Simulation d'erreurs d'OCR dans les systèmes de TAL pour le traitement de données anachroniques

Baptiste Blouin, Benoit Favre, Jeremy Auguste

Simulation d'erreurs d'OCR dans les systèmes de TAL pour le traitement de données anachroniques

Baptiste Blouin[1, 2] Benoit Favre[1] Jeremy Auguste[2]
(1) LIS, Aix Marseille Université, Marseille, France
(2) Institut de Recherches Asiatiques, Aix Marseille Université, Aix, France
baptiste.blouin@lis-lab.fr, benoit.favre@lis-lab.fr,
jeremy.auguste@univ-amu.fr

RÉSUMÉ

L'extraction d'information offre de nouvelles perspectives au sein des recherches historiques. Cependant, la majorité des recherches liées à ce domaine s'effectue sur des données contemporaines. Malgré l'évolution constante des systèmes d'OCR, les textes historiques résultant de ce procédé contiennent toujours de multiples erreurs. Du fait d'un manque de ressources historiques dédiées au TAL, le traitement de ce domaine reste dépendant de l'utilisation de ressources contemporaines. De nombreuses études ont démontré l'impact négatif que pouvaient avoir les erreurs d'OCR sur les systèmes prêts à l'emploi contemporains. Mais l'évaluation des nouvelles architectures, proposant des résultats prometteurs sur des données récentes, face à ce problème reste encore très minime. Dans cette étude, nous quantifions l'impact des erreurs d'OCR sur trois tâches d'extraction d'information en utilisant plusieurs architectures de type *Transformers*. Au vu de ces résultats, nous proposons une approche permettant de réduire de plus de 50% cet impact sans avoir recours à des ressources historiques spécialisées.

ABSTRACT

Simulation of OCR errors in NLP systems for processing anachronistic data

Information extraction offers new perspectives in historical research. However, most of the research in this field is done on contemporary data. Despite the constant evolution of OCR systems, historical texts resulting from this process still contain multiple errors. Due to a lack of historical resources dedicated to NLP, the treatment of this domain remains dependent on the use of contemporary resources. Many studies have demonstrated the negative impact of OCR errors on contemporary off-the-shelf systems. But the evaluation of new architectures, offering promising results on recent data, against this problem is still very minimal. In this paper, we quantify the impact of OCR errors on three information extraction tasks using several *Transformers* architectures. Based on these results, we propose an approach to reduce the impact of OCR errors by more than 50% without using specialized historical resources.

MOTS-CLÉS : Données historiques, OCR, Transformers, Extraction d'information.

KEYWORDS: Historical data, OCR, Transformers, Information extraction.

1 Introduction

Contrairement aux documents nés numériques, les textes historiques sont les résultats d'une numérisation massive des ressources historiques. Étant donné que cette quantité de documents analogiques est encore assez importante malgré la récente évolution vers la création de documents numériques, des efforts considérables ont été déployés pour transformer ces documents papier en textes électroniques grâce à l'utilisation de système de reconnaissance optique de caractères (OCR). Bien que ces logiciels aient continuellement évolué et qu'ils puissent fonctionner correctement sur des textes modernes, leurs performances sur des documents anciens sont réduites par le manque de données d'entraînement sur ces types de documents. Les erreurs d'OCR, résultats de ces mauvaises performances, vont d'un caractère incorrect à des mots entiers et, par conséquent, à des phrases incorrectes.

De plus, dans le cadre de l'extraction automatique d'information, l'utilisation de systèmes alignés sur les données cibles est d'une importance capitale pour obtenir des résultats adéquats. Cependant, annoter des données est très coûteux. Le transfert de domaine est alors une approche rentable permettant de réduire considérablement l'effort humain. La majorité des systèmes actuels d'extraction d'information s'appuient sur des données annotées modernes. L'utilisation de ces ressources annotées contemporaines (OntoNotes, CoNLL) ou de modèles prêts à l'emploi (Spacy, Stanford NLP) est encore aujourd'hui l'approche principale au sein des applications d'extraction d'information pour les sciences humaines et sociales. De ce fait, de nombreux travaux ont étudié l'impact des entrées bruitées sur la recherche d'information et le traitement du langage naturel. Cet impact négatif motive l'amélioration des systèmes d'OCR, la correction automatique ou la mise en place de système de TAL robuste face à ce type d'erreur afin d'obtenir de meilleures performances lors de l'utilisation de ressources contemporaines pour le traitement des données historiques.

L'objectif de cette étude est d'évaluer la robustesse des modèles récents face aux erreurs d'OCR. Suite à cette évaluation, nous proposons une solution permettant de réduire l'impact des entrées bruitées sur ces modèles, grâce à une simulation des erreurs d'OCR lors de l'apprentissage des systèmes d'extraction d'information. Comparé aux autres approches cherchant à rendre les modèles plus robustes aux erreurs d'OCR et où les évaluations sont généralement effectuées sur une tâche et un jeu de données particulier ; nous proposons une approche indépendante des données ou de la tâche, s'appuyant sur des connaissances a priori sur les données que l'on souhaite traiter. Nous avons évalué cette approche sur la tâche de reconnaissance d'entité nommée (NER), l'étiquetage des parties du discours (POS) et l'extraction d'arguments d'événement (AE) sur plusieurs jeux de données en utilisant les architectures récentes des *Transformers*. Nos résultats montrent qu'une adaptation appropriée des données d'entraînement permet de réduire de 48% à 73%, en fonction de la tâche et du jeu de données, l'impact des erreurs d'OCR.

2 Travaux connexes

De nombreux travaux ont étudié l'impact des entrées bruitées sur le traitement du langage naturel. Packer *et al.* (2010) a expérimenté la reconnaissance de noms de personnes dans des textes OCR bruités à l'aide d'un modèle de Markov basé sur un dictionnaire, un modèle basé sur des regex, un modèle de Markov à entropie maximale et un modèle CRF, et a évalué les résultats par rapport à des données de test étiquetées manuellement. Grover *et al.* (2008) a construit un système de NER basé sur des règles pour reconnaître les noms de lieux et de personnes dans les documents numérisés

en se concentrant sur les problèmes causés par le niveau élevé de variance dans l'utilisation des lettres majuscules initiales des mots, ainsi que sur les problèmes liés à l'utilisation de la technologie OCR. Rodriguez *et al.* (2012) a évalué quatre outils de NER sur des textes historiques, dont OpenNLP (Kwartler, 2017) et Stanford NER. Ils ont montré que le système NER de Stanford avait la meilleure performance globale. Rodrigues Alves *et al.* (2018) montrent que l'incorporation de mots au niveau des caractères, combinée avec un modèle Bi-LSTM-CRF, peut aider à réduire l'impact des erreurs d'OCR et à traiter les mots rares dans les livres et revues savantes du 19-21C. Le NER n'est pas la seule tâche affectée par ce type d'erreur. Lin (2003) et Mieskes & Schmunk (2019) révèlent que le taux d'erreur de l'étiqueteur en partie de discours augmente linéairement avec celui de la sortie OCR. À partir de 5% de taux d'erreurs de caractères, les étiqueteurs en partie de discours subissent une dégradation significative de performances sur les données anglaises et allemandes. Il en va de même pour le résumé automatique (Jing *et al.*, 2003). En outre, il a été prouvé que les erreurs d'OCR ont une influence négative sur la détection des limites de phrases (Jing *et al.*, 2003; Lopresti, 2009; Strien *et al.*, 2020), sur la modélisation des sujets qui découvre les sujets abstraits latents dans une collection de documents d'entrée (Mutuvi *et al.*, 2018; Strien *et al.*, 2020), sur l'analyse des sentiments, sur la classification des textes (Murata *et al.*, 2006), et sur la liaison des entités nommées qui relie les entités nommées aux bases de connaissances externes (Linhares Pontes *et al.*, 2019). Plus récemment, Labusch *et al.* (2019) appliquent un modèle *Transformer* BERT multilingue, pré-entraîné sur de grandes quantités de données historiques allemandes non étiquetées, puis affiné sur plusieurs jeux de données NER. Ils montrent qu'un modèle BERT correctement pré-entraîné offre de bonnes performances dans une variété de contextes. Cependant, malgré ces évolutions récentes offrant des perspectives prometteuses, l'utilisation des représentations de mots, telles que BERT, est encore en questionnement, car sa capacité à gérer des données bruitées reste un point à clarifier quant à sa robustesse (Sun *et al.*, 2020). Finalement, Hamdi *et al.* (2020) ont étudié l'évolution des performances des systèmes de NER sur des données bruitées par l'OCR en utilisant des systèmes de NER s'appuyant sur les réseaux de neurones les plus récents. Ils montrent que la précision du NER chute de 90% à 50% lorsque le taux d'erreur sur les mots passe de 8% à 50%. Additionné à ces approches, de nouvelles campagnes d'évaluations se sont focalisées sur le traitement de textes historiques. HIPE (Ehrmann *et al.*, 2020) est une campagne d'évaluation du traitement des entités nommées sur des journaux historiques en français, allemand et anglais, qui vise à renforcer et à comparer la robustesse des approches sur des entrées non standard.

En comparaison à ces travaux, nous étudions la robustesse des modèles récents de TAL, sans les modifier, face aux erreurs d'OCR. Nous montrons que sans chercher à corriger ces erreurs, les modèles récents sont robustes lors d'un apprentissage sur des données similaires.

3 Jeu de données

Dans cette étude, nous avons utilisé quatre jeux de données permettant de couvrir trois tâches différentes. Pour le NER, tâche qui consiste à rechercher des objets textuels catégorisables dans des classes telles que noms de personnes, noms d'organisations, noms de lieux, etc, trois jeux de données sont utilisés. **OntoNotes 5.0** (Weischedel *et al.*, 2013) est un jeu de données en anglais, en arabe et en chinois, composé de textes provenant d'une grande variété de sources. Le jeu de données comprend 18 catégories d'entités nommées. Étant l'une des majeures sources contemporaines concernant le NER, son utilisation permet d'avoir une bonne estimation comparée à l'utilisation de systèmes prêts à l'emploi. **LitBank** (Bamman *et al.*, 2019) et **ACE 2005** (Walker *et al.*, 2006) sont deux jeux de

données partageant le même guide d'annotation dans deux domaines différents. Cet alignement permet d'évaluer un transfert de données contemporain vers le domaine historique. LitBank est un ensemble de données annotées de 100 textes littéraires de langue anglaise du domaine public provenant du Projet Gutenberg. Tous les textes ont été publiés avant 1923. La majorité des textes ayant été publiées entre 1852 et 1911. ACE a été développé par le Linguistic Data Consortium (LDC) et contient environ 1 800 documents de textes de genres mixtes en anglais, arabe et chinois annotés pour les entités, les relations et les événements. Ces documents datent de 2003 à 2004. Pour le POS, nous avons réalisé des expériences sur le Wall Street Journal du jeu de données Penn Treebank (Marcus *et al.*, 1993). L'extraction d'arguments d'événement (AE) consiste à identifier et à classifier le rôle des arguments (lieu, participants, date, etc.) liés à un événement. Cette tâche est évaluée sur le jeu de données ACE. Toutes les expériences ont été réalisées sur la partie anglaise de ces jeux de données.

4 Configuration

Nous avons utilisé trois systèmes différents qui sont représentatifs des systèmes qui pourraient être mis en œuvre dans une solution fournie par l'industrie pour l'extraction d'information sur des données historiques compte tenu de la technologie actuelle. **BERT** (Devlin *et al.*, 2019) ou **RoBERTa** (Liu *et al.*, 2019), avec une couche supplémentaire pour la prédiction. Les représentations contextuelles pré-entraînées sont affiné sur les jeux de données en utilisant l'approche proposée dans leurs papiers. Il s'agit d'un *de facto* de base pour les systèmes de TAL, et ils sont implémentés avec la bibliothèque *Transformers* (Wolf *et al.*, 2020). Nous avons également utilisé un modèle BERT (**BERT-Hist**) (Hosseini *et al.*, 2021) affinées sur 47 685 livres (5,1B tokens) en anglais de l'année 1760 à 1900 du Corpus Microsoft British Library. Ce modèle, comparé à BERT-Base appris à partir de Wikipédia et BookCorpus, nous permet de questionner l'impact d'un pré-entraînement de ces modèles sur des données historiques. **CharBERT** (Ma *et al.*, 2020) qui tient compte des caractères en plus des tokens BPE. Le modèle est pré-entraîné avec un objectif de modèle de langage bruité pour obtenir des représentations robustes au niveau des caractères. Nous nous attendons à ce qu'un tel modèle soit performant face aux erreurs d'OCR. Nous nous appuyons sur l'implémentation disponible sur le site `https://github.com/wtma/CharBERT`. BERT et CharBERT sont utilisés comme modèles de pré-entrainement sur les tâches de classification de tokens. RoBERTa est utilisé comme modèle de pré-entrainement sur la tâche d'extraction d'arguments d'événement.

Tous les résultats présentés sont des moyennes sur 10 initialisations aléatoires. L'objectif n'étant pas de maximiser les performances sur un jeu de données particulier ou sur une architecture particulière, les mêmes hyper-paramètres sont utilisés pour toutes les expériences d'un système donné. Le taux d'apprentissage est initialisé à 3e-5. Pour les trois architectures, la taille des couches cachées est fixée à 512 et l'apprentissage est effectué sur trois itérations. Le reste des paramètres sont les valeurs par défaut proposées par les bibliothèques pour les différents modèles. Les performances sont obtenues avec le score F1 calculé comme $2 \times \frac{\text{precision} \times \text{rappel}}{\text{precision} + \text{rappel}}$.

5 Quantification de l'impact des erreurs d'OCR

Afin d'évaluer l'impact des erreurs d'OCR sur les tâches TAL, une correction/alignement des erreurs OCR doit être incluse dans les corpus annotés. Afin de ne pas être limité dans l'évaluation de l'impact

du bruit sur les tâches TAL, nous avons décidé de simuler les erreurs OCR, pour l'appliquer sur plusieurs jeux de données, avec différents niveaux de bruit et sur plusieurs tâches.

En suivant les travaux précédents de Pruthi *et al.* (2019), nous modifions la séquence de caractères originale en supprimant, insérant et substituant des caractères à l'intérieur de celle-ci. Les probabilités d'erreurs sont calculées à partir du corpus ICDAR-17 (Nayef *et al.*, 2017), qui contient à la fois des OCR de plusieurs systèmes et des textes de référence. Provenant de trois collections digitales de presses anglaises, de la British Library, allant de 1744 à 1911, les données de ce corpus sont similaires à ce que peuvent traiter les historiens. De plus, avec ses six millions de caractères résultant de l'OCR alignés sur leurs valeurs de références, l'utilisation de cette source est raisonnable afin d'obtenir une distribution d'erreurs d'OCR. Cette approche permet de contrôler la quantité de bruit appliquée au jeu de données, simulant ainsi différents paramètres de difficulté de l'OCR. Trois tâches, la reconnaissance d'entités nommées, l'étiquetage de parties du discours et l'extraction d'arguments d'événements, sont évaluées face à l'impact du bruit envers les modèles récents de TAL. Dans un premier temps, nous avons évalué trois modèles dans un scénario de transfert, où ces modèles sont appris sur une tâche de NER sur ACE et évalués sur LitBank avec différents niveaux de bruit.

La quantité de bruit varie de 0% à 10% calculée comme le taux d'erreur sur les caractères (CER). Les résultats sont présentés dans le tableau 1.

CER	BERT-Hist	BERT	CharBERT
0%	63.91%	70.44%	67.38%
1%	64.37%	67.73%	66.17%
2.5%	61.61%	64.03%	63.86%
5%	58.29%	57.81%	60.58%
10%	47.71%	47.76%	48.50%

TABLE 1 – F1 obtenus sur le jeu de test de LitBank en fonction de la quantité de bruit injectée dans LitBank pour des modèles entraînés sur ACE.

CER	ICDAR-17	Uniform
0%	67.38%	67.38%
2.5%	63.86%	62.31%
5%	58.29%	57.55%
10%	48.50%	46.92%

TABLE 2 – F1 obtenus sur le jeu de test de LitBank en fonction de la quantité et la distribution de bruit injectée dans LitBank pour le modèle CharBERT entraîné sur ACE.

Les deux systèmes sont sensibles au bruit, et plus la quantité de bruit augmente, plus le F1 diminue. On peut voir que, en combinant un système appris sur des données propres avec des données de tests bruitées, CharBERT est moins sensible au bruit que BERT, plus les données sont dégradées. Et il en va de même pour BERT-Hist qui a été appris à partir de texte créé par OCR. Les deux systèmes montrent un comportement différent en ce qui concerne le bruit. Lorsque les données cibles sont bruitées à 10%, nous observons une forte baisse du F1, due à une baisse du rappel pour BERT et à une baisse de la précision pour CharBERT et BERT-Hist. Les résultats obtenus par BERT-Hist montrent que l'entraînement d'un modèle directement sur des données historiques bruitées permet d'atténuer l'impact du bruit sur des données également bruitées.

Comme la distribution d'erreurs est apprise sur le corpus ICDAR-17, elle correspond uniquement à un cas de type d'erreur potentiel appris par un système OCR. Afin d'évaluer le cas extrême, nous avons évalué le modèle CharBERT appris sur ACE et testé sur une version uniformément bruitée de LitBank. Les résultats obtenus dans le tableau 2, montrent qu'il est préférable de suivre une distribution provenant des erreurs d'OCR standard. Malgré cela, l'utilisation d'une distribution totalement uniforme a peu d'impact sur les résultats. La quantité de bruit présente dans les données à plus d'influence que la distribution utilisée.

Afin de s'assurer que la dégradation des résultats due aux erreurs d'OCR n'est pas un artefact de

ces jeux de données appliqués à cette tâche, nous avons évalué cet impact obtenu par CharBERT sur d'autres tâches/jeux de données tout en gardant une distribution cohérente des erreurs d'OCR (celle apprise sur ICDAR-17). Toujours dans le cadre de l'extraction d'information, nous avons évalué ce phénomène également sur l'étiquetage de parties de discours et l'extraction d'arguments d'événements. Ces deux autres tâches, n'ayant pas un alignement parfait entre les domaines contemporain et historique, sont alors évaluées dans les mêmes domaines. C'est-à-dire qu'étant donné un jeu de données, nous n'avons bruité que la partie test et laissé la partie d'entraînement telle quelle. Le tableau 3 montre les résultats obtenus par CharBERT, pour les tâches de classification de tokens, et RoBERTa, pour l'EA, en fonction de la quantité de bruit sur différentes tâches. Bien que toutes ces tâches différentes partagent la même métrique d'évaluation, les jeux de données, les typologies et la quantité de données/étiquettes ne sont pas les mêmes entre les résultats. Par conséquent, l'évaluation de la différence de score F1 entre ces différents résultats n'est pas pertinente. Il est alors plus intéressant de regarder l'évolution de l'augmentation du taux d'erreur en fonction de la quantité de bruit par tâche. Nous pouvons voir ici que, indépendamment de la tâche et du jeu de données, la diminution du score F1 est proportionnelle à la quantité de bruit.

Tâche	CER		0%	2.5%	5%	10%
	Train	**Test**	F1	F1	F1	F1
NER	ACE	LIT	67.38%	63.86%	60.58%	48.50%
NER	OntoNotes	OntoNotes	85.81%	78.42%	70.48%	52.06%
EA	ACE	ACE	72.86%	66.10%	60.52%	47.88%
POS	TreeBank	TreeBank	97.80%	95.76%	93.51%	88.34%

TABLE 3 – F1 obtenus en fonction de la quantité de bruit injectée dans les corpus **test** pour des modèles entrainés uniquement sur le **train** non bruité.

6 Réduction de l'impact des erreurs d'OCR

Afin de réduire l'impact des erreurs d'OCR, plusieurs solutions sont possibles. La post-correction de l'OCR, l'entraînement de modèles de langue sur des données historiques, l'utilisation d'une tokenisation des sous-mots jusqu'au niveau des caractères ou l'utilisation de données d'entraînement du domaine cible. Cependant, toutes ces approches, hormis l'entraînement des modèles de langage, nécessitent des données d'entraînement annotées pour obtenir des résultats satisfaisants.

Nous proposons une approche qui consiste à transférer automatiquement la distribution du bruit sur les données d'entraînement. Pour ce faire, la même fonction de bruit définie dans la partie précédente est appliquée directement sur les données utilisées comme sources de transfert. Dans un premier temps, la performance de cette approche compte tenu des modèles utilisés est évaluée. Et dans un scénario optimiste où nous avons la même distribution d'erreur OCR dans les données d'entraînement et de test.

Par rapport aux résultats précédents, lorsque la même distribution de bruit est appliquée au jeu de donnée d'entrainement (table 4), une grande amélioration est observée lorsque les modèles entraînés sur ACE bruité sont directement évalués sur LitBank bruité. En effet, la difficulté du rappel de BERT lors du passage de sans bruit à bruité est considérablement réduite lorsque nous bruitons le jeu de données source, tout en gardant une précision équivalente. Cependant, dans ce scénario, BERT a

appris aussi bien que CharBERT jusqu'à $2,5\%$ de CER. Au-delà, plus les données sont bruitées, plus CharBERT est robuste par rapport à BERT, avec une amélioration des performances de $5,93$ points de F1 avec 10% de CER. Réduite à $4,83$ points de F1 avec BERT-Hist. Ces résultats encouragent l'utilisation de modèles de langage utilisant la tokenisation au niveau des caractères pour le traitement de données bruitées.

CER	*BERT-Hist*	*BERT*	*CharBERT*
0%	65.67%	70.44%	67.38%
1%	64.95%	68.03%	66.76%
2.5%	62.46%	65.23%	65.52%
5%	59.15%	60.73%	62.25%
10%	52.80%	51.70%	57.63%

TABLE 4 – F1 obtenus sur le jeu de test de LitBank en fonction de la même quantité de bruit injectée dans les corpus **train (ACE) et test (LitBank)**.

CER	ICDAR-17	Uniforme	A $\neq$ B
0%	67.38%	67.38%	67.38%
2.5%	65.52%	64.78%	65.14%
5%	62.25%	62.10%	62.25%
10%	57.63%	57.16%	57.70%

TABLE 5 – F1 obtenus sur le jeu de test de LitBank en fonction de la quantité et la distribution de bruit injectée dans les corpus **train (ACE) et test (LitBank)** pour le modèle CharBERT.

Comme mentionné ci-dessus, ce scénario est cependant optimiste. Disposer des informations sur la distribution des erreurs d'OCR dans le corpus cible n'est pas toujours réalisable. Le tableau 5 montre les résultats de l'expérience précédente utilisant CharBERT avec deux autres scénarios de distribution des erreurs. *Uniforme* représente une distribution d'erreurs uniforme appliquée aux deux parties. Et $A \neq B$ représente des variantes de la distribution ICDAR appliquées de manière non-équivalente sur les deux parties. En effet, le fait d'avoir des variantes de la distribution ICDAR nous permet de rester proches de la distribution d'erreurs que nous pouvons obtenir des systèmes OCR. Et le fait d'avoir deux données d'entraînement et d'inférence différentes avec le même niveau de bruit facilite l'adaptation à de nouvelles données. Ces résultats montrent la stabilité de cette approche. Entraîner un système sur un niveau de bruit, indépendant de la distribution utilisée, permet de réduire l'impact des erreurs d'OCR.

Tâche	CER		0%	2.5%	5%	10%
	Train	**Test**				
NER	ACE	LIT	67.38%	65.52%	62.25%	57.63%
NER	OntoNotes	OntoNotes	85.81%	83.28%	81.00%	76.60%
EA	ACE	ACE	72.86%	70.82%	67.77%	62.38%
POS	TreeBank	TreeBank	97.80%	96.91%	95.64%	93.94%

TABLE 6 – F1 obtenus sur le test LitBank en fonction de la même quantité de bruit injectée dans les **Train** et **Test**.

Étant donné que nous disposons de la base de référence de ces jeux de données dans le cas de l'apprentissage sur des données propres (table 3), la réduction obtenue par cette approche appliquée aux autres tâches est comparée dans le tableau 6. Les résultats montrent que, l'impact des erreurs liées à l'OCR, avec 10% de bruit, peut être minimisé de 48% à 73% en fonction de la tâche et du jeu de données.

Au vu des résultats prometteurs de cette approche, une question reste à éclaircir. Les scénarios présentés ci-dessus supposent que la quantité de bruit est homogène sur l'ensemble des phrases à

traiter. Cependant, les documents textuels historiques ne présentent pas la même quantité de bruit entre les années, les documents, les pages, les phrases. Il est alors nécessaire de s'assurer que cette approche est fonctionnelle dans le contexte d'applications historiques. Pour ce faire, les modèles entraînés avec différents niveaux de bruit sont évalués sur la tâche d'EA sur plusieurs niveaux de bruit. Le tableau 7 montre les résultats de cette expérience.

Train \ Test	0%	2.5%	5%	10%
0%	**72.86%**	67.12%	63.16%	49.72%
2.5%	68.47%	**70.82%**	66.65%	56.71%
5%	68.30%	68.79%	**67.77%**	59.15%
10%	67.36%	68.58%	67.22%	**62.38%**

TABLE 7 – F1 obtenus sur le test en fonction de la quantité de bruit injectée dans le **Train** et **Test** d'ACE sur la tâche d'EA.

Les résultats montrent que pendant l'entraînement, plus la quantité de bruit est proche de celle des documents d'inférence, plus les résultats sont élevés. Cependant, le fait d'entraîner un système sur des données bruitées pour l'utiliser sur des données propres à moins d'impact que l'inverse. Seule une différence de $5,5\%$ du F1 est observée lors de l'application d'un système à 10% sur les données d'origine. En comparaison, une différence de $23,14\%$ du F1 est observée lorsqu'on applique un système original sur des données contenant 10% de CER.

7 Conclusion

Les erreurs d'OCR constituent un biais majeur pour le traitement automatique des textes. À travers les expériences proposées, l'impact que le bruit pouvait avoir sur les tâches d'extraction d'information est évalué avec l'utilisation des architectures *Transformers*. La configuration d'expérimentation reste cependant artificielle, mais les résultats de réduction de cet impact permettent d'identifier des clés majeures quant à l'amélioration des performances que l'on peut attendre. En effet, nous avons pu constater dans un premier temps qu'il est possible d'améliorer les performances des systèmes de TAL en injectant du bruit dans les données d'entraînement. Cette notion de bruit est ici définie par une distribution d'erreurs et une quantité d'injection au niveau des caractères. Nous avons vu que cette distribution a un léger impact sur les résultats obtenus, qu'elle diffère entre l'entraînement et l'inférence ou qu'elle soit totalement uniforme. En revanche, la quantité de bruit joue un rôle majeur quant aux résultats obtenus. Ces résultats sont cependant limités par le fait que nous utilisons la même distribution sur les données d'entraînement et sur celles de test. Mais au vu des résultats présentés dans le tableau 7, obtenir une estimation du CER présent dans les données est suffisant pour minimiser l'impact de ce type d'erreur. Une estimation qui pourrait être obtenue grâce à la sortie complète du système OCR proposant le score de confiance de celui-ci. En plus de pouvoir entraîner des modèles plus robustes pour ce type de document, avoir la sortie OCR nous permettrait d'identifier les documents traitables. Il est important de considérer le fait que les résultats obtenus par cette approche, ne corrigeant pas les erreurs, ne soient pas exploitables par les historiens. Afin de répondre à ce problème, nous voulons continuer cette recherche afin de combiner l'approche proposée à une correction ciblée sur les informations extraites par nos systèmes.

Références

BAMMAN D., POPAT S. & SHEN S. (2019). An annotated dataset of literary entities. In *Proceedings of the 2019 Conference of the North*, p. 2138–2144, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1220.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

EHRMANN M., ROMANELLO M., FLUCKIGER A. & CLEMATIDE S. (2020). Extended Overview of CLEF HIPE 2020 : Named Entity Processing on Historical Newspapers. p.̃38.

GROVER C., GIVON S., TOBIN R. & BALL J. (2008). Named entity recognition for digitised historical texts. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco : European Language Resources Association (ELRA).

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 M. HALL, T. MERČUN, T. RISSE & F. DUCHATEAU, Éds., *Digital Libraries for Open Knowledge*, p. 87–101, Cham : Springer International Publishing.

HOSSEINI K., BEELEN K., COLAVIZZA G. & ARDANUY M. C. (2021). Neural language models for nineteenth-century english. *CoRR*, **abs/2105.11321**.

JING H., LOPRESTI D. & SHIH C. (2003). Summarization of noisy documents : A pilot study. In *Proceedings of the HLT-NAACL 03 Text Summarization Workshop*, p. 25–32.

KWARTLER T. (2017). *The OpenNLP Project*, In *Text Mining in Practice with R*, chapitre 8, p. 237–269. John Wiley and Sons, Ltd. DOI : https ://doi.org/10.1002/9781119282105.ch8.

LABUSCH K., NEUDECKER C. & ZELLHÖFER D. (2019). Bert for named entity recognition in contemporary and historic german. In *KONVENS*.

LIN X. (2003). Impact of imperfect ocr on part-of-speech tagging. p. 284– 288 vol.1. DOI : 10.1109/ICDAR.2003.1227674.

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.

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. *CoRR*, **abs/1907.11692**.

LOPRESTI D. (2009). Optical character recognition errors and their effects on natural language processing. *IJDAR*, **12**, 141–151. DOI : 10.1145/1390749.1390753.

MA W., CUI Y., SI C., LIU T., WANG S. & HU G. (2020). CharBERT : Character-aware pre-trained language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, p. 39–50, Barcelona, Spain (Online) : International Committee on Computational Linguistics. DOI : 10.18653/v1/2020.coling-main.4.

MARCUS M. P., SANTORINI B. & MARCINKIEWICZ M. A. (1993). Building a large annotated corpus of English : The Penn Treebank. *Computational Linguistics*, **19**(2), 313–330.

MIESKES M. & SCHMUNK S. (2019). OCR quality and NLP preprocessing. In *Proceedings of the 2019 Workshop on Widening NLP*, p. 102–105, Florence, Italy : Association for Computational Linguistics.

MURATA M., BUSAGALA L. S. P., OHYAMA W., WAKABAYASHI T. & KIMURA F. (2006). The impact of ocr accuracy and feature transformation on automatic text classification. In H. BUNKE & A. L. SPITZ, Éds., *Document Analysis Systems VII*, p. 506–517, Berlin, Heidelberg : Springer Berlin Heidelberg.

MUTUVI S., DOUCET A., ODEO M. & JATOWT A. (2018). Evaluating the Impact of OCR Errors on Topic Modeling. In *Maturity and Innovation in Digital Libraries. 20th International Conference on Asia-Pacific Digital Libraries, ICADL 2018, Hamilton, New Zealand, November 19-22, 2018, Proceedings*, p. 3 – 14. DOI : 10.1007/978-3-030-04257-8_1, HAL : hal-03025563.

NAYEF N., YIN F., BIZID I., CHOI H., FENG Y., KARATZAS D., LUO Z., PAL U., RIGAUD C., CHAZALON J., KHLIF W., LUQMAN M. M., BURIE J.-C., LIU C.-L. & OGIER J.-M. (2017). Icdar2017 robust reading challenge on multi-lingual scene text detection and script identification - rrc-mlt. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 01, p. 1454–1459. DOI : 10.1109/ICDAR.2017.237.

PACKER T. L., LUTES J. F., STEWART A. P., EMBLEY D. W., RINGGER E. K., SEPPI K. D. & JENSEN L. S. (2010). Extracting person names from diverse and noisy ocr text. In *Proceedings of the Fourth Workshop on Analytics for Noisy Unstructured Text Data*, AND '10, p. 19–26, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/1871840.1871845.

PRUTHI D., DHINGRA B. & LIPTON Z. C. (2019). Combating Adversarial Misspellings with Robust Word Recognition. *arXiv :1905.11268 [cs]*. arXiv : 1905.11268.

RODRIGUES ALVES D., COLAVIZZA G. & KAPLAN F. (2018). Deep reference mining from scholarly literature in the arts and humanities. *Frontiers in Research Metrics and Analytics*, **3**, 21. DOI : 10.3389/frma.2018.00021.

RODRIGUEZ K. J., BRYANT M., BLANKE T. & LUSZCZYNSKA M. (2012). Comparison of named entity recognition tools for raw ocr text. DOI : 10.13140/2.1.2850.3045.

STRIEN D., BEELEN K., COLL ARDANUY M., HOSSEINI K., MCGILLIVRAY B. & COLAVIZZA G. (2020). Assessing the impact of ocr quality on downstream nlp tasks. DOI : 10.5220/0009169004840496.

SUN L., HASHIMOTO K., YIN W., ASAI A., LI J., YU P. & XIONG C. (2020). Adv-BERT : BERT is not robust on misspellings ! Generating nature adversarial samples on BERT. *arXiv :2003.04985 [cs]*. arXiv : 2003.04985.

WALKER C., STRASSEL S., MEDERO J. & MAEDA K. (2006). ACE 2005 Multilingual Training Corpus. Type : dataset, DOI : 10.35111/MWXC-VH88.

WEISCHEDEL R., PALMER M., MARCUS M., HOVY E., PRADHAN S., RAMSHAW L., XUE N., TAYLOR A., KAUFMAN J., FRANCHINI M., EL-BACHOUTI M., BELVIN R. & HOUSTON A. (2013). OntoNotes Release 5.0. type : dataset, DOI : 10.35111/XMHB-2B84.

WOLF T., DEBUT L., SANH V., CHAUMOND J., DELANGUE C., MOI A., CISTAC P., RAULT T., LOUF R., FUNTOWICZ M., DAVISON J., SHLEIFER S., VON PLATEN P., MA C., JERNITE Y., PLU J., XU C., SCAO T. L., GUGGER S., DRAME M., LHOEST Q. & RUSH A. M. (2020). Transformers : State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, p. 38–45, Online : Association for Computational Linguistics.