Oui mais.. ChatGPT peut-il identifier des entités dans des documents historiques ?

Carlos-Emiliano González-Gallardo[1] Emanuela Boros[2*] Nancy Girdhar[1]
Ahmed Hamdi[1] Jose G. Moreno[3] Antoine Doucet[1]
(1) La Rochelle Université, L3i, 17000 La Rochelle, France
(2) EPFL, Digital Humanities Laboratory, Lausanne, Suisse
(3) Université de Toulouse, IRIT UMR 5505 CNRS, 31000 Toulouse, France
carlos.gonzalez_gallardo@univ-lr.fr,{prénom.nom}@univ-lr.fr,
emanuela.boros@epfl.ch, jose.moreno@irit.fr

RÉSUMÉ

Les modèles de langage de grande taille (LLM) sont exploités depuis plusieurs années maintenant, obtenant des performances de l'état de l'art dans la reconnaissance d'entités à partir de documents modernes. Depuis quelques mois, l'agent conversationnel ChatGPT a suscité beaucoup d'intérêt auprès de la communauté scientifique et du grand public en raison de sa capacité à générer des réponses plausibles. Dans cet article, nous explorons cette compétence à travers la tâche de reconnaissance et de classification d'entités nommées (NERC) dans des sources primaires (des journaux historiques et des commentaires classiques) d'une manière *zero-shot* et en la comparant avec les systèmes de pointe basés sur des modèles de langage. Nos résultats indiquent plusieurs lacunes dans l'identification des entités dans le texte historique, allant de l'uniformité des directives d'annotation des entités, de la complexité des entités et du *code-switching*, à la spécificité de l'invite. De plus, comme prévu, la relative absence sur Internet des archives historiques et donc dans le corpus d'entraînement de ChatGPT a également un impact sur sa performance.

ABSTRACT

Yes but.. Can ChatGPT Identify Entities in Historical Documents ?

Large language models (LLM) have been leveraged for several years now, obtaining state-of-the-art performance in recognizing entities from modern documents. For the last few months, the conversational agent ChatGPT has "prompted" a lot of interest in the scientific community and public due to its capacity of generating plausible-sounding answers. In this paper, we explore this ability by probing it in the named entity recognition and classification (NERC) task in primary sources (i.e., historical newspapers and classical commentaries) in a zero-shot manner and by comparing it with state-of-the-art LM-based systems. Our findings indicate several shortcomings in identifying entities in the historical text that range from the consistency of entity annotation guidelines, entity complexity, and code-switching, to the specificity of prompting. Moreover, as expected, the inaccessibility of historical archives to the public (and thus on the Internet) also impacts its performance.

MOTS-CLÉS : Reconnaissance et classification d'entités nommées, Modèles de langage de grande taille, Transformeur génératif pré-entraîné, Documents historiques.

KEYWORDS: Named entity recognition and classification, Large language models, Generative pretrained transformer, Historical documents.

∗. Ce travail a été réalisé à l'Université de La Rochelle, à La Rochelle, France.

1 Introduction

Depuis qu'OpenAI a lancé ChatGPT lors du trente-sixième colloque sur les systèmes neuronaux de traitement de l'information (NeurIPS) en novembre 2022, sa capacité à fournir des réponses d'apparence humaines et plausibles a rendu le modèle extrêmement populaire au-delà de la communauté des chercheurs, avec plus d'un million d'utilisateurs en moins d'une semaine. ChatGPT est un agent conversationnel basé sur GPT-3.5 (Transformeur génératif pré-entraîné), un grand modèle de langage avec plus de 175 milliards de paramètres (Ouyang *et al.*, 2022). Étant donné sa grande popularité et son accessibilité, la question de savoir comment ce modèle hautement médiatisé se comporte dans différentes tâches de traitement du langage naturel (TAL) s'est déjà posée dans plusieurs domaines (Biswas, 2023; Pavlik, 2023).

Les modèles de langage de grande taille (LLM) sont exploités depuis plusieurs années maintenant, obtenant des performances de pointe dans la majorité des tâches de TAL, en étant généralement affinés sur des tâches en aval telles que la reconnaissance et la classification d'entités nommées (NERC) et moins dans des paramètres *zero-shot* (Li *et al.*, 2020). Ainsi, pour la reconnaissance et la classification d'entités nommées, mais aussi de manière générale, les efforts sont consacrés à la manière de transférer efficacement les connaissances pour l'adaptation au domaine en développant des systèmes robustes inter-domaines et en explorant l'apprentissage *zero-shot* ou *few-shot* pour traiter la cohérence et l'inadéquation des domaines et des annotations dans des contextes inter-domaines (Ehrmann *et al.*, 2020c, 2022). Simultanément, dans les documents historiques comme la presse ancienne, la tâche de NERC est confrontée à de nouveaux défis, outre l'hétérogénéité des domaines, tels que le bruit des entrées, la dynamique de la langue et le manque de ressources (Ehrmann *et al.*, 2021; Schweter & Baiter, 2019; González-Gallardo *et al.*, 2023; Boroş *et al.*, 2020; Najem-Meyer & Romanello, 2022; Schweter *et al.*, 2022; Boros *et al.*, 2022).

Dans ce court travail préliminaire, nous menons une étude exploratoire pour étudier le potentiel de ChatGPT, qui a été entraîné sur une quantité massive de données Internet (e.g., Common Crawl, WebText2, Wikipedia) (Brown *et al.*, 2020) et des ensembles de données d'invites pour l'apprentissage par renforcement à partir des préférences humaines (RLHF) (Ouyang *et al.*, 2022). Nous menons cette étude sur la tâche de NERC dans une configuration *zero-shot* et en comparant les performances de ChatGPT avec celles des systèmes de pointe.

2 Méthodologie

Nous avons suivi une approche de type *zero-shot* pour récupérer les entités nommées extraites par ChatGPT via son interface web officielle [1] entre le 11 janvier et le 7 février 2023. Une mise à jour a été publiée le 30 janvier pour améliorer la factualité et les capacités mathématiques du modèle [2], cependant, nous n'avons pas perçu de différence en ce qui concerne la capacité du modèle à détecter les entités.

1. https://chat.openai.com
2. https://help.openai.com/en/articles/6825453-chatgpt-release-notes

Jeux de données Nous avons sélectionné trois collections de documents historiques englobant une période d'environ 200 ans. Ces ensembles comprennent des commentaires classiques et des journaux historiques provenant de bibliothèques impliquées dans divers projets de recherche internationaux, notamment *NewsEye* [3] et *impresso* [4].

	Tokens	Entités	PERS	LOC	ORG	HumanProd	TIME	SCOPE
NewsEye	30 458	1 298	463	597	217	21	–	–
hipe-2020	48 854	1 600	502	854	130	61	53	–
ajmc	5 390	360	139	9	–	80	3	129

TABLE 1 – Description statistique des jeux de donnés (PERS = personne, LOC = emplacement, ORG = organisation, HumanProd = ouvrage/production, TIME = date/intervalle, SCOPE = partie spécifique de l'ouvrage).

Le jeu de données NewsEye (Hamdi *et al.*, 2021a) a été collecté auprès des bibliothèques nationales de France (BnF), d'Autriche (ONB) et de Finlande (NLF) [5]. Il se compose de quatre corpus (français, allemand, finnois et suédois). Le corpus français est constitué de textes provenant des archives numérisées de neuf journaux, notamment *L'Oeuvre*, *La Fronde*, *La Presse*, *Le Matin*, *Marie-Claire*, *Ce soir*, *Marianne*, *Paris Soir* et *Regards*, couvrant la période de 1854 à 1946.

Le jeu de données hipe-2020 (Ehrmann *et al.*, 2020b) est composé d'articles de journaux suisses, luxembourgeois et états-uniens en français, allemand et anglais couvrant les XIX^{e} et XX^{e} siècles. Il a été collecté principalement auprès de la Bibliothèque nationale suisse (BN), de la Bibliothèque nationale du Luxembourg (BnL), de la Médiathèque et des Archives d'Etat du Valais et des Archives économiques suisses (AES) [6] dans le cadre du projet *impresso*.

Le jeu de données ajmc (Romanello *et al.*, 2021) se compose de commentaires classiques provenant du projet *Ajax Multi-Commentary*. Ce projet rassemble des commentaires numérisés du XIX^{e} siècle rédigés en français, en allemand et en anglais. Ces commentaires offrent une analyse approfondie ainsi qu'une explication détaillée de la tragédie grecque *Ajax* de Sophocle.

Les collections mentionnées ont été annotées à deux niveaux de granularité (grossier et fin) pour la tâche de NERC. Les entités identifiées comprennent des catégories universelles telles que les personnes, les lieux et les organisations ; ainsi que des entités spécifiques au domaine telles que des références bibliographiques à la littérature primaire et secondaire. Les données ont été réparties en ensembles d'entraînement, de développement et de test.

Au cours de ce travail préliminaire, seules les partitions de test en français avec une granularité grossière ont été prises en compte pour simplifier la complexité des invites. Le tableau 1 fournit des informations sur le nombre et le type d'entités identifiées dans les ensembles de données spécifiés.

Nous avons défini les trois invites présentées dans le tableau 2 en fonction des différents types d'entités entre les ensembles de données et pour respecter la casse des étiquettes correspondantes. Même si le format IOB/BIO [7] est explicitement demandé pour chaque mot, la tokénisation par ChatGPT était

3. https://www.newseye.eu/
4. https://impresso-project.ch/
5. BnF : https://bnf.fr ; ONB : https://onb.ac.at ; NLF : https://kansalliskirjasto.fi
6. BN : https://www.nb.admin.ch ; BnL : https://bnl.public.lu ; AES : https://wirtschaftsarchiv.ub.unibas.ch
7. https://en.wikipedia.org/wiki/Inside%E2%80%93outside%E2%80%93beginning_(tagging)

incohérente avec les fichiers de jeux de données tokénisés IOB. Ainsi, à des fins d'évaluation, une vérification a été nécessaire pour assurer la cohérence de l'évaluation.

Des études antérieures ont démontré que l'inclusion d'instructions détaillées et d'exemples dans les invites ont un impact sur la qualité des résultats obtenus (Wang & Jin, 2023). Cependant, l'introduction d'une telle complexité dans les invites aurait été en contradiction avec notre étude de NERC dans une configuration *zero-shot*. Par conséquent, nous avons pris la décision de maintenir les invites aussi simples que possible afin de minimiser toute influence potentielle sur les résultats.

NewsEye	hipe-2020	ajmc
Quels sont les emplacements (LOC), les personnes (PER), organisations (ORG) et productions humaines (HumanProd) présents dans le texte historique suivant ? ***{PHRASE}*** Répondez, pour chaque mot, en utilisant IOB ou BIO séparé par tabulation. Si un mot n'a pas d'entité, ajoutez O.	Quels sont les emplacements (loc), les personnes (pers), organisations (org), produits (prod) et périodes (temps) présents dans le texte historique suivant ? ***{PHRASE}*** Répondez, pour chaque mot, en utilisant IOB ou BIO séparé par tabulation. Si un mot n'a pas d'entité, ajoutez O.	Quels sont les emplacements (loc), les personnes (pers), les périodes de temps (date), œuvres humaines (HumanProd), objets physiques (objet) et partie spécifique des travaux (étendue - *scope*) présents dans le texte historique suivant ? ***{PHRASE}*** Répondez, pour chaque mot, en utilisant IOB ou BIO séparé par tabulation. Si un mot n'a pas d'entité, ajoutez O.

TABLE 2 – Invites du jeu de données utilisées pour collecter les prédictions.

3 Résultats

Le tableau 3 présente les performances de ChatGPT par rapport au NERC avec une granularité grossière (types d'entités de haut niveau) en termes de précision (P), de rappel (R) et de F-mesure (F1) à micro-niveau *strict* et *fuzzy*, évaluées avec CLEF-HIPE-2020-scorer [8]. Nous présentons également les performances de deux systèmes NERC de l'état de l'art basés sur modèles de langage qui ont été entraînés avec l'ensemble d'entraînement des jeux de données figurant à la section 2.

	NewsEye			hipe-2020			ajmc		
	P	R	F1	P	R	F1	P	R	F1
strict									
Stacked NERC	**75,0**	70,6	**72,7**	–	–	–	–	–	–
Temporal NERC	–	–	–	**76,5**	**76,5**	**76,5**	**84,8**	**83,9**	**84,4**
ChatGPT	70,9	**72,3**	71,6	32,5	50,0	39,4	21,8	26,1	23,8
fuzzy									
Stacked NERC	**85,4**	**80,5**	**82,9**	–	–	–	–	–	–
Temporal NERC	–	–	-	**86,7**	**86, 7**	**86,7**	**90,2**	**89,2**	**89,7**
ChatGPT	77,8	79,4	78,6	49,0	75,4	59,4	25,5	30,6	27,8

TABLE 3 – Résultats comparatifs utilisant les trois jeux de données (micro).

Stacked NERC est basé sur le modèle pré-entraîné BERT proposé par (Devlin *et al.*, 2019) avec un empilement de deux blocs de Transformeur au-dessus, finalisé par une couche de prédiction à champ aléatoire conditionnel (CRF). *Stacked NERC* a démontré une performance accrue dans les documents historiques, tout en ne dégradant pas la performance sur les données modernes (Boros *et al.*, 2020; Boroş *et al.*, 2020). La même architecture a été utilisée comme base de référence dans la description du jeu de données NewsEye (Hamdi *et al.*, 2021a).

8. https://github.com/hipe-eval/HIPE-scorer

Temporal NERC s'appuie sur *Stacked NERC*, et il comprend une amélioration au niveau des données en exploitant des graphes de connaissances temporelles pour générer des informations temporelles contextuelles supplémentaires et une amélioration au niveau du modèle qui incorpore ces informations avec des contextes regroupés par la moyenne (González-Gallardo *et al.*, 2023). *Temporal NERC* a prouvé l'importance de la temporalité pour les journaux historiques et les commentaires classiques, en fonction des intervalles de temps et du taux d'erreur de numérisation.

D'après le tableau 3, il est clair que la capacité de ChatGPT à identifier des entités nommées dépend fortement de l'ensemble de données et du type d'entités. Des performances nettement inférieures en termes de F1 sont observées pour `ajmc`, avec une diminution de plus de 71% pour la métrique *strict* et de plus de 69% pour la métrique *fuzzy*. Pour `hipe-2020`, les performances ont diminué de manière moins radicale, avec plus de 48% et 31,48% respectivement. En ce qui concerne `NewsEye`, les scores sont légèrement similaires, avec une baisse d'environ 1,5% et un peu plus de 5% respectivement. Bien que les résultats soient globalement équilibrés, nous observons également un rappel plus élevé dans le cas de `hipe-2020`, ce qui pourrait indiquer que la complexité de l'annotation des entités permet à ChatGPT de les détecter, mais pas de les classer correctement [9]. La section suivante présente une analyse des faiblesses de ChatGPT dans le processus de reconnaissance des entités dans des textes historiques, en tenant compte de la définition des entités nommées, de leur complexité et des erreurs liées au processus de numérisation.

4 Analyse des erreurs

Définition des entités nommées L'annotation d'entités nommées suit des directives d'annotation bien définies (*guidelines*) pour décrire la nature et les limites des types d'entités, cependant il est nécessaire de faire confiance à l'intuition et à la conscience linguistiques de l'annotat·eur·rice (Hamdi *et al.*, 2021b; Ehrmann *et al.*, 2020a; Romanello & Najem-Meyer, 2022). Alors que les définitions des types d'entités universelles sont similaires entre les directives d'annotation, les types d'entités spécifiques à un domaine sont très variables. Celles de `hipe-2020` (Ehrmann *et al.*, 2020a) et `NewsEye` (Hamdi *et al.*, 2021b) définissent une entité de type « production humaine » comme étant tout ce qui est diffusé dans la presse, à la radio ou à la télévision, comme les journaux, les magazines, les émissions ou les catalogues de vente (e.g., *Die Zeit*, *Le Figaro*, *Le sept à huit*, *La ferme célébrités*) et excluent les produits médiatiques tels que les films et les téléfilms, ainsi que les doctrines politiques, philosophiques et religieuses/sectaires, comme *Der Sozialismus*, *Theheravada Buddhismus*, *Le socialisme*, *Le bouddhisme theravâda*. De même, l'entité de type « ouvrage » est décrite par les directives d'annotation pour `ajmc` (Romanello & Najem-Meyer, 2022) comme une entité désignant une création humaine, qu'elle soit intellectuelle ou artistique, qui peut être désignée par son titre. Pour le FRBR [10], une œuvre *« est une création intellectuelle ou artistique distincte »*, notamment les œuvres littéraires, les œuvres religieuses, les éditions de sources papyrologues et épigraphiques (e.g., *IG2*, *P.Oxy 1.119*), et les revues.

Étant donné que nous explorons la tâche de NERC dans une configuration *zero-shot* et bien que nous convenions que l'uniformité des annotations est une préoccupation majeure en raison de l'ambiguïté de la langue, nous ne pouvons que supposer que la variété des définitions crée nécessairement une difficulté pour ChatGPT, qui n'est pas entraîné sur des jeux de données annotés par des annotateurs.

9. Toutes les prédictions sont disponibles sur https://github.com/cic4k/NERC_ChatGPT.
10. https://www.oclc.org/research/activities/frbr.html

Complexité des entités Les directives d'annotation de `NewsEye` définissent à une entité nommée comme un objet du monde réel désignant un individu unique avec un nom propre. Historiquement, le nom d'une personne a joué un rôle influent en reflétant les attributs clés de son travail ou de sa vie, la majorité des entités ont également été annotées avec l'intitulé du métier de la personne. Prenons pour illustration l'exemple suivant : *« Becthoven. - Par le Quatuor de la fondation Beethoven : MM.A, Géioso1er violon ; A. Tracol ; 2e violon ;P. Monteux. aito, F. Schnéklud, violoncelle ; César Geloso, pianiste »*. « César Geloso, le pianiste » est considéré comme une entité « personne », cependant, ChatGPT n'a pas été capable de détecter au-delà de la mention du prénom et du nom. Si des métiers doivent être détectés à l'intérieur des entités, nous supposons que des informations supplémentaires doivent être ajoutées dans l'invite.

L'écriture bicamérale semble également problématique. Alors qu'il est courant que les noms d'organisations et les titres d'articles de journaux soient tout en majuscules (e.g., *LE SPORT*, *NOUVELLES BREVES*), ChatGPT les a tous identifiés comme des organisations. Les démonymes, les lieux et les personnes (e.g., *Carcassonnais, Russe, Mexicains, « Italien, 17 3/4 », « Japon 1899, 73 », « Portugais 3 %, 2 1/4 », « Russe 1906 »*) ont également posé des problèmes. Il n'est cependant pas clair si la confusion provient du fait que ces mots commencent par des majuscules ou si elle est due à un autre élément de contexte, car cette limitation se retrouve couramment dans des systèmes de NERC de l'état de l'art.

Les directives d'annotation de `NewsEye` et `impresso` considèrent des adresses telles que *« 130, rue de la Courselle »* et *« 56, rue de la Montagne-Sainte-Geneviève, 5e »* comme des emplacements, mais ChatGPT ne semble pas capturer ce niveau fin de granularité. Étant donné que le mot « emplacement » définit l'espace d'une manière sémantiquement plus générique qu'un lieu où une position spécifique, une adresse fait référence aux particularités d'un lieu qui, s'il n'est pas spécifié dans l'invite, ne peut pas être identifié correctement par ChatGPT.

Erreurs d'océrisation D'un point de vue quantitatif, ChatGPT n'a identifié que 7% des entités nommées comportant des erreurs de reconnaissance optique de caractères (OCR) dans le jeu de données `ajmc`, tandis que *Temporal NERC* a correctement identifié environ 40% de ces entités bruitées. *Temporal NERC* a reconnu des entités nommées dont jusqu'à 70% des caractères avaient subi une erreur d'océrisation (i.e., suppression, insertion et substitution).

Cependant, ces erreurs ne devraient pas dépasser 20% pour que les entités puissent être reconnuespar ChatGPT. Notamment, ChatGPT s'est montré incapable de reconnaître les entités nommées comportant des erreurs de segmentation. Par exemple, dans *« 13659 - 4360.Hxépra. ... . elloug. Ulysse paraît faire allusion à l'amertume des peroles que vient de prononcer À gamemnon ; Agamemnon répond comme si Ulysse avait eu en vue l'amertume de ses propres remontrances, 4866. 'EvOaë' Tlouet, j'en arriverai là, c'est-à-dire, je mourrai, .Dindorf : Kai'* αὐτὸς ἴξομαι πρὸς τὸ θάπτειν αὐτόν. », le nom d'une personne *Agamemnon* a été correctement identifié. En revanche, *« À gamemnon »* n'a pas été reconnu. Il n'est pas surprenant de constater que lorsque l'entité est fortement impactée par le processus d'océrisation, comme *Becthoven* au lieu de *Beethoven* dans *« XIe Quatuor (op. 95) Becthoven »*, ChatGPT ne parvient pas à détecter ces entités. Enfin, en raison d'une quantité abondante de bruit, comme dans *« m _'¦'— "..'i -i' i' —'i m Nota. Les Avis à insérer dans cette Feuille, qui ne seront pas remis au Bureau le mardi matin, à neuf-heures au plus tard, seront renvoyés irrévocablement _. [...] »*, ChatGPT « abandonne » avec cette déclaration : *« Le texte que vous avez fourni n'est pas dans un langage cohérent, et il est difficile de comprendre ce qu'il essaie de communiquer. Le texte contient principalement de la ponctuation et des caractères spéciaux sans entités reconnaissables. »*.

Alternance codique L'alternance codique (*code-switching*) est un phénomène qui consiste à alterner entre deux ou plusieurs langues dans une même phrase, une même expression ou un même mot. Le biais monolingue dans le multilinguisme est un type de biais qui peut se produire dans les modèles de langage (Talat *et al.*, 2022; Choudhury & Deshpande, 2021). `ajmc` présente de l'alternance codique entre le français et le grec ancien. GPT-3.5 a été entraîné dans plus de 100 langues, l'anglais étant surreprésenté avec 93% du nombre de mots, 1,82% pour le français et 0,032% pour le grec moderne, tandis que le grec ancien n'est pas représenté (Brown *et al.*, 2020). ChatGPT a été entraîné à l'aide d'autres ensembles de données pour un réglage fin, un retour humain et des messages-guides (Ouyang *et al.*, 2022), mais on peut s'attendre à une distribution similaire des langues. Ainsi, pour un exemple tel que *« À la marge d'un exemplaire de Sophocle, on lit la traduction suivante de ces deux vers, due à notre ÆRacine : " O mon fils, sois un jour plas beareux que ton pèren ; Da reste avec honneur tu peux lui ressembler". »*, ChatGPT répond *« Je suis désolé, mais je ne suis pas en mesure de comprendre le texte que vous avez fourni. Le texte semble être un mélange de grec ancien et de français, avec quelques références à la littérature ancienne et des annotations, ce qui rend difficile l'extraction d'informations significatives. De plus, le texte n'est pas un texte historique, mais plutôt un texte littéraire, ce qui rend également difficile l'extraction d'entités historiques. »*.

5 Conclusions et perspectives

ChatGPT rencontre plusieurs difficultés dans la reconnaissance des entités dans les documents historiques, qui relèvent de la cohérence des directives d'annotation des entités, de la complexité des entités, du multilinguisme, de l'alternance codique et de la spécificité de l'invite. De plus, alors qu'une quantité sans précédent de documents historiques est disponible en format numérique, peu de choses sont disponibles gratuitement avec de nombreuses archives historiques qui restent inaccessibles au public. Par exemple, les sources primaires telles que les journaux et les articles de magazines (comme dans le cas de la majorité des ensembles de données de cette étude) sont disponibles à la fois en ligne et en bibliothèque, mais néanmoins, elles sont généralement filigranées ou derrière un mur payant.

Par conséquent, ChatGPT est « ignorant » (pour l'instant) de ces connaissances, ce qui contribue au degré de confusion du modèle en ce qui concerne les documents historiques. Même si ces ressources deviennent accessibles et que les systèmes basés sur les LLM améliorent leurs capacités de « compréhension » des documents historiques, leur mise en œuvre dans les bibliothèques numériques devra être prise avec précaution pour éviter les réponses biaisées et hors domaine.

Déclaration éthique

Bien qu'il puisse générer un texte à la consonance plausible, le contenu généré par ChatGPT n'est pas nécessairement vrai. Néanmoins, nous considérons qu'il ne concerne pas la tâche de reconnaissance d'entités nommées, car nous n'ajoutons aucune autre considération éthique que celles posées par ChatGPT. Nous sommes conscients de la position intentionnelle (Dennett, 2009) de termes comme « abandonne », « compréhension » et « ignorant » lorsqu'ils sont appliqués à un agent conversationnel, cependant, dans ce papier, nous les adoptons pour souligner la capacité de ChatGPT à interagir avec un utilisateur.

Remerciements

Ce travail a été soutenu par les projets ANNA (2019-1R40226), TERMITRAD (AAPR2020-2019-8510010), Pypa (AAPR2021-2021-12263410) et Actuadata (AAPR2022-2021-17014610) financés par la Région Nouvelle-Aquitaine, France.

Références

BISWAS S. (2023). Chatgpt and the future of medical writing.

BOROS E., GONZÁLEZ-GALLARDO C.-E., GIAMPHY E., HAMDI A., MORENO J. G. & DOUCET A. (2022). Knowledge-based contexts for historical named entity recognition & linking. *Conference and Labs of the Evaluation Forum (CLEF 2020)*.

BOROŞ E., HAMDI A., PONTES E. L., 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.

BOROS E., PONTES E. L., CABRERA-DIEGO L. A., HAMDI A., MORENO J. G., SIDÈRE N. & DOUCET A. (2020). Robust named entity recognition and linking on historical multilingual documents. In *Conference and Labs of the Evaluation Forum (CLEF 2020)*, volume 2696, p. 1–17 : CEUR-WS Working Notes.

BROWN T., MANN B., RYDER N., SUBBIAH M., KAPLAN J. D., DHARIWAL P., NEELAKANTAN A., SHYAM P., SASTRY G., ASKELL A. *et al.* (2020). Language models are few-shot learners. *Advances in neural information processing systems*, **33**, 1877–1901.

CHOUDHURY M. & DESHPANDE A. (2021). How linguistically fair are multilingual pre-trained language models ? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, p. 12710–12718.

DENNETT D. (2009). Intentional Systems Theory. In *The Oxford Handbook of Philosophy of Mind*. Oxford University Press. DOI : 10.1093/oxfordhb/9780199262618.003.0020.

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, WATTER, ROMANELLO, CLEMATIDE & FLÜCKIGER (2020a). *Impresso Named Entity Annotation Guidelines*. DOI : 10.5281/zenodo.3604227.

EHRMANN M., HAMDI A., PONTES E. L., ROMANELLO M. & DOUCET A. (2021). Named entity recognition and classification on historical documents : A survey. *arXiv preprint arXiv :2109.11406*.

EHRMANN M., ROMANELLO M., CLEMATIDE S., STRÖBEL P. & BARMAN R. (2020b). Language resources for historical newspapers : the impresso collection.

EHRMANN M., ROMANELLO M., FLÜCKIGER A. & CLEMATIDE S. (2020c). Extended overview of clef hipe 2020 : named entity processing on historical newspapers. In *CEUR Workshop Proceedings*, volume 2696 : CEUR-WS.

EHRMANN M., ROMANELLO M., NAJEM-MEYER S., DOUCET A. & CLEMATIDE S. (2022). Extended overview of hipe-2022 : Named entity recognition and linking in multilingual historical documents. In *Proceedings of the 13th International Conference of the CLEF Association (Lecture Notes in Computer Science)*, volume 13390 : Springer.

GONZÁLEZ-GALLARDO C.-E., BOROS E., GIAMPHY E., HAMDI A., MORENO J. G. & DOUCET A. (2023). Injecting temporal-aware knowledge in historical named entity recognition. In *Advances in Information Retrieval : 45th European Conference on IR Research, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part I*, p. 65–79 : Springer.

HAMDI A., LINHARES PONTES E., BOROS E., NGUYEN T. T. H., HACKL G., MORENO J. G. & DOUCET A. (2021a). A multilingual dataset for named entity recognition, entity linking and stance detection in historical newspapers. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 2328–2334.

HAMDI A., PONTES E. L. & DOUCET A. (2021b). Annotation Guidelines for Named Entity Recognition, Entity Linking and Stance Detection. DOI : 10.5281/zenodo.4574199.

LI J., SUN A., HAN J. & LI C. (2020). A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, **34**(1), 50–70.

NAJEM-MEYER S. & ROMANELLO M. (2022). Page layout analysis of text-heavy historical documents : a comparison of textual and visual approaches. In *Proceedings of the Computational Humanities Research Conference 2022 Antwerp, Belgium, December 12-14, 2022.*, p. 36–54.

OUYANG L., WU J., JIANG X., ALMEIDA D., WAINWRIGHT C. L., MISHKIN P., ZHANG C., AGARWAL S., SLAMA K., RAY A. *et al.* (2022). Training language models to follow instructions with human feedback. *arXiv preprint arXiv :2203.02155*.

PAVLIK J. V. (2023). Collaborating with chatgpt : Considering the implications of generative artificial intelligence for journalism and media education. *Journalism & Mass Communication Educator*, p. 10776958221149577.

ROMANELLO M. & NAJEM-MEYER S. (2022). *Guidelines for the Annotation of Named Entities in the Domain of Classics*. DOI : 10.5281/zenodo.6368101.

ROMANELLO M., NAJEM-MEYER S. & ROBERTSON B. (2021). Optical character recognition of 19th century classical commentaries : The current state of affairs. In *The 6th International Workshop on Historical Document Imaging and Processing*, HIP '21, p. 1–6, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3476887.3476911.

SCHWETER S. & BAITER J. (2019). Towards robust named entity recognition for historic German. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, p. 96–103, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/W19-4312.

SCHWETER S., MÄRZ L., SCHMID K. & ÇANO E. (2022). hmbert : Historical multilingual language models for named entity recognition. *Conference and Labs of the Evaluation Forum (CLEF 2020)*.

TALAT Z., NÉVÉOL A., BIDERMAN S., CLINCIU M., DEY M., LONGPRE S., LUCCIONI S., MASOUD M., MITCHELL M., RADEV D. *et al.* (2022). You reap what you sow : On the challenges of bias evaluation under multilingual settings. In *Proceedings of BigScience Episode# 5–Workshop on Challenges & Perspectives in Creating Large Language Models*, p. 26–41.

WANG S. & JIN P. (2023). A brief summary of prompting in using gpt models.