Au-delà du CER et du WER : l'impact réel de l'OCR sur la recherche d'information

Alexandre Jaud[1] Ahmed Hamdi[2] Antoine Doucet[1,3] Adam Jatowt[4] Mickaël Coustaty[1]

(1) L3i, La Rochelle université, France
(2) IRIT, Université de Toulouse, France
(3) University de Ljubljana, Slovénie
(4) Université d'Innsbruck, Autriche

alexandre.jaud@univ-lr.fr, ahmed.hamdi@irit.fr, antoine.doucet@univ-lr.fr, mickael.coustaty@univ-lr.fr, adam.jatowt@uibk.ac.at

RÉSUMÉ

La numérisation massive de documents repose sur l'OCR pour rendre le contenu des documents accessible, mais les erreurs de reconnaissance, notamment sur les documents dégradés, impactent, entre autres, la tâche de recherche d'information. Les métriques classiques (CER, WER) traitent toutes les erreurs de façon identique, ignorant l'importance des entités nommées. Cet article analyse l'impact des erreurs d'OCR sur les performances de la reconnaissance d'entités nommées et par conséquent sur la recherche d'information. Via un cadre d'évaluation simulant divers types de corruption du texte, nous montrons qu'altérer même légèrement les entités nommées dégrade significativement les performances de recherche. Ces résultats appellent à une évaluation de l'OCR prenant en compte l'importance de certains termes. Nous proposons des stratégies pratiques : correction sélective des termes critiques, indexation pondérée par confiance, et modèles préservant l'intégrité des entités, pour des systèmes de recherche plus robustes vis à vis de corpus bruités.

ABSTRACT

Beyond CER and WER : How Does OCR Really Impact Information Retrieval ?

Large-scale digitization relies on OCR for document accessibility, but OCR errors can harm information retrieval. Traditional metrics like CER and WER treat all errors equally, overlooking the importance of named entities. This paper examines how OCR errors affect named entity recognition and information retrieval, focusing on the differential impact on key terms versus less informative words. Experiments simulating random and entity-specific corruption show that minor entity errors significantly hurt retrieval performance, even when CER/WER scores remain unchanged. We advocate for entity-aware OCR evaluation and propose practical strategies : selective correction of critical terms, confidence-weighted indexing, and retrieval models preserving entity integrity, enabling more robust search over noisy collections.

MOTS-CLÉS : Recherche d'information, analyse de documents, reconnaissance optique de caractères, taux d'erreurs, métriques.

KEYWORDS: Information retrieval, document analysis, Optical character recognition, Character and word error rates.

1 Introduction

Depuis plusieurs décennies, des campagnes de numérisation massives transforment des collections de documents en ressources accessibles numériquement. La reconnaissance optique des caractères (OCR) est au cœur de cette transformation, permettant la conversion automatique de textes imprimés en format exploitable par machine. Les progrès récents des techniques d'apprentissage profond ont considérablement amélioré les performances des systèmes OCR, notamment sur des documents propres (Appalaraju *et al.*, 2021; Xu *et al.*, 2020). Parallèlement, les systèmes de recherche d'information (RI) s'appuient sur ces collections numérisées pour offrir une recherche plein texte, une indexation d'entités et des analyses sémantiques (Marinai *et al.*, 2011). Cependant, les erreurs d'OCR restent fréquentes, particulièrement sur les documents dégradés, les écritures anciennes ou les mises en page complexes (Gupta *et al.*, 2015). Ces erreurs peuvent significativement impacter l'efficacité de la RI, surtout quand elles concernent des termes clés.

Les entités nommées (noms de personnes, lieux, organisations) jouent un rôle crucial en RI. Elles constituent souvent le point d'entrée principal des requêtes utilisateurs (Gefen, 2014). Dans Gallica, par exemple, le portail numérique de la bibiothèque nationale de France, environ 80% des requêtes contiennent au moins une entité nommée (Chiron *et al.*, 2017). Leur préservation est donc essentielle pour l'efficacité des systèmes et la satisfaction des utilisateurs. Les métriques traditionnelles d'évaluation de l'OCR : *Character Error Rate* (CER) et *Word Error Rate* (WER) (Neudecker *et al.*, 2021) traitent toutes les erreurs d'une manière uniforme et ignorent leur rôle sémantique ou leur impact sur la RI. C'est intuitivement problématique puisqu'une erreur sur un terme pertinent peut rendre un document introuvable, tandis qu'une erreur sur d'autres mots peut rester sans conséquence. Par exemple, dans un scénario de RI dans une collection scientifique, une erreur d'OCR sur un nom de protéine ou une molécule peut faire qu'un document soit complètement absent des résultats de RI, là où une erreur sur un mot vide n'aurait probablement aucune conséquence pratique.

Malgré les progrès récents dans l'analyse des systèmes de RI face aux contenus bruités (Rao *et al.*, 2022), rares sont les études qui ont systématiquement examiné comment différents types d'erreurs d'OCR influencent la qualité de la RI. En particulier, la plupart des approches existantes considèrent des taux d'erreur agrégés sans distinguer si les erreurs se produisent sur des termes critiques pour la RI ou sur du texte non pertinent. Le CER et le WER représentent donc peut-être de façon inadéquate l'impact des erreurs d'OCR sur les applications finales.

Nous proposons ici une évaluation plus nuancée de l'impact de l'OCR sur la reconnaissance d'entités nommées ainsi que la recherche d'information. Notre cadre expérimental simule divers scénarios de bruit (aléatoire ou ciblé sur les entités) pour quantifier l'impact différentiel selon que des entités ou des mots non critiques sont affectés. Nos expériences montrent que (1) les métriques OCR sont faiblement corrélées avec la performance de la RI, surtout quand les entités sont altérées, et (2) même une corruption modérée des entités dégrade significativement la performance. Nous proposons enfin des recommandations pour améliorer cette situation.

La suite de l'article est organisée comme suit. La section 2 présente les travaux connexes. La section 3 décrit ensuite notre méthodologie. Les résultats sont présentés en section 4. Enfin, la section 5 conclut l'article et expose les perspectives.

2 Travaux connexes

L'impact de la qualité de l'OCR sur des tâches de traitement automatique des langues a été largement étudié. Les premières recherches ont montré les effets en cascade des erreurs sur des modules comme l'étiquetage morpho-syntaxique et l'analyse syntaxique (Lopresti, 2008, 2005). La segmentation en phrases et l'analyse syntaxique sont particulièrement sensibles (Van Strien *et al.*, 2020). Plusieurs travaux ont spécifiquement évalué l'effet du bruit OCR sur la reconnaissance d'entités nommées. Hamdi *et al.* (2020) observent une chute de 30 points de F1 quand le CER passe de 7% à 20%. La tâche CLEF-HIPE 2020 (Ehrmann *et al.*, 2020) montre qu'un bruit minime (distance d'édition de 0,1) peut réduire de moitié les performances. Les modèles transformers comme BERT résistent mieux au bruit que les architectures traditionnelles (Jiang *et al.*, 2021; Nguyen *et al.*, 2020; Todorov & Colavizza, 2022; Boros *et al.*, 2022), mais se dégradent quand les entités sont touchées (Giamphy *et al.*, 2023). Les grands modèles de langue, malgré leur robustesse, peinent aussi avec les entrées bruitées (González-Gallardo *et al.*, 2024b, 2023).

En recherche d'information, Taghva *et al.* (1996) ont montré que précision et rappel se dégradent avec les erreurs OCR. Chiron *et al.* (2017) soulignent que la corruption des entités nommées cause des pertes disproportionnées. Pourtant, l'évaluation OCR reste dominée par des métriques agrégées (CER, WER) qui ignorent ces variations d'impact. Certaines recherches proposent des évaluations sémantiques, comme le taux d'erreur caractère pondéré (Oard, 2003) ou les distances d'édition pondérées (Guo *et al.*, 2025), mais ces approches restent peu intégrées dans des expériences contrôlées. Des travaux récents sur la génération augmentée par récupération (RAG) montrent que les erreurs d'OCR affectent directement performance de récupération, raisonnement et génération (Zhang *et al.*, 2025; González-Gallardo *et al.*, 2024a; Tran *et al.*, 2025).

Dans les moteurs de recherche, les utilisateurs utilisent généralement des entités, titres et auteurs. Les erreurs sur ces termes clés compromettent gravement la recherche. Pourtant, peu d'études distinguent erreurs sur tokens pertinents vs non pertinents, et aucun jeu de données public ne permet une évaluation fine standardisée. Notre article comble ces lacunes par une analyse contrôlée simulant des scénarios de dégradation réalistes, corrompant sélectivement les entités et les autres mots, pour mesurer l'impact différentiel sur la performance de la RI.

3 Modélisation du bruit OCR

Un obstacle majeur à l'avancement de la recherche à l'intersection de l'OCR, de la reconnaissance d'entités nommées et de la recherche d'information est l'absence de jeux de données de référence adaptés. À notre connaissance, aucun corpus public ne fournit simultanément des textes dégradés par OCR, leur vérité terrain propre, et des entités nommées annotées dans un format approprié pour évaluer la recherche d'information dans des conditions de bruit réalistes. Les jeux de données existants dans la littérature ne répondent généralement qu'à une partie de ce besoin : certains contiennent des annotations d'entités nommées de haute qualité mais manquent de versions bruitées correspondantes, tandis que d'autres fournissent des sorties OCR alignées et du texte de référence sans aucune annotation au niveau des entités. Cette lacune rend difficile l'étude de l'impact des erreurs d'OCR de manière contrôlée et reproductible.

Pour surmonter cette limitation, nous avons construit un benchmark synthétique. À partir de sources propres, nous avons généré de multiples versions bruitées en appliquant un processus de dégradation

OCR réaliste qui simule les erreurs de reconnaissance courantes observées dans les sorties OCR réelles. Notre méthode rapproche notre évaluation de conditions opérationnelles réelles, capturant non seulement les schémas de bruit courants mais aussi les distorsions OCR fréquentes dans les pipelines de traitement de documents. Ces versions bruitées couvrent une large gamme de taux d'erreur et de scénarios de contamination, nous permettant d'étudier leur impact différentiel sur les performances de RI. En faisant varier la quantité de bruit synthétisé affectant les entités nommées et les autres tokens, notre benchmark permet de démêler l'impact différentiel de la dégradation de l'OCR sur les performances de RI.

3.1 Jeux de données

Nos expériences incluent deux tâches : la reconnaissance d'entités nommées et la recherche d'information.

Pour la reconnaissance d'entités nommées, nous avons utilisé le corpus anglais du jeu de données CoNLL-2003 (Sang & De Meulder, 2003). Les données proviennent du corpus Reuters (août 1996–août 1997) et consistent en trois partitions annotées manuellement pour l'entraînement, le développement et le test. Les phrases sont tokenisées, et chaque forme (token) est associée à l'une des quatre étiquettes (labels) d'entités nommées (PER, LOC, ORG, MISC). Les marqueurs *B-label* indiquent le début d'une entité (notamment lorsqu'elle est adjacente à une autre entité du même type), tandis que *I-label* signale la continuité d'une entité nommée. Toute forme ne faisant pas partie d'une entité est annotée avec l'étiquette *O*. Dans ce corpus, l'ensemble d'entraînement comprend 946 articles (14987 phrases, 203621 tokens), l'ensemble de développement 216 articles (3466 phrases, 51362 tokens) et l'ensemble de test 231 articles (3684 phrases, 46435 tokens). Dans ces partitions, les entités sont bien équilibrées entre les quatre catégories. En préservant à la fois les limites des tokens et les annotations IOB, CoNLL-03 constitue un jeu de données idéal pour que notre pipeline d'injection de bruit corrompe sélectivement soit les tokens d'entité, soit les tokens de contexte, tout en laissant intacts les labels de référence.

Pour la RI, nous avons utilisé le jeu de données DBPedia (Lehmann *et al.*, 2015). Nous évaluons l'impact du bruit OCR sur la recherche ad-hoc en utilisant la portion DBPedia-Entity du benchmark BEIR (Thakur *et al.*, 2021) pour l'évaluation zero-shot à travers diverses tâches de RI. Son ensemble de test DBPedia-Entity comprend exactement 400 requêtes sur un corpus de 4635922 documents. Pour rendre l'évaluation traitable et se concentrer sur la distinction entre signal pertinent et bruit, nous avons d'abord extrait toutes les paires (requête, document) marquées comme pertinentes. Nous avons ensuite complété avec des paires non pertinentes, jusqu'à atteindre un total de 40724 documents pour les 400 requêtes. Ce sous-ensemble stratifié préserve la distribution originale des requêtes tout en fournissant un mélange équilibré d'exemples positifs et négatifs pour une évaluation robuste de la RI sous corruption OCR contrôlée. Pour mesurer l'impact du bruit OCR à différents niveaux linguistiques, nous avons appliqué un modèle de reconnaissance d'entités nommées pour annoter chaque document avec les spans d'entités et leur contexte. Nous avons ensuite produit plusieurs variantes bruitées du jeu de données en introduisant du bruit selon trois configurations distinctes : (i) sur les seules entités nommées, (ii) sur le texte hors entités, et (iii) sur l'intégralité du document.

3.2 Injection de bruit OCR

Afin de simuler de manière réaliste le bruit OCR dans ces jeux de données, nous avons d'abord identifié les schémas d'erreurs typiques de l'OCR, de manière à reproduire des conditions de bruit authentiques. Des corpus de texte propre ont été transformés en images de documents synthétiques, garantissant que les dégradations résultantes reflètent fidèlement celles observées dans des scénarios réels. Ces images ont ensuite été dégradées à l'aide de l'outil AuGraphy [1], une bibliothèque open-source flexible conçue pour générer des distorsions d'image au niveau document. Nous avons appliqué une combinaison de techniques d'augmentation qui imitent les facteurs de détérioration couramment rencontrés dans les collections numérisées des bibliothèques et des archives. Plus précisément, les dégradations suivantes ont été introduites : texturisation de la luminosité, simulation de rouleaux encrassés (appliquée deux fois pour amplifier les artefacts de stries), ajout de bruit léger, gradient d'illumination et insertion de lignes périodiques d'encre affaiblie. L'objectif est de modéliser des phénomènes fréquemment observés dans les documents numérisés, notamment l'illumination non uniforme, l'altération de l'encre, la dégradation du support papier et le bruit de numérisation. Ces différents facteurs représentent des sources de dégradation fréquemment rencontrées dans les processus de numérisation à grande échelle. Un exemple des images résultantes est fourni en figure 1. Les images dégradées ont ensuite été traitées par plusieurs moteurs OCR performants, tels que Tesseract [2], Cloud Vision AI [3], et Amazon Textract [4], pour extraire le texte bruité. Cette stratégie multi-moteurs nous a permis de capturer un large spectre de types d'erreurs OCR et d'analyser leur fréquence et distribution. Les schémas d'erreur résultants ont informé la conception de nos expériences contrôlées, garantissant que les scénarios de bruit synthétique que nous avons ensuite appliqués étaient fondés sur des observations empiriques de pipelines de dégradation OCR réalistes.

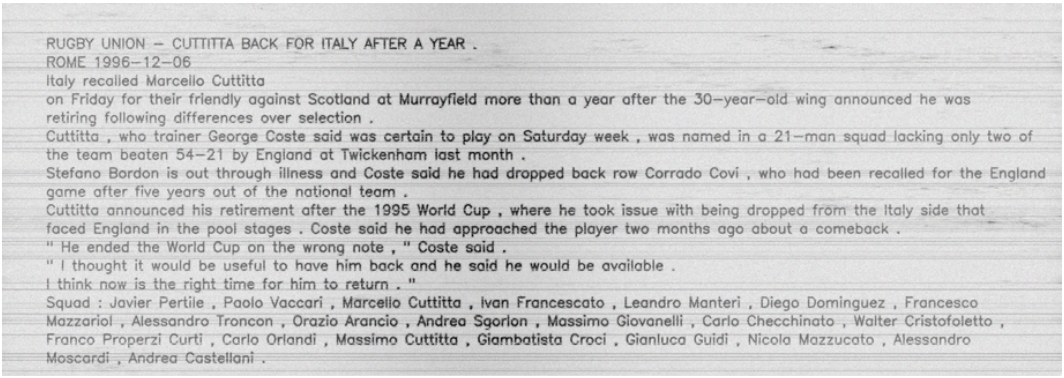
RUGBY UNION – CUTTITTA BACK FOR ITALY AFTER A YEAR .
ROME 1996–12–06
Italy recalled Marcello Cuttitta
on Friday for their friendly against Scotland at Murrayfield more than a year after the 30–year–old wing announced he was retiring following differences over selection .
Cuttitta , who trainer George Coste said was certain to play on Saturday week , was named in a 21–man squad lacking only two of the team beaten 54–21 by England at Twickenham last month .
Stefano Bordon is out through illness and Coste said he had dropped back row Corrado Covi , who had been recalled for the England game after five years out of the national team .
Cuttitta announced his retirement after the 1995 World Cup , where he took issue with being dropped from the Italy side that faced England in the pool stages . Coste said he had approached the player two months ago about a comeback .
" He ended the World Cup on the wrong note , " Coste said .
" I thought it would be useful to have him back and he said he would be available .
I think now is the right time for him to return . "
Squad : Javier Pertile , Paolo Vaccari , Marcello Cuttitta , Ivan Francescato , Leandro Manteri , Diego Dominguez , Francesco Mazzariol , Alessandro Troncon , Orazio Arancio , Andrea Sgorlon , Massimo Giovanelli , Carlo Checchinato , Walter Cristofoletto , Franco Properzi Curti , Carlo Orlandi , Massimo Cuttitta , Giambatista Croci , Gianluca Guidi , Nicola Mazzucato , Alessandro Moscardi , Andrea Castellani .

FIGURE 1 – Exemple d'image synthétique dégradée pour l'identification des schémas d'erreur OCR

Les erreurs OCR peuvent généralement être catégorisées en substitutions, insertions et suppressions, chacune pouvant déformer le texte de différentes manières. Les erreurs de substitution se produisent lorsqu'un caractère est pris pour un autre (par exemple, "O" ↔ "0"), souvent en raison d'une similarité visuelle. Les insertions surviennent lorsqu'un caractère supplémentaire est introduit de

1. https://github.com/sparkfish/augraphy
2. https://github.com/tesseract-ocr/tesseract
3. https://cloud.google.com/use-cases/ocr
4. https://aws.amazon.com/textract

manière erronée, tandis que les suppressions correspondent à la disparition d'un caractère réel.

Bien que notre cadre se concentre sur les substitutions, certaines séquences visuellement confuses, notamment "rn" et "m", peuvent produire des comportements émergents d'insertion ou de suppression. Par exemple, un vrai bigramme "rn" peut être réduit au monogramme "m" (suppression plus substitution), ou un "m" peut être faussement divisé en "rn" (insertion plus substitution), simulant des erreurs OCR plus complexes sans modélisation explicite de tous les types d'erreur.

Une analyse sensible à la tâche doit tenir compte non seulement des taux d'erreur bruts, mais aussi de l'impact linguistique de ces erreurs. Les substitutions au sein d'entités nommées (par exemple, "Paris" $\leftrightarrow$ "Par1s") peuvent gravement corrompre les tâches en aval de reconnaissance d'entités et de recherche, alors que la même substitution dans des mots vides peut être beaucoup moins dommageable. Par conséquent, distinguer où les erreurs se produisent est essentiel pour comprendre la dégradation des performances des tâches.

Afin d'évaluer rigoureusement l'impact des erreurs OCR sur les performances en aval, nous avons développé un script modulaire d'injection de bruit opérant directement sur des entrées au format CoNLL, tout en préservant les annotations d'entités au schéma IOB. Cette approche nous permet de générer des versions bruitées du jeu de données avec différents niveaux de bruit et de contrôler quels types de tokens sont affectés, une caractéristique essentielle de notre benchmark qui n'est pas disponible dans d'autres jeux de données. Cela nous permet, à notre tour, de mesurer précisément l'impact du bruit sur les entités nommées et non nommées.

À la lecture de chaque document au format CoNLL, le script identifie d'abord tous les spans de tokens annotés comme entités nommées et, par complément, ceux hors des spans d'entités. Dans nos expériences, nous utilisons ce marquage pour dégrader soit les tokens annotés comme des entités nommées, soit les autres, soit l'ensemble du texte. Pour chaque token choisi, une ou plusieurs positions de caractères sont sélectionnées aléatoirement. Les substitutions sont effectuées à l'aide d'un dictionnaire préchargé de 107 confusions OCR courantes, garantissant que les clés multi-caractères sont appariées préférentiellement pour simuler des erreurs de fusion ou de division réalistes. Une fois les substitutions appliquées, les tokens modifiés conservent leurs étiquettes IOB d'origine, de sorte que les évaluations en aval de la reconnaissance d'entités nommées et de de la recherche d'information ne reflètent que l'impact de la corruption textuelle, et non celui d'un éventuel désalignement des annotations. Enfin, les séquences de tokens corrompus sont réassemblées en fichiers CoNLL-03, prêts à être utilisés par les systèmes testés. Pour étudier systématiquement la robustesse des systèmes, nous appliquons le bruit sur un large spectre de conditions : entre 10% et 100% des mots de chaque document sont perturbés, selon trois scénarios distincts : (i) corruption restreinte aux entités nommées, (ii) corruption restreinte aux mots hors entités, et (iii) corruption de mots sélectionnés aléatoirement. Pour chaque scénario, nous faisons varier le degré de corruption en modifiant entre un et cinq caractères par mot, résultant en un total de 150 versions bruitées de chaque corpus.

Ce cadre configurable et reproductible permet ainsi des études ciblées sur la façon dont le bruit OCR centré sur les entités versus centré sur le contexte affecte différemment les performances des tâches. Dans le cadre de ce travail et pour soutenir la reproductibilité, nous publions le code source et les jeux de données bruités générés lors de nos expériences pour un usage public [5].

5. https://github.com/kadol-coder/noisy-datasets-ner-ir

TABLE 1 – Performances des systèmes de reconnaissance d'entités nommées à différents niveaux de bruit (score F1)

Système	Err./mot	Type bruit	Bruit 0%	Bruit 40%	Bruit 70%	Bruit 100%
FLAIR	1	NE	0,929	0,871	0,822	0,763
		non-NE	0,929	0,902	0,871	0,808
		ALL	0,929	0,826	0,686	0,454
	5	NE	0,929	0,829	0,720	0,575
		non-NE	0,929	0,865	0,799	0,687
		ALL	0,929	0,732	0,477	0,145
LUKE	1	NE	0,943	0,897	0,858	0,805
		non-NE	0,943	0,935	0,926	0,912
		ALL	0,943	0,879	0,790	0,621
	5	NE	0,943	0,836	0,728	0,560
		non-NE	0,943	0,930	0,916	0,892
		ALL	0,943	0,773	0,535	0,103
ACE	1	NE	0,948	0,909	0,876	0,830
		non-NE	0,948	0,940	0,930	0,901
		ALL	0,948	0,897	0,819	0,644
	5	NE	0,948	0,850	0,741	0,538
		non-NE	0,948	0,940	0,921	0,891
		ALL	0,948	0,791	0,551	0,082
CLNER L2	1	NE	0,940	0,888	0,844	0,790
		non-NE	0,940	0,913	0,890	0,859
		ALL	0,940	0,852	0,749	0,590
	5	NE	0,940	0,826	0,703	0,536
		non-NE	0,940	0,896	0,867	0,836
		ALL	0,940	0,748	0,515	0,114

4 Analyse empirique et résultats

Dans cette section, nous présentons une évaluation systématique de l'impact du bruit OCR sur la recherche d'information. Dans la mesure où cette tâche repose fortement sur la bonne reconnaissance des entités nommées, nous analysons dans un premier temps l'impact du bruit d'OCR sur la reconnaissance d'entités nommées. Nous étudions ensuite ses conséquences sur les performances globales de la recherche d'information.

4.1 Reconnaissance d'entités nommées

Nous avons sélectionné quatre architectures représentatives : Flair (Akbik *et al.*, 2019), LUKE (Yamada *et al.*, 2020), ACE (Wang *et al.*, 2021a), et CLNER L2 (Wang *et al.*, 2021b). Ces systèmes ont été évalués sur les versions bruitées du corpus CoNLL-2003 selon les trois scénarios de corruption : entités seulement (NE), hors entités (non-NE), ou tous les tokens (ALL).

L'analyse des résultats (tableau 1) révèle une disparité frappante dans la manière dont les systèmes de reconnaissance d'entités nommées réagissent au bruit OCR selon qu'ils traitent des tokens appar-

tenant à des entités ou des tokens (non-NE). Ces derniers démontrent une résilience remarquable face à la dégradation du texte. Même dans les conditions les plus extrêmes (bruit à 100% avec 5 erreurs par mot), les systèmes maintiennent des scores F1 impressionnants, oscillant entre 0,836 et 0,892. Cette robustesse s'explique probablement par la prédominance statistique des non-entités dans les corpus d'entraînement, ainsi que par la relative simplicité de leur classification : il s'agit essentiellement de reconnaître ce qui n'appartient pas aux catégories d'entités prédéfinies. Les patrons linguistiques réguliers du langage courant, même altérés, conservent suffisamment de caractéristiques discriminables pour permettre une identification correcte.

À l'inverse, les tokens appartenant à des entités subissent une dégradation beaucoup plus sévère. L'écart de performance entre NE et non-NE se creuse proportionnellement au niveau de bruit. À 0% de bruit, les scores sont naturellement identiques pour les deux catégories. Dès 40% de bruit avec 1 erreur par mot, un fossé apparaît : par exemple pour ACE, on observe 0,909 pour les NE contre 0,940 pour les non-NE. À 100% de bruit avec 5 erreurs par mot, cet écart atteint son paroxysme, avec parfois plus de 35 points d'écart (0,538 contre 0,891 pour ACE). Cette divergence croissante suggère que les mécanismes d'attention et les représentations contextuelles sur lesquels reposent ces modèles sont particulièrement perturbés lorsqu'il s'agit d'identifier des entités dans un texte dégradé.

Cette asymétrie dans la robustesse a des implications pratiques importantes pour les applications réelles traitant des documents numérisés. Elle suggère que les systèmes NER actuels, bien que capables de maintenir une bonne performance globale grâce à la reconnaissance des non-entités, deviennent rapidement inutilisables pour la tâche spécifique d'identification des entités lorsque la qualité OCR se dégrade. Pour des applications comme l'extraction d'information ou l'indexation sémantique, où les entités constituent l'information critique, ces résultats soulignent l'importance cruciale d'un prétraitement OCR de qualité ou du développement d'architectures spécifiquement conçues pour être robustes à ce type de dégradation.

4.2 Recherche d'information

Pour les expériences de recherche d'information, nous utilisons Elasticsearch avec deux variantes BM25 (standard et floue) et un modèle dense (Sentence-Transformer all-MiniLM-L6-v2).

L'analyse des performances (tableau 2) des trois systèmes de recherche (BM25 standard, BM25 flou et modèle dense) révèle des comportements contrastés face au bruit OCR, avec une vulnérabilité particulière des requêtes contenant des entités nommées. Le modèle dense surpasse nettement les approches BM25 en l'absence de bruit (0,558 contre 0,390 en nDCG@10), mais cette avance s'amenuise avec la dégradation du texte. À 100% de bruit avec 1 erreur par mot, BM25 flou maintient étonnamment mieux les performances sur les entités (0,341) que le modèle dense (0,307), suggérant une meilleure tolérance des approches lexicales approximatives face aux altérations locales.

L'impact différencié entre entités (NE) et non-entités (non-NE) est frappant. Pour BM25 standard, les performances sur les NE chutent de 63% à 100% de bruit contre seulement 25% pour les non-NE. Le modèle dense amplifie cet écart relatif, confirmant que les entités, pourtant cruciales pour la spécificité des requêtes, sont les plus vulnérables aux erreurs OCR. Cette fragilité s'explique par leur nature souvent plus courte et leur faible redondance contextuelle comparée aux mots courants.

La recherche floue démontre son efficacité mais avec des limites claires : avec 1 erreur par mot, elle stabilise remarquablement les performances sur les entités (perte de seulement 15% contre 63% pour BM25 standard). En revanche, à 5 erreurs par mot, cet avantage disparaît complètement, tous les

TABLE 2 – Performances des systèmes de recherche à différents niveaux de bruit (nDCG@10)

Système	Err./mot	Type bruit	Bruit 0%	Bruit 40%	Bruit 70%	Bruit 100%
BM25	1	NE	0,389	0,355	0,316	0,145
		non-NE	0,397	0,385	0,357	0,297
		ALL	0,378	0,348	0,265	0,017
	5	NE	0,392	0,356	0,304	0,135
		non-NE	0,395	0,382	0,353	0,298
		ALL	0,374	0,346	0,267	0,002
BM25 flou	1	NE	0,401	0,376	0,351	0,341
		non-NE	0,413	0,355	0,323	0,295
		ALL	0,388	0,350	0,315	0,285
	5	NE	0,402	0,365	0,301	0,148
		non-NE	0,401	0,389	0,364	0,300
		ALL	0,391	0,355	0,269	0,001
Dense	1	NE	0,558	0,499	0,432	0,307
		non-NE	0,558	0,524	0,486	0,379
		ALL	0,558	0,475	0,361	0,115
	5	NE	0,558	0,489	0,387	0,202
		non-NE	0,558	0,509	0,467	0,333
		ALL	0,558	0,436	0,311	0,004

systèmes s'effondrant à des niveaux similaires. Le seuil de tolérance des approximations lexicales est donc rapidement atteint.

Les scores globaux (ALL) qui tombent à près de zéro à 100% de bruit avec 5 erreurs par mot, alors que les sous-catégories NE et non-NE conservent des performances non nulles. Cette chute drastique illustre l'effet combiné de la dégradation simultanée des différents composants d'une requête, qui détruit sa spécificité et rend impossible le rapprochement avec les documents pertinents.

Ces résultats suggèrent que la robustesse en recherche d'information pourrait être améliorée par des stratégies de pondération adaptative selon la nature sémantique des termes, accordant plus de flexibilité aux entités tout en maintenant une certaine rigueur sur les non-entités.

4.3 Taux d'erreur sur entités nommées

Le décalage critique où la corruption OCR touche uniquement les entités nommées produit un CER/WER plus faible que la corruption hors entités, mais dégrade bien plus la recherche. Inversement, la corruption hors entités gonfle le CER/WER tout en impactant moins la recherche. Pour mieux capturer l'impact sémantique, nous définissons le taux d'erreur sur entités nommées (NEER) :

$$\text{NEER} = \frac{\text{\# tokens NE corrompus}}{\text{\# total tokens NE}}.$$

Contrairement au CER/WER qui traitent tous les tokens uniformément, le NEER isole les erreurs affectant les termes les plus importants pour la recherche. La figure 2 montre que la performance RI diminue modérément (0,4 à 0,3) quand le WER hors entités augmente jusqu'à 100% si les entités

restent intactes, mais chute brutalement à 0,15 quand les entités sont totalement corrompues, pour seulement 16% de WER au niveau du document.

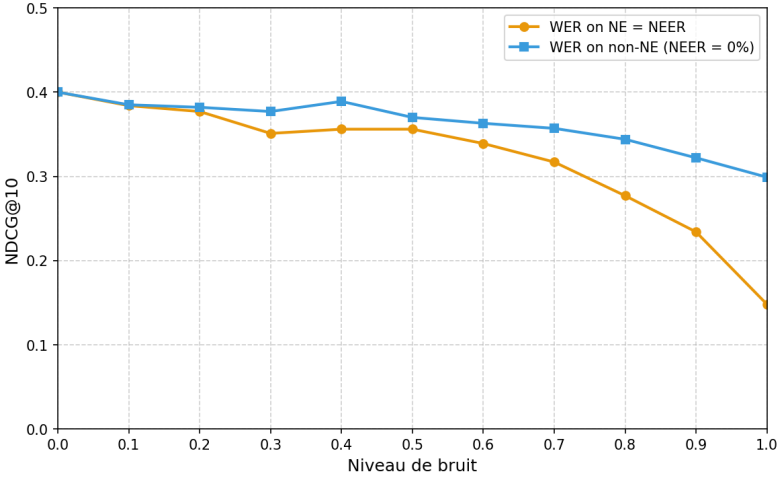

FIGURE 2 – Impact de la qualité OCR dans les performances de recherche d'information (nDCG@10), 1 erreur par mot, avec BM25.

Ces résultats ont des implications directes en recherche d'information : les systèmes devraient prioriser la préservation de l'intégrité des entités via une correction sélective, une indexation pondérée par confiance, ou une recherche floue orientée entités, plutôt que de viser uniquement la réduction globale des erreurs.

5 Conclusion

Les campagnes de numérisation massive ont transformé les documents en ressources numériques exploitables, mais l'utilisabilité de ces collections dépend de la qualité de l'OCR. Les métriques traditionnelles de qualité d'OCR (CER, WER) échouent à capturer cette relation car elles traitent toutes les erreurs uniformément. Nos expériences montrent que les erreurs sur les entités nommées ont un impact très important sur les performances de recherche même à faibles taux d'erreur globaux, tandis que le bruit sur les tokens non-NE a peu d'effet. Nous avons démontré que le taux d'erreur sur entités nommées est un indicateur bien plus fiable que CER/WER.

Nos résultats appellent à prioriser la préservation des entités dans le post-traitement OCR, l'indexation pondérée par la confiance, et des approches hybrides combinant méthodes lexicales et sémantiques. Notre étude ouvre plusieurs perspectives prometteuses, notamment le développement de modèles de correction OCR spécifiquement dédiés à la préservation des entités nommées, ainsi que l'extension de notre cadre d'évaluation aux erreurs de mise en page et de segmentation qui affectent fréquemment les documents historiques. La conception de modèles de recherche capables d'exploiter dynamiquement les signaux de confiance OCR et les sorties des systèmes de reconnaissance d'entités nommées pourra significativement améliorer la robustesse des pipelines existants et contribuer à rendre les collections

numérisées véritablement exploitables pour la recherche et le grand public.

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

AKBIK A., BERGMANN T., BLYTHE D., RASUL K., SCHWETER S. & VOLLGRAF R. (2019). FLAIR : An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, p. 54–59.

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.

BOROS E., NGUYEN N. K., LEJEUNE G. & DOUCET A. (2022). Assessing the impact of OCR noise on multilingual event detection over digitised documents. *International Journal of Digital Libraries*, **23**(3), 241–266.

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.

GEFEN A. (2014). Les enjeux épistémologiques des humanités numériques. *Socio-La nouvelle revue des sciences sociales*, (4), 61–74.

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., DOUCET A. *et al.* (2024a). Retrieval augmented generation for historical newspapers. In *ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*.

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

GUO S., WANG Y., YE J., ZHANG A., ZHANG P. & XU K. (2025). Semantic importance-aware communications with semantic correction using large language models. *IEEE Transactions on Machine Learning in Communications and Networking*.

GUPTA A., GUTIERREZ-OSUNA R., CHRISTY M., CAPITANU B., AUVIL L., GRUMBACH L., FURUTA R. & MANDELL L. (2015). Automatic assessment of ocr quality in historical documents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

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.

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.

LEHMANN J., ISELE R., JAKOB M., JENTZSCH A., KONTOKOSTAS D., MENDES P. N., HELLMANN S., MORSEY M., VAN KLEEF P., AUER S. *et al.* (2015). Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, **6**(2), 167–195.

LOPRESTI D. (2005). Performance evaluation for text processing of noisy inputs. In *Proceedings of the 2005 ACM symposium on Applied computing*, p. 759–763 : ACM.

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.

MARINAI S., MIOTTI B. & SODA G. (2011). Digital libraries and document image retrieval techniques : A survey. In *Learning Structure and Schemas from Documents*, p. 181–204. Springer.

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.

NGUYEN T. T. H., JATOWT A., NGUYEN N.-V., COUSTATY M. & DOUCET A. (2020). 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*, p. 333–336.

OARD D. W. (2003). Balanced query methods for improving ocr-based retrieval. In *Proceedings 2003 Symposium on Document Image Understanding Technology*, p. 181 : UMD.

RAO J., WANG F., DING L., QI S., ZHAN Y., LIU W. & TAO D. (2022). Where does the performance improvement come from ? -a reproducibility concern about image-text retrieval. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, p. 2727–2737.

SANG E. F. & DE MEULDER F. (2003). Introduction to the conll-2003 shared task : Language-independent named entity recognition. *arXiv preprint cs/0306050*.

TAGHVA K., BORSACK J. & CONDIT A. (1996). Effects of ocr errors on ranking and feedback using the vector space model. *Information processing & management*, **32**(3), 317–327.

THAKUR N., REIMERS N., RÜCKLÉ A., SRIVASTAVA A. & GUREVYCH I. (2021). BEIR : A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

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

TRAN T., GONZÁLEZ-GALLARDO C.-E. & DOUCET A. (2025). Génération augmentée de récupération pour les journaux historiques. In *Actes de la 20e Conférence en Recherche d'Information et Applications (CORIA)*, p. 131–134.

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.

WANG X., JIANG Y., BACH N., WANG T., HUANG Z., HUANG F. & TU K. (2021a). Automated Concatenation of Embeddings for Structured Prediction. In *the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (**ACL-IJCNLP 2021**)* : Association for Computational Linguistics.

WANG X., JIANG Y., BACH N., WANG T., HUANG Z., HUANG F. & TU K. (2021b). Improving Named Entity Recognition by External Context Retrieving and Cooperative Learning. In *the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (**ACL-IJCNLP 2021**)* : Association for Computational Linguistics.

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.

YAMADA I., ASAI A., SHINDO H., TAKEDA H. & MATSUMOTO Y. (2020). LUKE : Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.523.

ZHANG J., ZHANG Q., WANG B., OUYANG L., WEN Z., LI Y., CHOW K.-H., HE C. & ZHANG W. (2025). Ocr hinders rag : Evaluating the cascading impact of ocr on retrieval-augmented generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, p. 17443–17453.