Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Injection de connaissances temporelles dans la reconnaissance d'entités nommées historiques

Carlos-Emiliano González-Gallardo[1] Emanuela Boros[2*] Edward Giamphy[1,3]
Ahmed Hamdi[1] José G. Moreno[4] Antoine Doucet[1]
(1) La Rochelle Université, L3i, 17000 La Rochelle, France
(2) EPFL, Digital Humanities Laboratory, Lausanne, Suisse
(3) Preligens, 75009 Paris, France
(4) 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, edward.giamphy@preligens.com
jose.moreno@irit.fr

RÉSUMÉ

Dans cet article nous abordons la reconnaissance d'entités nommées (NER) dans des documents historiques multilingues. Cette tâche présente des multiples défis, tels que les erreurs générées suite à la numérisation et de la reconnaissance optique des caractères de ces documents. De plus, ces collections sont distribuées sur une période de temps assez longue et suivent plusieurs conventions orthographiques qui évoluent au fil du temps. Pour répondre à ce défi nous récupérons des contextes supplémentaires, sémantiquement pertinents en exploitant des graphes de connaissances temporelles à partir des informations temporelles fournies par les collections historiques. Ces contextes sont ensuite inclus en tant que représentations mises en commun dans un modèle NER basé sur des Transformeurs. Nous menons des expérimentations avec deux collections historiques multilingues récentes en anglais, français et allemand, composées de journaux historiques (XIX^{e} - XX^{e} siècles) et de commentaires classiques (XX^{e} siècle). Les résultats démontrent l'efficacité de l'injection de connaissances temporelles dans ces ensembles de données.

ABSTRACT

Injecting Temporal-aware Knowledge in Historical Named Entity Recognition.

In this paper we address the detection of named entities in multilingual historical collections. This task presents multiple challenges as a result of digitization and optical caracter recogonition processes. In addition, these collections are distributed over a fairly long period of time and are affected by changes and evolution of natural language. To address this challenge we retrieve semantically-relevant additional contexts from temporal knowledge graphs by extracting the time information provided on historical data collections and include them as mean-pooled representations in a Transformer-based NER model. We experiment with two recent multilingual historical collections in English, French, and German, consisting of historical newspapers (19C-20C) and classical commentaries (19C). The results show the effectiveness of injecting temporal-aware knowledge into the different datasets.

MOTS-CLÉS : Reconnaissance d'entités nommées, Extraction d'informations temporelles, Humanités numériques.

KEYWORDS: Named entity recognition, Temporal information extraction, Digital humanities.

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

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 4 : articles déjà soumis ou acceptés en conférence internationale, pages 65–73.

1 Introduction

Ces dernières décennies ont vu la mise à disposition d'un nombre croissant de corpus textuels pour les sciences humaines et sociales. Des exemples représentatifs proviennent de *Gallica*, la bibliothèque numérique de la Bibliothèque nationale de France [1], et de *Trove*, l'agrégateur de bases de données et service de documents en texte intégral, d'images numériques et de stockage de données provenant de la Bibliothèque nationale d'Australie [2]. L'accès à ces données massives offre de nouvelles perspectives à un nombre croissant de disciplines, allant de l'histoire sociopolitique et culturelle à l'histoire économique, ainsi que de la linguistique à la philologie.

Des milliards d'images de documents historiques, y compris des documents manuscrits numérisés, des registres médiévaux et des journaux anciens numérisés, sont désormais stockés et leur contenu est transcrit, soit manuellement grâce à des interfaces dédiées, soit automatiquement en utilisant la reconnaissance optique de caractères (OCR) ou la reconnaissance de texte manuscrit. Le processus de numérisation en masse, initié dans les années 1980 par des projets internes à petite échelle, a conduit à la montée en puissance de la numérisation, qui a atteint une certaine maturité au début des années 2000 avec des campagnes de numérisation à grande échelle dans toute l'industrie (Ehrmann *et al.*, 2020a,c).

Alors que ce processus de numérisation de masse se poursuit de plus en plus d'approches du domaine du traitement du langage naturel (TAL) sont dédiées aux documents historiques, offrant de nouveaux moyens d'accéder à des archives enrichies sémantiquement en texte intégral (Oberbichler *et al.*, 2022), tels que la reconnaissance d'entités nommées (NER) (Boroş *et al.*, 2020a; Ehrmann *et al.*, 2023; Hamdi *et al.*, 2021), l'annotation sémantique (Linhares Pontes *et al.*, 2022) et la détection d'événements (Boros *et al.*, 2022; Nguyen *et al.*, 2020).

La NER est une tâche d'extraction d'information dédiée à l'identification d'entités d'intérêt dans les textes, généralement de type personne, organisation et lieu. Ces entités agissent comme des ancrages référentiels qui sous-tendent la sémantique des textes et guident leur interprétation. Par exemple, en Europe, à l'époque médiévale, la plupart des personnes étaient identifiées par un simple mononyme ou un seul nom propre. Les noms de famille ou patronymes ont commencé à être utilisés à partir du XIII^{e} siècle, mais beaucoup plus tard dans certaines régions ou classes sociales (XVII^{e} siècle pour les Gallois). De nombreuses personnes partageaient le même nom et la même orthographe dans les langues vernaculaires et latines, mais aussi au sein d'une même langue (e.g., Guillelmus, Guillaume, Willelmus, Guillaume, Wilhelm). Les lieux ont pu disparaître ou changer complètement. Pour ceux qui ont survécu de la préhistoire jusqu'au XXI^{e} siècle (e.g., l'Écosse, le Pays de Galles, l'Espagne), ils sont très ambigus et possèdent de très différentes orthographes, ce qui rend leur identification très difficile (Boroş *et al.*, 2020b).

Dans cet article, nous nous concentrons sur l'exploration de la temporalité dans la NER à partir de collections historiques. Nous proposons une nouvelle technique pour injecter des connaissances temporelles supplémentaires en s'appuyant sur Wikipédia et Wikidata pour fournir des informations contextuelles sémantiquement proches.

1. https://gallica.bnf.fr/
2. https://trove.nla.gov.au/

2 Contextes basés sur des connaissances temporelles

Plusieurs travaux ont montré que les erreurs d'OCR peuvent avoir un impact sur des tâches en TAL (van Strien *et al.*, 2020), et plus particulièrement sur la NER (Hamdi *et al.*, 2022). Pour remédier à cela, des efforts ont été déployés pour élaborer des corpus et/ou des systèmes de NER adaptés (Ehrmann *et al.*, 2023). Dans ce travail, nous proposons d'introduire des contextes externes grammaticalement corrects dans les systèmes de NER. Ces contextes supplémentaires contribuent à améliorer les performances des systèmes de NER en dépit les erreurs d'OCR (Wang *et al.*, 2022). De plus, l'inclusion de tels contextes, en tenant compte de la temporalité, pourrait encore améliorer la détection des entités, qui sont particulièrement sensibles au contexte temporel. Ainsi, nous proposons plusieurs configurations pour inclure ces contextes supplémentaires, basés sur Wikidata5m [3] (Wang *et al.*, 2021), un graphe de connaissances (KG) comportant cinq millions d'entités Wikidata [4]. Wikidata5m contient des entités du domaine général (e.g., des célébrités, des événements, des concepts, des objets) qui sont alignées sur une description correspondant au premier paragraphe de sa page Wikipédia."

2.1 Intégration de l'information temporelle

Nous agrégeons la temporalité dans Wikidata5m en utilisant le graphe de connaissances temporelles (TKG) créé par (Leblay & Chekol, 2018) et mis au point par (García-Durán *et al.*, 2018) [5]. Ce TKG contient plus de 11 000 entités, 150 000 faits, et un champ temporel couvrant les années 508 à 2017. Pour une entité donnée, il fournit un ensemble de faits décrivant les interactions de l'entité dans le temps. Il est donc nécessaire de combiner ces faits en un seul élément en utilisant un opérateur d'agrégation sur leurs éléments temporels.

Nous effectuons une transformation sur les informations temporelles de chaque fait d'une entité afin de les combiner en un seul élément d'information temporelle. Soit e une entité décrite par les faits : $F_e i = 1^n = (e, r1, e_1, t_1), \ldots (e, r_i, e_i, t_i), \ldots (e, r_n, e_n, t_n)$, où le fait (e, r_i, e_i, t_i) est composé de deux entités e et e_i reliées par la relation r_i et l'horodatage t_i. Un horodatage est un point discret dans le temps qui correspond à une période (une année dans ce travail). L'opérateur d'agrégation est la fonction $AGG \rightarrow t_e$ qui prend en entrée l'information temporelle de F_e et génère l'information temporelle associée à e. Plusieurs opérateurs sont possibles (moyenne, médiane, minimum et maximum). Le minimum d'un ensemble de faits est défini par le fait le plus ancien tandis que le maximum correspond au fait le plus récent. Si une entité est associée à quatre faits s'étendant sur les années 1891, 1997, 2006 et 2011, l'opérateur d'agrégation minimum consiste à conserver le plus ancien, ce qui fait de l'année 1891 l'information temporelle de l'entité.

Étant donné que nos ensembles de données correspondent à des documents entre le XIX^{e} et le XX^{e} siècles, l'opérateur d'agrégation minimum est plus susceptible de créer un contexte temporel approprié pour les entités. Il met en évidence les entités correspondant à une période en accentuant les faits plus anciens. À la fin de l'opération d'agrégation, 8 176 entités de Wikidata5m ont été associées à une année comprise entre 508 et 2001, ce qui permet de filtrer la plupart des faits survenus au cours du XXI^{e} siècle.

3. https://deepgraphlearning.github.io/project/wikidata5m
4. https://www.wikidata.org/
5. https://github.com/mniepert/mmkb/tree/master/TemporalKGs/wikidata

2.2 Recherche de contexte

Notre système de base de connaissances repose sur une instance locale d'ElasticSearch [6] et utilise une correspondance de similarité sémantique multilingue, ce qui présente un avantage pour les requêtes multilingues. Cette correspondance est réalisée avec des index de champ vectoriel dense. Ainsi, à partir d'un vecteur de requête, une API de recherche des k plus proches voisins (k-NN) récupère les k vecteurs les plus proches et renvoie les documents correspondants en tant que résultats de recherche.

Pour chaque entité Wikidata5m, nous créons une entrée ElasticSearch comprenant un identifiant, un champ de description et un champ contenant le vecteur dense de la description, obtenus à l'aide du modèle multilingue pré-entraîné Sentence-BERT (Reimers & Gurevych, 2019, 2020). Nous construisons un index sur l'identifiant de l'entité et un index vectoriel dense sur les vecteurs de description. Nous proposons deux configurations différentes pour la récupération du contexte :

— `non-temporelle` : cette configuration n'utilise aucune information temporelle. Lors de la recherche de contexte pour une phrase d'entrée, nous commençons par obtenir la représentation vectorielle dense correspondante avec le même modèle Sentence-BERT utilisé pendant la phase d'indexation. Ensuite, nous interrogeons la base de connaissances afin de récupérer les entités les plus proches sur le plan sémantique en utilisant une recherche de similarité cosinus via l'algorithme des *k-NN* sur l'index du vecteur dense de la description. Le contexte C est finalement composé de k descriptions d'entités.

— `temporelle-`δ : cette configuration intègre les informations temporelles. Après la récupération des entités sémantiquement similaires avec `non-temporelle`, nous appliquons une opération de filtrage pour garder ou exclure les entités du contexte. En utilisant l'année t_{input} liée aux métadonnées de la phrase d'entrée lors de la recherche de contexte, nous conservons une entité si son année associée t_e se situe dans l'intervalle $t_{input} - \delta \leq t_e \leq t_{input} + \delta$, où δ est le seuil de l'intervalle d'années. Sinon, l'entité est rejetée. La valeur de t_e correspond à l'année la plus ancienne parmi tous les faits de l'entité e dans le TKG, conformément à l'opération d'agrégation *AGG*. Si t_e est absent, l'entité e est également conservée. Cette opération est répétée jusqu'à ce que $|C| = k$.

2.3 Architecture

Notre modèle de base se compose d'une approche d'apprentissage hiérarchique et multitâche, avec un encodeur ajusté basé sur BERT. Ce modèle comprend un encodeur avec deux couches de Transformeur (Vaswani *et al.*, 2017) dotées de modules adaptateurs (Houlsby *et al.*, 2019; Pfeiffer *et al.*, 2020) au-dessus du modèle pré-entraîné BERT. Les adaptateurs sont ajoutés à chaque couche de Transformeur après la projection suivant l'attention multitêtes et ils s'adaptent non seulement à la tâche, mais aussi à l'entrée bruitée, ce qui a prouvé augmenter les performances de la reconnaissance d'entités dans de telles conditions spéciales (Boroş *et al.*, 2020a). Enfin, la couche de prédiction multitâche est constituée de couches distinctes de champs conditionnels aléatoires (CRF).

Pour inclure les contextes supplémentaires, nous introduisons les *jokers contextuels*. Chaque contexte supplémentaire passe par l'encodeur pré-entraîné [7] générant un $JokerTokRep$ qui est ensuite réduit

6. https://www.elastic.co/guide/en/elasticsearch/reference/8.1/release-highlights.html

7. Dans ce cas, nous n'utilisons pas les couches de Transformeur supplémentaires avec les adaptateurs, car ceux-ci ont été spécifiquement proposés pour le texte bruité/non standard et n'apportent aucune amélioration des performances sur le texte

à la moyenne le long de l'axe de la séquence. Nous appelons cette représentation le *joker contextuel*. Nous les considérons comme des jokers insérés discrètement dans la représentation de la phrase actuelle pour améliorer la reconnaissance des entités. Cependant, nous considérons également que ces jokers peuvent affecter les résultats d'une manière qui n'est pas immédiatement apparente et peuvent nuire aux performances d'un système de NER.

Configuration expérimentale Notre configuration expérimentale comprend un modèle de base et quatre configurations avec différents niveaux de contextes basés sur les connaissances :

— *sans-contexte* : dans cette configuration de base, aucun contexte n'est ajouté aux représentations des phrases d'entrée.
— *non-temporelle* : les *jokers contextuels* sont générés et intégrés avec la première configuration de recherche de contexte sans information temporelle.
— *temporelle-(50|25|10)* : les *jokers contextuels* sont générés et intégrés à l'aide de la deuxième configuration de recherche de contexte avec un seuil d'intervalle d'année $\delta \in \{50, 25, 10\}$.

L'évaluation est réalisée en termes de [P]récision, de [R]appel et de mesure F1 au niveau micro (Ehrmann *et al.*, 2020a) dans un cadre strict (correspondance exacte des limites) [8].

Jeux de données Nous avons sélectionné deux collections de documents historiques comprenant des journaux historiques et des commentaires classiques.

— `hipe-2020` (Ehrmann *et al.*, 2020b) : couvre les XIX^{e} et XX^{e} siècles et rassemble des articles de journaux suisses, luxembourgeois et états-uniens en français, en allemand et en anglais provenant de diverses sources telles que la Bibliothèque nationale suisse (BN), la Bibliothèque nationale du Luxembourg (BnL), la Médiathèque et des Archives d'Etat du Valais et les Archives économiques suisses (AES) [9] dans le cadre du projet *impresso*.
— `ajmc` (Romanello *et al.*, 2021) : se compose de commentaires classiques rédigés en français, en allemand et en anglais provenant du projet *Ajax Multi-Commentary*. Ces commentaires, datant du XIX^{e} siècle, fournissent une analyse détaillée de la tragédie grecque *Ajax* de Sophocle datant du début de la période médiévale [10].

3 Résultats

Le tableau 1 présente nos résultats pour les trois langues et les deux ensembles de données. Les meilleurs résultats sont en gras. Nous pouvons observer que les modèles avec des *jokers contextuels* présentent une amélioration par rapport au modèle de base sans contextes supplémentaires. De plus, l'inclusion d'informations temporelles conduit à de meilleurs résultats que les contextes non temporels. Les scores sur `ajmc` s'avèrent plus élevés que ceux obtenus sur `hipe-2020`, quel que

standard, comme l'ont observé Boroş *et al.* (2020a).

8. Nous avons utilisé l'évaluateur HIPE disponible sur `https://github.com/hipe-eval/HIPE-scorer`.

9. BN : `https://www.nb.admin.ch`; BnL : `https://bnl.public.lu`; AES : `https://wirtschaftsarchiv.ub.unibas.ch`

10. Bien que la date exacte de sa première représentation soit inconnue, la plupart des spécialistes la datent du début de la carrière de Sophocle (peut-être la plus ancienne pièce de Sophocle encore existante), quelque part entre 450 et 430 avant J.-C., peut-être vers 444 avant J.-C.

Français						Allemand						Anglais					
hipe-2020			ajmc			hipe-2020			ajmc			hipe-2020			ajmc		
P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
sans-contexte																	
0,755	0,757	0,756	0,829	0,806	0,817	0,754	0,730	0,742	0,910	0,877	0,893	0,604	0,563	0,583	0,789	0,859	0,823
non-temporelle																	
0,762	**0,767**	**0,765**	0,829	0,783	0,806	0,759	**0,767**	**0,763**	**0,930**	0,898	0,913	0,565	0,601	0,583	0,828	0,871	0,849
temporelle-50																	
0,765	0,765	**0,765**	0,839	0,822	0,830	0,748	0,756	0,752	0,921	**0,911**	**0,916**	**0,643**	0,617	**0,630**	0,855	0,882	0,868
temporelle-25																	
0,759	0,756	0,757	**0,848**	**0,839**	**0,844**	0,757	0,743	0,750	0,925	0,903	0,914	0,621	0,630	0,625	0,833	0,876	0,854
temporelle-10																	
0,762	0,764	0,763	**0,848**	**0,839**	**0,844**	**0,760**	0,765	0,762	0,917	0,898	0,907	0,605	**0,646**	0,625	**0,866**	**0,888**	**0,877**

TABLE 1 – Résultats sur le français, l'allemand et l'anglais, pour les deux jeux de données.

	Français		Allemand		Anglais	
	train	test	train	test	train	test
temporelle-50 / 25 / 10						
hipe-2020	120 / 154 / 217	42 / 47 / 61	325 / 393 / 482	12 / 14 / 14	192 / 222 / 246	77 / 85 / 96
ajmc	10 /12 / 12	0 / 0 / 0	71 / 71 / 73	20 / 20 / 20	2 / 2 / 2	0 / 0 / 0

TABLE 2 – Nombre de contextes remplacés par période.

soit la langue et les contextes utilisés. Nous expliquons ce comportement par la faible diversité de certains types d'entités dans ajmc. Par exemple, les dix entités les plus fréquentes du type "personne" représentent respectivement 55%, 51,5% et 62,5% de toutes les entités "personne" dans les ensembles d'entraînement, de développement et de test. Il existe également une intersection de 80% entre les dix entités les plus fréquentes des ensembles d'entraînement et de test, ce qui signifie que huit des dix entités les plus fréquentes apparaissent à la fois dans les ensembles d'entraînement et de test. Le jeu de données hipe-2020 en anglais présente les scores les plus bas par rapport au français et à l'allemand, indépendamment des contextes. Nous attribuons cette baisse de performance à l'absence d'un corpus d'entraînement en anglais.

Impact des intervalles de temps Le jeu de données ajmc en allemand contient des commentaires provenant de deux années (1853 et 1894), le ajmc en anglais provient également de deux années (1881 et 1896), tandis que le ajmc en français ne concerne qu'une seule année (1886). En raison de la taille de la collection, hipe-2020 couvre un plus grand nombre d'années. En termes de couverture, les articles en français ont été collectés de 1798 à 2018, les articles en allemand de 1798 à 1948, et les articles en anglais de 1790 à 1960. Ainsi, nous avons examiné la différence entre les contextes récupérés par les configurations temporelles et non temporelles.

Le tableau 2 résume ces différences pour les ensembles d'entraînement et de test et indique le nombre de contextes qui ont été filtrés et remplacés par non-temporelle pour chaque intervalle de temps. En général, plus l'intervalle d'années est court, plus le nombre de contextes remplacés est élevé. Nous remarquons que le nombre de contextes remplacés est plus faible pour ajmc que pour hipe-2020. Cela s'explique par la taille limitée de la plage temporelle et le manque de diversité des entités pendant cette période. En comparant avec les résultats du tableau 1, nous pouvons déduire qu'en général, l'utilisation d'intervalles de temps plus courts, tels que $\delta = 10$, est bénéfique. En effet, la configuration temporelle-10 présente les scores F1 les plus élevés.

Impact des erreurs de numérisation Les commentaires sur la littérature grecque classique de `ajmc` présentent les difficultés typiques de l'océrisation historique. Avec des mises en page complexes, souvent composées de plusieurs colonnes et lignes de texte, la qualité de la numérisation des commentaires peut avoir un impact significatif sur la NER et d'autres tâches en aval, telles que la liaison d'entités. Statistiquement, environ 10% des entités sont affectées par des erreurs d'OCR dans les corpus `ajmc` en anglais et en allemand, tandis que ce chiffre s'élève à 27,5% dans le corpus en français. Les modèles intégrant un contexte supplémentaire, en particulier les approches temporelles, contribuent à la reconnaissance correcte des entités nommées, qu'elles soient contaminées ou non par des erreurs d'OCR. Cette amélioration est particulièrement significative pour la reconnaissance des entités nommées contaminées, par rapport à celles qui ne le sont pas (même si ces dernières sont plus fréquentes). Par exemple, dans le corpus en allemand, la configuration `temporelle-50` apporte une amélioration d'environ 14 points de pourcentage par rapport au modèle de base pour les entités contaminées, tandis que cette amélioration est de seulement 2 points de pourcentage pour les entités non contaminées. En outre, les trois quarts des entités présentant un taux d'erreur sur les caractères de 67% sont correctement reconnus, tandis que le modèle de base n'en reconnaît qu'un quart. Enfin, les entités avec des taux d'erreur supérieurs à 70% ne sont pas du tout reconnues par tous les modèles.

4 Conclusions et perspectives

Dans cet article, nous avons exploré l'apport de l'injection d'informations temporelles dans la tâche de reconnaissance d'entités nommées à partir de collections historiques. Nos résultats ont démontré que l'injection de *jokers contextuels* sur de courtes périodes offre de meilleurs résultats pour les collections présentant une diversité d'entités limitée et des intervalles de temps restreints. De manière symétrique, l'utilisation de *jokers contextuels* sur une période plus longue est plus bénéfique pour les intervalles d'années plus larges. Nous avons également montré que notre approche est performante dans la détection des entités affectées par des erreurs de numérisation, même lorsque le taux d'erreur des caractères atteint 67%. Enfin, nous avons observé que la qualité des contextes récupérés dépend de l'adéquation entre la collection historique et la base de connaissances. Ainsi, dans de futures recherches, il serait intéressant d'inclure des informations sur la temporalité en prédisant les intervalles d'années à partir d'un large ensemble de pages Wikipédia, afin de les utiliser comme contextes complémentaires.

Limitations Idéalement, le système requiert des métadonnées indiquant l'année de rédaction des ensembles de données, ou du moins un intervalle temporel. Dans le cas contraire, il sera nécessaire de recourir à d'autres systèmes pour prédire l'année de publication (Rastas *et al.*, 2022). Cependant, les erreurs générées par ces systèmes se propageront et pourront influencer les résultats de la reconnaissance d'entités nommées.

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

BOROŞ E., HAMDI A., PONTES E. L., CABRERA-DIEGO L.-A., MORENO J. G., SIDERE N. & DOUCET A. (2020a). 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., NGUYEN N. K., LEJEUNE G. & DOUCET A. (2022). Assessing the impact of ocr noise on multilingual event detection over digitised documents. *International Journal on Digital Libraries*, p. 1–26.

BOROŞ E., ROMERO V., MAARAND M., ZENKLOVÁ K., KŘEČKOVÁ J., VIDAL E., STUTZMANN D. & KERMORVANT C. (2020b). A comparison of sequential and combined approaches for named entity recognition in a corpus of handwritten medieval charters. In *2020 17th International conference on frontiers in handwriting recognition (ICFHR)*, p. 79–84 : IEEE.

EHRMANN M., HAMDI A., LINHARES PONTES E., ROMANELLO M. & DOUVET A. (2023). A Survey of Named Entity Recognition and Classification in Historical Documents. *ACM Computing Surveys*.

EHRMANN M., ROMANELLO M., BIRCHER S. & CLEMATIDE S. (2020a). Introducing the CLEF 2020 HIPE shared task : Named entity recognition and linking on historical newspapers. In J. M. JOSE, E. YILMAZ, J. MAGALHÃES, P. CASTELLS, N. FERRO, M. J. SILVA & F. MARTINS, Éds., *Advances in information retrieval*, p. 524–532, Cham : Springer International Publishing.

EHRMANN M., ROMANELLO M., CLEMATIDE S., STRÖBEL P. B. & BARMAN R. (2020b). Language resources for historical newspapers : the impresso collection. In *Proceedings of The 12th Language Resources and Evaluation Conference*, p. 958–968.

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

GARCÍA-DURÁN A., DUMANČIĆ S. & NIEPERT M. (2018). Learning sequence encoders for temporal knowledge graph completion. *arXiv preprint arXiv :1809.03202*.

HAMDI A., LINHARES PONTES E., BOROS E., NGUYEN T. T. H., HACKL G., MORENO J. G. & DOUCET A. (2021). 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., SIDERE N., COUSTATY M. & DOUCET A. (2022). In-depth analysis of the impact of ocr errors on named entity recognition and linking. *Natural Language Engineering*, p. 1–24.

HOULSBY N., GIURGIU A., JASTRZEBSKI S., MORRONE B., DE LAROUSSILHE Q., GESMUNDO A., ATTARIYAN M. & GELLY S. (2019). Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, p. 2790–2799 : PMLR.

LEBLAY J. & CHEKOL M. W. (2018). Deriving validity time in knowledge graph. In *Companion Proceedings of the The Web Conference 2018*, p. 1771–1776.

LINHARES PONTES E., CABRERA-DIEGO L. A., MORENO J. G., BOROS E., HAMDI A., DOUCET A., SIDERE N. & COUSTATY M. (2022). Melhissa : a multilingual entity linking architecture for historical press articles. *International Journal on Digital Libraries*, **23**(2), 133–160.

NGUYEN N. K., BOROS E., LEJEUNE G. & DOUCET A. (2020). Impact analysis of document digitization on event extraction. In *4th workshop on natural language for artificial intelligence (NL4AI 2020) co-located with the 19th international conference of the Italian Association for artificial intelligence (AI* IA 2020)*, volume 2735, p. 17–28.

OBERBICHLER S., BOROŞ E., DOUCET A., MARJANEN J., PFANZELTER E., RAUTIAINEN J., TOIVONEN H. & TOLONEN M. (2022). Integrated interdisciplinary workflows for research on historical newspapers : Perspectives from humanities scholars, computer scientists, and librarians. *Journal of the Association for Information Science and Technology*, **73**(2), 225–239.

PFEIFFER J., VULIĆ I., GUREVYCH I. & RUDER S. (2020). MAD-X : An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 7654–7673, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.617.

RASTAS I., RYAN Y. C., TIIHONEN I., QARAEI M., REPO L., BABBAR R., MÄKELÄ E., TOLONEN M. & GINTER F. (2022). Explainable publication year prediction of eighteenth century texts with the bert model. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, p. 68–77.

REIMERS N. & GUREVYCH I. (2019). Sentence-BERT : Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 3982–3992, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1410.

REIMERS N. & GUREVYCH I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 4512–4525.

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*, p. 1–6.

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

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER Ł. & POLOSUKHIN I. (2017). Attention is all you need. *Advances in neural information processing systems*, **30**.

WANG X., GAO T., ZHU Z., ZHANG Z., LIU Z., LI J. & TANG J. (2021). Kepler : A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, **9**, 176–194.

WANG X., SHEN Y., CAI J., WANG T., WANG X., XIE P., HUANG F., LU W., ZHUANG Y., TU K. *et al.* (2022). Damo-nlp at semeval-2022 task 11 : A knowledge-based system for multilingual named entity recognition. *arXiv preprint arXiv :2203.00545*.