Reconnaissance d'entités nommées sur des sorties OCR bruitées : des pistes pour la désambiguïsation morphologique automatique

Caroline Koudoro-Parfait, Gaël Lejeune, Richy Buth

Reconnaissance d'entités nommées sur des sorties OCR bruitées : des pistes pour la désambiguïsation morphologique automatique.

Caroline Koudoro-Parfait[1, 2, 3] Gaël Lejeune[2] Richy Buth[2]
(1) ObTIC, Sorbonne Université, 1 Rue Victor Cousin, 75005 Paris, France
(2) STIH, Sorbonne Université, 1 Rue Victor Cousin, 75005 Paris, France
(3) SCAI, Campus Pierre et Marie Curie, 4 place Jussieu 75005 Paris, France
caroline.parfait@sorbonne-universite.fr,
gael.lejeune@sorbonne-universite.fr, buth_richy@hotmail.com

RÉSUMÉ

La variation dans les données textuelles, en particulier le bruit, est un facteur limitant la performance des systèmes de Reconnaissance d'Entités Nommées (REN). Les systèmes de REN sont en effet généralement entraînés sur des données « propres », non-bruitées, ce qui n'est pas le cas des données des humanités numériques obtenues par reconnaissance optique de caractères (OCR). De fait, la qualité des transcriptions OCR est souvent perçue comme la source principale des erreurs faites par les outils de REN. Cependant, des résultats obtenus avec différents systèmes REN sur des transcriptions OCR d'un corpus du 19ème siècle (ELTeC) tendent à montrer une certaine robustesse, modulo la présence de formes bruitées, parfois dites « contaminées ». La difficulté, est alors de lier ces formes contaminées avec leur forme de référence, par exemple, pour rapprocher la chaîne « Parisl »et la chaîne « Paris ». Il s'agit de modéliser le fait que différentes variations se rapprochent du même terme. Des questions quant à l'automatisation de cette tâche et sa généralisation à toutes les variations d'un même terme restent ouvertes. Nous montrons dans cet article différentes expériences visant à traiter ce problème sous l'angle de la désambiguïsation morphologique des entités nommées (EN) en aval de la chaîne de traitement, plutôt que par la correction en amont des données de l'OCR.

ABSTRACT

Resolution of entity linking issues on noisy OCR output : automatic disambiguation tracks.

Textual variability is perceived as a significant limitation to the performance of Named Entity Recognition (NER) systems. NER systems are trained on clean data, which is not the case for OCR corpora. In fact, the quality of OCR transcriptions is often perceived as the main source of errors made by NER tools. However, results obtained with different REN systems on OCR transcriptions of a 19th century corpus (ELTeC) tend to show a certain robustness, modulo the presence of so-called "contaminated" forms. The difficulty, from now on, is to link the contaminated forms with their reference form, for example, to bring "Parisl" and "Paris" together. It is a question of modeling the fact that different variations come close to the same term. Questions about the automation of this task and its generalization to all variations of the same term remain open. Our first approach is to deal with this problem from the perspective of morphological disambiguation of named entities.

MOTS-CLÉS : Reconnaissance d'entités nommées, Reconnaissance optique de caractères, .

KEYWORDS: Named entity recognition, Optical character recognition.

1 Introduction

La numérisation en masse des fonds des bibliothèques a mis à disposition des chercheurs des données textuelles en grande quantité. Aujourd'hui, ces communautés sont confrontées à la nécessité d'extraire des informations de ces documents numérisés afin de les rendre interrogeables. Selon (Linhares Pontes *et al.*, 2019) et (Hamdi *et al.*, 2020), la Reconnaissance d'entités nommées (REN) est une technologie clé pour accéder aux connaissances contenues dans ces vastes corpus. En effet, la majorité des requêtes des utilisateurs comportent au moins une entité nommée (EN), en particulier des noms de lieux (van Strien *et al.*, 2020). De sorte que l'identification des EN serait un moyen efficace d'améliorer l'accès à l'information et de valoriser les données préalablement numérisées. Par ailleurs, la désambiguïsation des toponymes est un enjeu important (variation linguistique, diachronique ou diatopique, entre autres exemples) pour pouvoir établir des rapprochements entre des textes faisant partie de sous-corpus différents (période historique, qualité d'impression ou de numérisation ...).

Le problème patent, auquel les chercheurs sont confrontés, est d'appliquer les outils informatiques existants, généralement entraînés sur des données textuelles correctement orthographiées (Eshel *et al.*, 2017), à des textes bruités par la numérisation, puis par l'application de la reconnaissance optique de caractères (OCR). Le bruit désigne toutes les erreurs produites par le système OCR : l'insertion, la suppression, mais aussi la substitution d'un ou plusieurs caractères. Les outils automatiques produisent certaines erreurs de manière systématique (Stanislawek *et al.*, 2019) et l'agent humain peut alors les modéliser et produire un programme de correction automatique. Néanmoins, lorsqu'il s'agit d'erreurs singulières un tel dispositif reste difficile à mettre en place. Dans de nombreux cas, les erreurs commises par les systèmes de REN sont attribuées au caractère bruité des sorties OCR, ce qui suggère de corriger les données en entrée pour améliorer les résultats obtenus en sortie. Pour pallier le problème lié à la qualité des transcriptions OCR les utilisateurs mettent en place des stratégies de nettoyage des textes. Néanmoins, comme le souligne (Huynh *et al.*, 2020), s'il est effectivement possible d'améliorer les résultats de REN en corrigeant automatiquement les sorties OCR, cette correction peut produire ses propres erreurs. Les performances des outils de REN présentent encore des limites en raison des variations des sources figurant en entrée, tant dans leur qualité textuelle (transcription OCR ou ou d'écriture manuscrite - HTR) que dans leur genre (littéraire, critique) ou leur ancienneté.

Dès lors, la question principale concerne l'évaluation de l'incidence des erreurs d'OCR sur la reconnaissance d'entités nommées spatiales (Baledent *et al.*, 2020), et l'influence de ce bruit sur les usages consécutifs (van Strien *et al.*, 2020) de ses données. Dans le même temps, (Koudoro-Parfait *et al.*, 2021) produisent une analyse automatique des sorties d'outils de REN tels que SPACY (Honnibal & Montani, 2017). Ils démontrent que les outils de REN prêts à l'emploi sont capables de repérer des EN spatiales, y compris sous des formes contaminées (Hamdi *et al.*, 2022), c'est-à-dire des formes impactées par une ou plusieurs erreurs d'OCR. Leurs analyses s'appuient sur le corpus ELTeC en français et comparent la REN sur cette version de référence, aux sorties obtenues par transcription OCR. Ils utilisent des mesures de distances comme Jaccard et Cosinus et calculent ensuite les intersections entre l'ensemble des entités de la référence et celui des entités de la version OCR. Cette analyse reste toutefois limitée à un point de vue ensembliste, qui ignore donc l'effectif des différentes entités et ne rapproche pas les formes contaminées des formes de référence, du fait de problèmes d'alignement entre les tokens de la version de référence et les tokens obtenus par l'OCR. Au contraire, nous proposons de produire une analyse automatique plus précise et de résoudre les problèmes d'alignements entre les différentes sorties en utilisant l'outil NERVAL [1].

1. https://gitlab.com/teklia/nerval

La suite de cette contribution est organisée de la façon suivante : dans la section 2, nous proposons une revue de la littérature sur la désambiguïsation des entités nommées sur des transcriptions OCR bruitées, puis dans la section 3, nous présentons le corpus littéraire en français et les outils utilisés pour notre étude, puis nous proposons une analyse des résultats de REN qui vient dépasser la problématique d'alignement pour l'évaluation F-score ainsi que des pistes pour la désambiguïsation dans la section 4, enfin nous exposons nos perspectives d'utilisation des résultats de désambiguïsation dans la section 5.

2 Désambiguïsation des EN extraites de transcriptions OCR

(Lopresti, 2009) démontrent que le bruit de l'OCR a un impact plutôt négatif sur les tâches de traitement automatique des langues (TAL) réalisées en aval. Si les résultats récents de (van Strien *et al.*, 2020) vont dans ce sens, pour les tâches de segmentation de phrases et d'analyse syntaxique en dépendance, il semblerait que cet impact soit plus faible sur la REN. Dans le même temps, (Hamdi *et al.*, 2020) et (Hamdi *et al.*, 2022) démontrent que s'il existe bien une relation entre la dégradation de la qualité de la transcription OCR et la perte de qualité de la REN, celle-ci, sans être négligeable, n'est pas non plus catastrophique. La tâche de REN appliquée à des textes anciens et transcrits par OCR est une des manières d'aborder la problématique actuelle du traitement des textes bruités (Boros *et al.*, 2020), (Ehrmann *et al.*, 2021). Du côté des humanités numériques spatialisées, (Koudoro-Parfait *et al.*, 2021) ont évalué différents systèmes de REN sur des transcriptions de textes romanesques, des 19ème et 20ème siècles, obtenues via différents outils OCR (Kraken, Tesseract). Ils soulignent le fait que les outils de REN, prêts à l'emploi, sont plutôt efficaces sur des textes bruités et que le problème se situe plus sur le liage entre les formes correctes et les formes contaminées. L'enjeu pour l'utilisateur est désormais d'avoir un outil de désambiguïsation automatique de ces formes variantes pour les lier à la forme d'origine de l'EN et obtenir des sorties exploitables et publiables dans les éditions savantes.

(Bousmaha *et al.*, 2013) soulignent que la tâche de désambiguïsation des entités nommées (Named Entity Disambiguation, NED) comporte plusieurs versants, d'une part la désambiguïsation morphologique qui consiste à lier deux mots qui ont la même morphologie ou une morphologie proche, d'autre part, la désambiguïsation lexicale (Word Sense Disambiguation - WSD en anglais) (Ehrmann, 2008) qui consiste à choisir le sens pertinent dans le contexte d'apparition. Enfin, la désambiguïsation sémantique qui consiste à rapprocher des termes de leur sens, de définition et d'article dans des bases de données (Moro *et al.*, 2014). La tâche de désambiguïsation sémantique, aussi appelée Named Entity Linking (NEL) est dépendante des deux autres types. (Brando *et al.*, 2016) la définisse comme une tâche spécifique de celle de NED parce qu'elle permet l'enrichissement sémantique des textes.

En TAL, la tâche de désambiguïsation automatique est couramment définie comme une tâche de classification consistant à associer différentes formes rencontrées dans des textes à des formes de référence. Autrement dit, il s'agit d'associer des variantes à un terme vedette. Cette association peut se faire de façon supervisée, si les termes vedettes sont déjà connus, car présents dans une ressource de référence, une base de données ou encore calculés hors ligne. Elle est au contraire non supervisée, si les termes vedettes sont découverts directement dans le corpus, s'ils sont calculés en ligne. Dans les deux cas, il est question de désambiguïsation puisque plusieurs formes font référence à la même entité et qu'une même forme pourrait être rapprochée de différentes entités candidates figurant dans un index, un gazetteer ou tout simplement d'autres entités découvertes dans le corpus.

L'utilisation des bases de données, des bases de connaissance et des ontologies, visent principalement

la désambiguïsation lexicale (WSD) et sémantique (NEL) des entités (Shen *et al.*, 2015). (Hoffart *et al.*, 2011) rapportent que les premières expériences de désambiguïsation s'appuyant sur les bases de données comme Wikipédia remontent aux années 2000 ((Bunescu & Paşca, 2006), (Milne & Witten, 2008), (Cucerzan, 2007) et (Nguyen & Cao, 2008) notamment). La majorité des expériences emploient des mesures de similarité et des poids pour déterminer la corrélation entre les termes trouvés et les termes des bases de données. Parmi les outils de désambiguïsation et liage sémantique, on trouve SOFIE [2] qui s'appuie sur la base de connaissance YAGO (Suchanek *et al.*, 2007) ou REDEN (Brando *et al.*, 2016) opérationnel dans les champs de la critique littéraire et la littérature scientifique qui s'appuie sur un système de graphe de connaissances.

Les approches supervisées nécessitent l'accès à une grande quantité de données annotées (Moro *et al.*, 2014) et de pouvoir s'appuyer pour l'évaluation sur un Gold Standard (étalon) , ce qui n'est pas toujours adapté à l'évaluation des systèmes de REN sur des textes bruités. En outre, la question de l'annotation de corpus en vue de l'entrainement des modèles pour la REN soulève aussi la question de la (sur)adaptation de ces modèles au type de corpus pour lequel ils ont été préparés (Vigier *et al.*, 2020). Aujourd'hui, certains jeux de données standardisés de référence existent et ont été enrichis avec le temps, CoNLL (CoNLL-02,CoNLL-03, AIDA-CoNLL-YAGO (Hoffart *et al.*, 2011)) et ACE (ACE2004 (Guo & Barbosa, 2014)) qui sont sans doute les plus utilisés. Malgré tout, la tâche reste encore complexe à mettre en place sur des données non standards et bruitées. (Eshel *et al.*, 2017) utilisent le jeu de données WIKILINKSNED conçue à partir du corpus WIKILINKS, pour pratiquer la désambiguïsation sur un corpus de textes courts bruités issus du web.

Pour tenter de dépasser les verrous dus à la conformité des jeux de données standards, certaines équipes s'intéressent aux approches non supervisées qui utilisent des systèmes à réseaux de neurones (convolutional neural network, CNN) (Sun *et al.*, 2015) pour apprendre au système des similarités entre le contexte, l'entité et l'entité candidate. Ces dernières années, avec la mise en avant des travaux de (Mikolov *et al.*, 2013), des systèmes d'alignements fondés sur des plongements de mots ou word embeddings ont aussi été élaborés (Yamada *et al.*, 2016). La principale qualité des systèmes fondés sur des word embeddings est de permettre l'alignement des entités par une comparaison de similarité entre les contextes d'apparitions de certain terme et de pouvoir déterminer une plus ou moins forte probabilité qu'il s'agisse du même terme. Si ces systèmes sont capables de reconnaitre qu'une chaine de caractères est identique à une autre, ils ne détectent pas encore la polysémie potentielle et connaissent une limitation dans leur application automatique pour la NEL, ils ne permettent pas à eux seule la désambiguïsation sémantique (Cuxac *et al.*, 2019). Néanmoins, des systèmes non supervisés comme celui de (Le & Titov, 2018) peuvent être entrainés et utilisés pour l'aide à la prise de décision dans la désambiguïsation des entités nommées contaminées par les erreurs d'OCR avec d'assez bons résultats, comme le démontrent (Hamdi *et al.*, 2022).

3 Jeux de données et systèmes de REN et d'OCR utilisés

À des fins de comparaison, nous empruntons le même corpus que (Koudoro-Parfait *et al.*, 2021) [3], qui compte une dizaine de romans français issus de la Collection européenne de textes littéraires - ELTeC [4]. Le choix d'un corpus français nous est paru pertinent, car il permet de tester les outils prêts

2. https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/sofie
3. Détails sur le corpus et les résultats sur https://github.com/These-SCAI2023/NER_GEO_COMPAR
4. Collection européenne de textes littéraires : https://www.distant-reading.net/eltec/

à l'emploi sur une langue différente de l'anglais. Nous travaillons sur les versions OCR générées avec Kraken[5], et Tesseract[6] (Smith, 2007). Kraken est un outil en cours de développement qui propose des interfaces pratiques pour les chercheurs en Humanités Numériques (Kiessling *et al.*, 2019) avec des modèles entrainés pour plusieurs langues (Weichselbaumer *et al.*, 2020) ou variantes, par exemple le français du 17ème siècle (Gabay *et al.*, 2020). Tesseract est un outil paramétrable pour des tâches particulières, pour plusieurs langues, mais qui présente de très bonnes performances dans sa configuration par défaut (Clausner *et al.*, 2020). Pour chaque roman, nous aurons donc quatre versions : la version de référence de ELTeC et trois transcriptions OCR différentes : Kraken, Tesseract par défaut (Tess) et Tesseract pour le français (Tess_fr).

Nous avons appliqué trois modèles SPACY pour le français[7] sur les quatre versions de chaque roman (la version originale et les trois versions OCR). SPACY est sans doute l'outil de REN le plus communément présenté sur les tutoriels en ligne, ce qui en fait un bon banc d'essai pour les problèmes qu'un chercheur en Humanités Numériques (HN) peut rencontrer. SPACY n'a pas été conçu dans le but de servir les domaines de la recherche académique et de l'édition savante, néanmoins (van Strien *et al.*, 2020) le décrivent comme un outil facilement utilisable par les chercheurs en HN et les professionnels des bibliothèques et de la culture. Dans le but de surmonter les problèmes d'alignement évoqués par (Koudoro-Parfait *et al.*, 2021), nous avons produit des sorties de SPACY au format IOB que nous avons pu exploiter avec l'outil NERVAL[8]. NERVAL est un package python conçu pour l'évaluation de la REN sur du texte bruité, typiquement pour des textes transcrits par OCR ou manuscrit (HTR). Le système s'appuie sur la distance de Levenshtein pour la comparaison de similarité et l'alignement.

4 Évaluation automatique de la reconnaissance d'entités nommées sur des transcriptions OCR imparfaites

4.1 Alignement en IOB pour différentes transcriptions OCR avec Nerval

Les entrées attendues par Nerval sont un fichier de vérité de terrain et un fichier de prédiction annotés au format IOB (Ramshaw & Marcus, 1995). Le format IOB (abréviation de inside, outside, begin) sert, entre autres, au balisage des tokens dans les tâches de REN. Le préfixe B- avant une balise indique que la balise est le début d'un bloc de tokens formant une entité nommée. Le préfixe I- indique que la balise est à l'intérieur de ce bloc. Une balise O indique qu'un token n'appartient à aucun bloc. La documentation de NERVAL précise que les documents ne doivent contenir aucune occurrence de "§", ce caractère ayant une signification particulière lors de l'évaluation, mais nous avons aussi remarqué que d'autres caractères ont posé problèmes durant nos premières expériences[9]. Ce problème a été surmonté en changeant le paramètre d'encodage par défaut de NERVAL en UTF-8.

Avant d'opérer l'alignement avec NERVAL, nous avons converti nos sorties SPACY au format d'annotation IOB. NERVAL opère ensuite l'alignement des balises des EN de la référence et des différentes transcriptions, caractère par caractère. Lors de l'alignement, il est nécessaire que les chaînes de caractères alignées aient la même taille, de ce fait NERVAL ajoute le caractère "-" (Tableau 1[a]), qui,

5. https://github.com/mittagessen/kraken
6. https://github.com/tesseract-ocr/tesseract
7. small (fr_core_news_sm), medium (fr_core_news_md) et large (fr_core_news_lg)
8. https://teklia.com/blog/202104-nerval/
9. Quelques uns des caractères problématiques pour NERVAL : à, ç, Ç, ï, fi, fl, “, ”,

Versions	Entités Réf.^{I}	Entités Align	Entités Réf.^{II}	Entités Align
Kraken		—^{a} merique		**eares**^{c}
Tess fr	l'Amérique	**dé**Amérique^{b}	Paris	**-aran**^{c}
Tess		–Amérique		**aran**^{c}
Kraken		Rio Gila		Laou^{d}
Tess fr	rio Gila	Bio Gz–	Laon	Laon
Tess		rio Gila		Laon

TABLE 1 – Alignement des entités (`spacy_lg`) de la référence et des versions OCR avec NERVAL, ^{I}*"Les trappeurs de l'Arkansas"*, Aimard, 1858 et ^{II}*"Le château de Pinon, vol. I.", Dash, 1844*

pendant l'étape d'alignement, prend la balise du caractère qui le précède dans la chaine de caractères. Le système cherche, ensuite, à déterminer le périmètre de l'entité (sa longueur). Lorsqu'il repère une nouvelle balise signalant une entité (P pour Personne, L pour Location par exemple), ce caractère est considéré comme le début de l'entité. Ensuite, le système compare les balises des deux caractères (annotation et prédiction) et s'ils se correspondent, le système retrace les balises dans la chaîne de prédiction pour détecter la première occurrence de l'étiquette marquant l'entité. Si le système ne trouvent pas de correspondance entre les balises des premiers caractères de l'entité recherchée et celles de la prédiction, il cherche jusqu'à la fin de l'entité dans l'annotation. Dans tous les cas, la dernière occurrence de la balise correspondant au premier caractère de l'entité détermine la fin de l'entité. En sortie de NERVAL, l'utilisateur n'a pas directement accès aux alignements, nous avons procédé à quelques modifications afin d'extraire les alignements (Tableau 1) et les observer. Pour exposer nos exemples de manière lisible, nous avons reconstitué les tokens. Nous notons bien que NERVAL permet l'alignement de termes qui pour un système de comparaison stricte ne seraient pas alignés, par exemple "Bio Gz–" (erreur OCR) et "rio Gila". Plus étonnant, il semble que du fait de la recherche du périmètre de l'entité par NERVAL, le système : i) prend des caractères d'un autre token et les ajoute au token suivant ^{b}, ii) concatène des caractères qui ne sont pas une seule et même chaine de caractères dans nos entrées^{c}. Nous déduisons ces deux points de manière heuristique, par comparaisons manuelles de nos sorties. i) Nous supposons, qu'il s'agit, dans l'exemple ^{b}, du couple de caractère "dé" du terme "décou-verte" qui apparait avant l'entité "Amérique". ii) Comme nous ne comprenions pas la présence des termes ^{c}, qui ne figurent ni dans les transcriptions OCR (il ne s'agit pas de bruit OCR), ni dans les sorties de NER comme tels, nous avons expliqué ces présences par des erreurs de calcul de la longueur de l'entité par NERVAL. Concernant, l'exemple, ^{d} nous nous sommes aussi interrogés quant à la génération de termes par le système, afin que l'alignement soit possible. En effet, la référence comporte 19 occurrences de « laon »et la version OCR 15, mais aucune ne correspond à "laou". La documentation à laquelle nous avons eu accès ne précise pas ce dernier point.

4.2 Évaluation avec NERVAL

Après avoir effectué l'alignement entre les termes de la référence et ceux de l'hypothèse, NERVAL propose de calculer la similarité des chaines de caractères appariées entre elles, avec la distance de Levenshtein. Si la distance est inférieure à 0.3, NERVAL considère l'entité comme reconnue. Dans leurs analyses (Koudoro-Parfait *et al.*, 2021) mentionnent que le meilleur système OCR est Tesseract français. Nous observons que les résultats F_1 *mesure* (Tableau 2 et Tableau 3) sont en effet supérieurs pour ce modèle. D'autre part, l'équipe a mis à disposition ses résultats en accès libre et nous avons pu observer que les schémas de distances pour le texte d'Alphonse Daudet montre qu'il s'agit d'un

	#Entités		Évaluation par NERVAL			
Version	Version OCR	Référence	Intersection	Précision	Rappel	F_1 mesure
Kraken	*1391*	965	576	*0.414*	*0.597*	0.489
Tess fr	980	965	713	0.728	0.739	**0.733**
Tess	1090	965	608	0.558	0.630	0.592

TABLE 2 – Comparaison des résultats de la reconnaissance d'entités nommées avec `spacy_lg` sur différentes versions de *"Le petit chose"*, Daudet, 1868, après alignement avec NERVAL

	#Entités		Évaluation par NERVAL			
Version	Version OCR	Référence	Intersection	Précision	Rappel	F_1 mesure
Kraken	*1144*	204	69	*0.060*	*0.338*	0.102
Tess fr	441	204	147	0.333	0.631	**0.456**
Tess	577	204	110	0.191	0.539	0.282

TABLE 3 – Comparaison des résultats de la reconnaissance d'entités nommées avec `spacy_lg` sur différentes versions de *"La petite Jeanne"*, Carraud, 1884, après alignement avec NERVAL

texte pour lequel l'OCR et la NER se sont bien déroulés. Nos résultats tendent à confirmer cette hypothèse. Par ailleurs, nous observons que l'OCR Kraken a connu des dysfonctionnements sur le texte de Zulma Carraud. Il apparait que `spacy_lg` a récupéré 1144 entités pour ce texte, pour 204 dans la référence, ce compte corrobore la thèse selon laquelle plus une sortie OCR est bruitée plus l'outil de REN récupèrera des résultats en très grande quantité, les mots inconnus semblant facilement étiquetées comme EN. Les résultats du tableau 3 montrent que `spacy_lg` amène beaucoup de faux Positifs. Ainsi, la combinaison de Kraken et SPACY n'est pas idéale. L'entièreté de nos résultats avec NERVAL sur ce corpus sont accessibles en ligne [10].

4.3 Des pistes pour une désambiguïsation automatique

De notre côté, nous avons conçu un outil [11], qui permet à l'utilisateur, de récupérer les différentes formes contaminées d'un même terme en exploitant les distances de Jaccard ou Cosinus. Afin, de procéder au calcul de similarité, nous transformons nos listes d'entités en ensembles, ce qui réduit la quantité des comparaisons, puis, nous comparons chaque entité de l'ensemble de référence aux entités de la transcription OCR, avec une représentation en bigrammes de caractères qui permet de modéliser la séquentialité dans les entités (par opposition à une approche en unigrammes) tout en conservant une certaine robustesse à l'insertion de caractères bruités dans l'entité (par opposition à une approche en quadrigrammes par exemple). De manière heuristique, et après avoir comparé les résultats obtenus avec Jaccard et Cosinus (Tableau 4), nous avons décidé d'utiliser la métrique Cosinus pour établir le seuil permettant de discriminer les formes contaminées des autres entités. Il semble que la distance Cosinus soit plus flexible et mette moins de poids sur les différences au grain caractère que la distance de Jaccard. Néanmoins, la distance Cosinus paraît sensible à un trop grand écart dans la différence entre le nombre des caractères des EN des deux groupes qu'elle compare. Nous avons fixé le seuil à une distance 0,35, ce qui tend à rejoindre au ratio utilisé pour la distance de Levenshtein dans NERVAL.

10. https ://github.com/anonymous
11. https ://github.com/anonymous

Versions	Entités Réf.	Entités Align	Jaccard	Cosinus
Kraken	Morlincourt	Mlorlincourt	0.1428	0.0715
		Mlorlincourtl	0.1818	0.1210
Tess fr	Morlincourt	Morlinco'urt	0.1818	0.0762
		Morlin	**0.4761**	**0.2788**
Kraken	Saint-Brunelle	Bruncll e	**0.5925**	**0.3244**
		Brunelle	0.4583	0.2012
Tess fr	Saint-Brunelle	Saint—Brunelle	0.2222	0.0909
		Saint—anelle	0.4642	0.2183

TABLE 4 – Résultats de la récupération automatique des formes contaminées récupérées par `spacy_lg` sur différentes versions de *"Mon village"*, Adam, 1860.

5 Conclusion et perspectives d'utilisation

Dans cette contribution, nous avons proposé une évaluation automatique de la qualité des sorties de REN sur un corpus de textes OCR bruités en utilisant l'outil NERVAL. Nous avons comparé les résultats de SPACY sur 4 versions de onze textes littéraires avec dans chaque cas une version de référence tirée du corpus ELTeC et trois transcriptions OCR (Tesseract, Tesseract Français et Kraken). Nous avons observé que les modèles de REN parviennent à détecter les entités nommées spatiales sur des textes issus d'OCR en quantité importante. Bien sûr, les comparaisons automatiques des résultats obtenus sur le texte de référence et sur ses transcriptions OCR indiquent qu'il existe des différences significatives entre ces différentes versions. Ces différences sont en partie dues au bruit dans le texte d'entrée donné aux systèmes REN, mais ces systèmes parviennent également à reconnaître des variantes orthographiques d'un terme qui est bien une entité nommée spatiale et que nous nommons entités contaminées, expression proposée par (Hamdi *et al.*, 2022).

Nous avons pu observer le fonctionnement de NERVAL pour l'évaluation des EN. Nous avons constaté d'une part que le système procède à l'alignement entre les balises IOB des entités de référence et les entités hypothèses, d'autre part, et bien que cet outil soit plutôt efficace, nous avons consigné quelques problèmes d'usage concernant la définition du périmètre des entités qui fait que certaines entités de la référence sont alignées sur des termes qui ne sont en fait pas des sorties REN, mais des assemblages de caractères précédents ou suivants l'entité que le système cherchent à aligner. Nous nous interrogeons donc sur l'existence d'un biais possible dans les évaluations NERVAL. Par ailleurs, nous avons vu qu'il est très instructif d'observer plusieurs mesures de distance, car elles ne vont pas toutes pénaliser les phénomènes de présence et d'absence (bruit et silence) de la même manière.

Dans le futur, nous voulons explorer trois directions différentes. Premièrement, mesurer le bruit dans les résultats en annotant manuellement différents résultats de REN. Deuxièmement, œuvrer à la résolution des problèmes d'alignement et de désambiguïsation des entités (diachronie, formes contaminées). Enfin, nous souhaitons améliorer la géolocalisation automatique des lieux, car en parallèle de ces travaux concernant l'évaluation des sorties de REN sur des OCR bruités, la désambiguïsation morphologique des EN contaminées pourra nous permettre d'effectuer un alignement automatique entre ces formes et des index géographiques. Ceci, nous permettra de produire des cartes présentant l'évocation des lieux dans la littérature française du 19ème siècle. Ainsi, pourrons-nous tenter de déterminer, entre autres, si les auteurs issus du romantisme et inspirés par l'orient évoquent ces parties du monde plus que l'Europe dont ils sont originaires.

Références

BALEDENT A., HIEBEL N. & LEJEUNE G. (2020). Dating Ancient texts : an Approach for Noisy French Documents. In *Language Technologies for Historical and Ancient Langues (LT4HLA) @LREC2020*. HAL : hal-02571633.

BOROS E., HAMDI A., LINHARES PONTES E., CABRERA-DIEGO L. A., MORENO J. G., SIDERE N. & DOUCET A. (2020). Alleviating digitization errors in named entity recognition for historical documents. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, p. 431–441, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.conll-1.35.

BOUSMAHA K., CHAREF-ABDOUN S., HADRICH BELGUITH L. & RAHMOUNI M. (2013). Une approche de désambiguïsation morpho-lexicale évaluée sur l'analyseur morphologique alkhalil. *Revue de l'Information Scientifique et Technique*, **21**(1), 26–40.

BRANDO C., FRONTINI F. & GANASCIA J.-G. (2016). REDEN : Named Entity Linking in Digital Literary Editions Using Linked Data Sets. *Complex Systems Informatics and Modeling Quarterly*, (7), 60 – 80. DOI : 10.7250/csimq.2016-7.04, HAL : hal-01396037.

BUNESCU R. & PAŞCA M. (2006). Using encyclopedic knowledge for named entity disambiguation. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, p. 9–16, Trento, Italy : Association for Computational Linguistics.

CLAUSNER C., ANTONACOPOULOS A. & PLETSCHACHER S. (2020). Efficient and effective ocr engine training. *International Journal on Document Analysis and Recognition (IJDAR)*, **23**(1), 73–88.

CUCERZAN S. (2007). Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, p. 708–716, Prague, Czech Republic : Association for Computational Linguistics.

CUXAC P., COLLIGNON A., GREGORIO S. & PARMENTIER F. (2019). Des bases de données massives au Web de données : désambiguïsation et alignement d'entités géographiques dans les textes scientifiques. In *12ème Colloque international d'ISKO-France : Données et mégadonnées ouvertes en SHS : de nouveaux enjeux pour l'état et l'organisation des connaissances ?*, Montpellier, France. HAL : hal-02307577.

EHRMANN M. (2008). *Les Entitées Nommées, de la linguistique au TAL : Statut théorique et méthodes de désambiguïsation*. Theses, Paris Diderot University. HAL : tel-01639190.

EHRMANN M., HAMDI A., PONTES E. L., ROMANELLO M. & DOUCET A. (2021). Named entity recognition and classification on historical documents : A survey. DOI : 10.48550/ARXIV.2109.11406.

ESHEL Y., COHEN N., RADINSKY K., MARKOVITCH S., YAMADA I. & LEVY O. (2017). Named entity disambiguation for noisy text. DOI : 10.48550/ARXIV.1706.09147.

GABAY S., CLÉRICE T. & REUL C. (2020). OCR17 : Ground Truth and Models for 17th c. French Prints (and hopefully more). working paper or preprint.

GUO Z. & BARBOSA D. (2014). Robust entity linking via random walks. In J. LI, X. S. WANG, M. N. GAROFALAKIS, I. SOBOROFF, T. SUEL & M. WANG, Éds., *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014, Shanghai, China, November 3-7, 2014*, p. 499–508 : ACM. DOI : 10.1145/2661829.2661887.

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 *Digital Libraries for Open Knowledge 24th International Conference on Theory and Practice of Digital Libraries, TPDL 2020, Lyon, France, August 25–27, 2020, Proceedings*, p. 87–101. DOI : 10.1007/978-3-030-54956-5_7, HAL : hal-03026931.

HAMDI A., LINHARES PONTES E., SIDÈRE N., COUSTATY M. & DOUCET A. (2022). In-Depth Analysis of the Impact of OCR Errors on Named Entity Recognition and Linking. *Natural Language Engineering*. DOI : 10.1017/S1351324922000110, HAL : hal-03615997.

HOFFART J., YOSEF M. A., BORDINO I., FÜRSTENAU H., PINKAL M., SPANIOL M., TANEVA B., THATER S. & WEIKUM G. (2011). Robust disambiguation of named entities in text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, p. 782–792, Edinburgh, Scotland, UK. : Association for Computational Linguistics.

HONNIBAL M. & MONTANI I. (2017). spacy 2 : Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, **7**(1), 411–420.

HUYNH V.-N., HAMDI A. & DOUCET A. (2020). When to Use OCR Post-correction for Named Entity Recognition ? In *22nd International Conference on Asia-Pacific Digital Libraries, ICADL 2020*, p. 33–42. DOI : 10.1007/978-3-030-64452-9_3, HAL : hal-03034484.

KIESSLING B., TISSOT R., STOKES P. & EZRA D. S. B. (2019). escriptorium : An open source platform for historical document analysis. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, p. 19–19 : IEEE.

KOUDORO-PARFAIT C., LEJEUNE G. & ROE G. (2021). Spatial named entity recognition in literary texts : What is the influence of OCR noise ? In L. MONCLA, C. BRANDO & K. MCDONOUGH, Éds., *GeoHumanities@SIGSPATIAL 2021 : Proceedings of the 5th ACM SIGSPATIAL International Workshop on Geospatial Humanities, Beijing, China, November 2 - 5, 2021*, p. 13–21 : ACM. DOI : 10.1145/3486187.3490206.

LE P. & TITOV I. (2018). Improving entity linking by modeling latent relations between mentions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 1595–1604, Melbourne, Australia : Association for Computational Linguistics. DOI : 10.18653/v1/P18-1148.

LINHARES PONTES E., HAMDI A., SIDÈRE N. & DOUCET A. (2019). Impact of OCR Quality on Named Entity Linking. In *International Conference on Asia-Pacific Digital Libraries 2019*, Kuala Lumpur, Malaysia. DOI : 10.1007/978-3-030-34058-2_11, HAL : hal-02557116.

LOPRESTI D. P. (2009). Optical character recognition errors and their effects on natural language processing. *Int. J. Document Anal. Recognit.*, **12**(3), 141–151. DOI : 10.1007/s10032-009-0094-8.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013). Efficient estimation of word representations in vector space. In Y. BENGIO & Y. LECUN, Éds., *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

MILNE D. N. & WITTEN I. H. (2008). Learning to link with wikipedia. In *CIKM '08*.

MORO A., RAGANATO A. & NAVIGLI R. (2014). Entity linking meets word sense disambiguation : a unified approach. *Transactions of the Association for Computational Linguistics*, **2**, 231–244. DOI : 10.1162/tacl_a_00179.

NGUYEN H. T. & CAO T. H. (2008). Named entity disambiguation on an ontology enriched by wikipedia. In *2008 IEEE International Conference on Research, Innovation and Vision for the Future in Computing and Communication Technologies*, p. 247–254. DOI : 10.1109/RIVF.2008.4586363.

RAMSHAW L. A. & MARCUS M. P. (1995). Text chunking using transformation-based learning. DOI : 10.48550/ARXIV.CMP-LG/9505040.

SHEN W., WANG J. & HAN J. (2015). Entity linking with a knowledge base : Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, **27**(2), 443–460. DOI : 10.1109/TKDE.2014.2327028.

SMITH R. (2007). An overview of the tesseract ocr engine. In *Ninth international conference on document analysis and recognition (ICDAR 2007)*, volume 2, p. 629–633 : IEEE.

STANISLAWEK T., WRÓBLEWSKA A., WÓJCICKA A., ZIEMBICKI D. & BIECEK P. (2019). Named entity recognition - is there a glass ceiling ? In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, p. 624–633. DOI : 10.18653/v1/K19-1058.

SUCHANEK F., KASNECI G. & WEIKUM G. (2007). Yago : a core of semantic knowledge. p. 697–706. DOI : 10.1145/1242572.1242667.

SUN Y., LIN L., TANG D., YANG N., JI Z. & WANG X. (2015). Modeling mention, context and entity with neural networks for entity disambiguation. In *IJCAI*.

VAN STRIEN D., BEELEN K., ARDANUY M., HOSSEINI K., MCGILLIVRAY B. & COLAVIZZA G. (2020). Assessing the Impact of OCR Quality on Downstream NLP Tasks. In *In Proceedings of the 12th International Conference on Agents and Artificial Intelligence - Volume 1 : ARTIDIGH*, p. 484 – 496. DOI : 10.5220/0009169004840496.

VIGIER D., MONCLA L., BRENON A., MCDONOUGH K. & JOLIVEAU T. (2020). Classification des entités nommées dans l'encyclopédie ou dictionnaire raisonné des sciences des arts et des métiers par une société de gens de lettres (1751-1772).

WEICHSELBAUMER N., SEURET M., LIMBACH S., DONG R., BURGHARDT M. & CHRISTLEIN V. (2020). New approaches to ocr for early printed books. *DigItalia*, **2**, 74–87.

YAMADA I., SHINDO H., TAKEDA H. & TAKEFUJI Y. (2016). Joint learning of the embedding of words and entities for named entity disambiguation. *CoRR*, **abs/1601.01343**.