.9	45.0	50	15.1	29.6
LIR	RoBERTa	4.41	66.6	53.5	0.71	68.0	54.5	1.58	67.8	54.4	5	53.0	66.2
		9.18	64.6	50.8	0.77	67.9	54.4	3.38	65.8	52.0	10	49.7	64.1
		13.88	62.2	46.7	0.88	67.2	53.6	5.22	61.2	47.2	15	44.7	59.5
		18.71	58.1	42.1	1.05	66.0	52.1	7.13	57.2	41.7	20	37.5	53.4
		23.45	52.8	38.4	1.25	65.1	50.8	8.97	53.6	36.2	25	31.7	47.9
		28.13	48.0	31.2	1.45	64.0	49.5	10.80	47.3	31.0	30	24.4	40.7
		32.87	43.2	27.5	1.65	63.2	48.5	12.65	43.6	26.2	35	18.6	34.6
		37.67	40.4	23.8	1.95	62.0	47.0	14.54	39.9	23.0	40	15.0	29.8
		42.36	36.3	19.9	2.14	61.0	45.9	16.33	35.2	19.2	45	10.9	24.7
		47.25	31.5	16.8	2.49	60.0	44.7	18.33	31.7	15.1	50	8.5	21.5
	LayoutLMv3	4.41	56.1	67.9	0.71	56.5	68.4	1.58	56.4	68.3	5	55.8	67.8
		9.18	55.8	67.9	0.77	56.2	68.0	3.38	55.3	67.4	10	54.2	66.7
		13.88	54.5	67.1	0.88	56.1	68.1	5.22	54.3	66.9	15	53.7	66.5
		18.71	51.5	64.7	1.05	55.6	67.8	7.13	51.9	65.3	20	50.3	64.2
		23.45	49.7	62.7	1.25	54.5	67.3	8.97	48.4	62.8	25	44.4	59.6
		28.13	46.0	59.6	1.45	54.1	67.0	10.80	44.6	60.0	30	40.9	56.8
		32.87	43.9	58.3	1.65	53.4	66.6	12.65	40.9	57.5	35	35.7	51.9
		37.67	40.3	54.8	1.95	52.3	66.0	14.54	38.3	55.0	40	35.7	51.7
		42.36	38.7	53.9	2.14	51.0	65.2	16.33	35.9	53.1	45	30.2	46.7
		47.25	35.6	50.9	2.49	49.6	64.2	18.33	31.9	49.3	50	26.8	43.2

contre 47% lorsque les erreurs touchent uniquement les mots non pertinents et 36% en cas de bruit global. Des tendances similaires sont observées avec LayoutLMv3, avec des scores F1 légèrement supérieurs.

En LIR, RoBERTa a atteint 60% d'AP à 2,5% de $\text{WER}_{\text{rel-stm}}$, comparable à 12% de WER_{irr} ou 9% de WER global. Avec 4% de $\text{WER}_{\text{rel-ftm}}$, l'AP est montée à 63%, mais a chuté à 31% avec 18,3% de WER. Un bruit sur les termes non pertinents au même niveau (4%) a permis d'obtenir une AP de 62%. LayoutLMv3 a obtenu des scores F1 de 64–66% à faible WER_{rel}, équivalant à 14–18% de WER_{irr}.

Pour la tâche LIR, RoBERTa atteint 60% d'AP avec seulement 2,5% de $\text{WER}_{\text{rel-stm}}$, un niveau comparable à celui obtenu avec 12% de WER sur les mots non pertinents ou 9% de WER global. Avec la définition basée sur le type de champ, 4% de $\text{WER}_{\text{rel-ftm}}$ produit une AP de 63%, mais celle-ci chute à 31% lorsque le WER atteint 18,3%. À niveau de bruit équivalent (4%), les erreurs affectant uniquement les mots non pertinents maintiennent une AP plus élevée (62%). LayoutLMv3 présente une évolution similaire des performances, avec des scores F1 de 64–66% à faible WER_{rel}, comparables à 14–18% de WER_{irr}.

Globalement, 2,5% de WER_{rel} équivaut approximativement à 16% de WER_{irr} ou 13% de WER global selon les tâches. Avec la seconde méthode, 18,3% de WER_{rel} correspond à 31% de WER_{irr} et jusqu'à 42% de WER global. Ces résultats confirment le caractère disproportionné de l'impact des erreurs OCR lorsqu'elles affectent des zones pertinentes : indépendamment de la densité globale de bruit, le facteur déterminant demeure leur localisation. Les deux modèles présentent d'ailleurs des schémas de dégradation cohérents. La robustesse varie également selon les types de champs et les modèles, soulignant la nécessité de protocoles d'évaluation adaptés aux tâches.

Globalement, l'impact de 2,5% de WER_{rel} correspond approximativement à l'impact de 16% de WER_{irr} ou 13% de WER global selon les tâches. Avec la seconde méthode, l'impact de 18,3% de WER_{rel} équivaut à l'impact de 31% de WER_{irr} et jusqu'à 42% de WER global.

Ces résultats mettent en évidence la forte sensibilité des performances aux erreurs affectant les zones pertinentes, indépendamment de leurs densités globales indiquées par les métriques standard WER/CER. Les deux modèles testés présentent des profils de dégradation cohérents, avec toutefois des variations selon les types de champs, soulignant la nécessité de protocoles d'évaluation adaptés aux spécificités des tâches.

5 Conclusion

Cette étude met en évidence le fait que les erreurs d'OCR affectant les zones pertinentes des documents commerciaux ont un impact nettement disproportionné sur les performances d'extraction d'information, comparativement au bruit localisé dans le contenu non pertinent. Elle montre ainsi que la localisation des erreurs constitue un facteur bien plus déterminant que leur volume global. Par ailleurs, LayoutLMv3 s'est révélé plus robuste que RoBERTa, grâce à l'exploitation conjointe des informations textuelles et de la mise en page, ce qui lui permet de mieux compenser certaines dégradations textuelles.

Ces résultats soulignent la nécessité de stratégies OCR et de post-traitement ciblant en priorité la protection des champs critiques. Les perspectives de recherche incluent l'analyse des erreurs d'espacement, l'étude de la sensibilité selon les types de champs, ainsi que la conception de systèmes

d'OCR plus adaptés aux exigences des tâches d'extraction d'information à partir de documents commerciaux.

Remerciements

Ce travail a été cofinancé par l'Union européenne via la subvention HORIZON-WIDERA-2023-TALENTS-01-01 n°101186647 - AI4DH. Les points de vue et opinions exprimés ici n'engagent toutefois que leurs auteurs et ne reflètent pas nécessairement ceux de l'Union européenne. Ni l'Union européenne ni l'autorité de financement ne peuvent en être tenues pour responsables.

Références

APPALARAJU S., JASANI B., KOTA B. U., XIE Y. & MANMATHA R. (2021). Docformer : End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, p. 993–1003.

CHIRON G., DOUCET A., COUSTATY M., VISANI M. & MOREUX J.-P. (2017). Impact of ocr errors on the use of digital libraries : towards a better access to information. In *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, p. 1–4 : IEEE.

EHRMANN M., ROMANELLO M., FLÜCKIGER A. & CLEMATIDE S. (2020). Overview of clef hipe 2020 : Named entity recognition and linking on historical newspapers. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, p. 288–310 : Springer.

GIAMPHY E., SANCHIS K., DASHYAN G., GUILLAUME J.-L., HAMDI A., SANSELME L. & DOUCET A. (2023). A quantitative analysis of noise impact on document ranking. In *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, p. 4612–4618 : IEEE.

GONZÁLEZ-GALLARDO C.-E., BOROS E., GIRDHAR N., HAMDI A., MORENO J. G. & DOUCET A. (2023). Yes but.. can chatgpt identify entities in historical documents ? In *2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, p. 184–189 : IEEE.

GONZÁLEZ-GALLARDO C.-E., TRAN H. T. H., HAMDI A. & DOUCET A. (2024). Leveraging open large language models for historical named entity recognition. In *International Conference on Theory and Practice of Digital Libraries*, p. 379–395 : Springer.

HAMDI A., CAREL E., JOSEPH A., COUSTATY M. & DOUCET A. (2021). Information extraction from invoices. In *Proceedings of the 16th International Conference on Document Analysis and Recognition, ICDAR 2021*, volume 12822 de *Lecture Notes in Computer Science*, p. 699–714, Lausanne, Switzerland : Springer.

HAMDI A., JEAN-CAURANT A., SIDERE N., COUSTATY M. & DOUCET A. (2019). An analysis of the performance of named entity recognition over ocred documents. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, p. 333–334 : IEEE.

HAMDI A., JEAN-CAURANT A., SIDÈRE N., COUSTATY M. & DOUCET A. (2020). Assessing and minimizing the impact of ocr quality on named entity recognition. In *International Conference on Theory and Practice of Digital Libraries*, p. 87–101 : Springer.

HAMDI A., PONTES E. L., SIDERE N., COUSTATY M. & DOUCET A. (2023). In-depth analysis of the impact of ocr errors on named entity recognition and linking. *Natural language engineering*, **29**(2), 425–448.

HUANG Y., ZHANG Y., MA J., JIANG Y., HUANG X., SHAN Y., TAN M. & YU D. (2022). Layoutlmv3 : Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*, p. 1915–1924 : Association for Computing Machinery.

JIANG M., HU Y., WORTHEY G., DUBNICEK R. C., UNDERWOOD T. & DOWNIE J. S. (2021). Impact of ocr quality on bert embeddings in the domain classification of book excerpts. *Proceedings http ://ceur-ws. org ISSN*, **1613**, 0073.

LI M., CUI L., HUANG S., WEI F., ZHOU M. & LI Z. (2020). Tablebank : Table benchmark for image-based table detection and recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 1918–1925.

LINHARES PONTES E., HAMDI A., SIDERE N. & DOUCET A. (2019). Impact of ocr quality on named entity linking. In *International Conference on Asian Digital Libraries*, p. 102–115 : Springer.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). Roberta : A robustly optimized bert pretraining approach. *arXiv preprint arXiv :1907.11692*, (1907.11692). cs.CL, DOI : 10.48550/arXiv.1907.11692.

LOPRESTI D. (2008). Measuring the impact of character recognition errors on downstream text analysis. In *Document Recognition and Retrieval XV*, volume 6815, p. 68150G : International Society for Optics and Photonics.

NEUDECKER C., BAIERER K., GERBER M., CLAUSNER C., ANTONACOPOULOS A. & PLETSCHACHER S. (2021). A survey of ocr evaluation tools and metrics. In *Proceedings of the 6th International Workshop on Historical Document Imaging and Processing*, p. 13–18.

ŠIMSA Š., ŠULC M., UŘIČÁŘ M., PATEL Y., HAMDI A., KOCIÁN M., SKALICKỲ M., MATAS J., DOUCET A., COUSTATY M. *et al.* (2023a). Docile benchmark for document information localization and extraction. In *International Conference on Document Analysis and Recognition*, p. 147–166 : Springer.

ŠIMSA Š., UŘIČÁŘ M., ŠULC M., PATEL Y., HAMDI A., KOCIÁN M., SKALICKỲ M., MATAS J., DOUCET A., COUSTATY M. *et al.* (2023b). Overview of docile 2023 : document information localization and extraction. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, p. 276–293 : Springer.

TODOROV K. & COLAVIZZA G. (2022). An assessment of the impact of ocr noise on language models. *arXiv preprint arXiv :2202.00470*.

VAN STRIEN D., BEELEN K., ARDANUY M. C., HOSSEINI K., MCGILLIVRAY B. & COLAVIZZA G. (2020). Assessing the impact of ocr quality on downstream nlp tasks.

XU Y., LI M., CUI L., HUANG S., WEI F. & ZHOU M. (2020). Layoutlm : Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, p. 1192–1200.