Atténuer les erreurs de numérisation dans la reconnaissance d'entités nommées pour les documents historiques

Emanuela Boros, Ahmed Hamdi, Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, Jose G. Moreno, Nicolas Sidere, Antoine Doucet

Atténuer les erreurs de numérisation dans la reconnaissance d'entités nommées pour les documents historiques

Emanuela Boros — Ahmed Hamdi[1] — Elvys Linhares Pontes[1] — Luis Adrián Cabrera-Diego[1] — Jose G. Moreno[1,2] — Nicolas Sidere[1] — Antoine Doucet[1]

[1] *Laboratoire L3i, Université de La Rochelle, France*
[2] *IRIT, Université de Toulouse, France*

RÉSUMÉ. *Cet article aborde la reconnaissance d'entités nommées (NER) appliquée aux textes historiques obtenus à partir du traitement d'images numériques de journaux à l'aide de techniques de reconnaissance optique de caractères (OCR). Nous soutenons que le principal défi pour cette tâche est que le processus OCR produit des textes contenant entre autres des fautes d'orthographe et des erreurs de syntaxes. De plus, des variations sémantiques peuvent être présentes dans les documents anciens, ce qui a un impact sur les performances de la reconnaissance d'entités nommées. Nous menons une évaluation comparative à l'état de l'art de deux ensembles de données historiques en allemand et en français, et nous proposons un modèle basé sur une pile hiérarchique de couches Transformer pour aborder la reconnaissance d'entités nommées dans des données historiques. Nos résultats montrent que le modèle proposé améliore clairement les résultats sur les deux ensembles de données.*

ABSTRACT. *This paper tackles the task of NER applied to historical texts obtained from processing digital images of newspapers using OCR techniques. The main challenge for this task is that the OCR process leads to misspellings and linguistic errors in the output text, which can impact the performance of the NER. We conduct a comparative evaluation on two historical datasets in German and French against previous state-of-the-art models, and we propose a model based on a hierarchical stack of Transformers to approach the NER task for historical data. Our findings show that the proposed model clearly improves the results on both historical datasets.*

MOTS-CLÉS : *Extraction d'information, reconnaissance d'entités nommées, données multilingues, données historiques.*

KEYWORDS: *Information extraction, Named entity recognition, Multilingual data, Historical data*

1. Introduction

Avec la numérisation à grande échelle de contenus patrimoniaux, le besoin de rendre efficacement accessible les documents historiques à l'aide de technologies appropriées a très fortement augmenté. Dans le même temps, il existe un intérêt croissant pour l'extraction d'informations pertinentes à partir de sources historiques. Dans cet article, nous abordons la tâche de la reconnaissance d'entités nommées (NER). qui vise à identifier des entités du monde réel, telles que les noms de personnes, d'organisations et de lieux à partir des textes bruts.

Alors que la plupart des travaux de recherche se concentrent sur les ensembles de données contemporains, les performances des systèmes NER ont augmenté à un rythme rapide, grâce à la capacité de représentation des réseaux de neurones. Plus récemment, les modèles NER basés sur des représentations contextuelles de mots et de chaînes de caractères fournis par Flair (Akbik *et al.*, 2018) ou BERT (Devlin *et al.*, 2019) ont permis des améliorations impressionnantes. Les architectures (Vaswani *et al.*, 2017) basées sur Transformer pour NER sont devenues populaires depuis la sortie du modèle BERT.

Pour extraire des entités de documents historiques, les outils NER sont confrontés à des défis supplémentaires. La majorité de ces documents est numérisée et traitée par un outil de reconnaissance optique de caractères (OCR) pour transcrire le texte. Cependant, la sortie de l'OCR peut potentiellement contenir des erreurs. Cela est principalement dû à la qualité de l'outil ou encore à la dégradation des documents numérisés en particulier pour les documents historiques. Cela conduit à des erreurs dans le texte transcrit, notamment des emplacements ou des noms de personnes mal orthographiés, ce qui est problématique puisque ce type d'entité nommée fait fréquemment partie des requêtes soumises aux collections patrimoniales. Pour relever ces défis nous proposons un modèle NER robuste basé sur une pile de *Transformers* qui comprend des encodeurs BERT affinés. Nous étudions l'impact d'un tel modèle, et nous concluons que ce type de modèle est adapté à l'extraction d'entités à partir de documents historiques. Le travail présenté ici est décrit plus en détail dans (Boroş *et al.*, 2020).

2. Ensembles de données

Des expériences ont été menées sur deux jeux de données issues de presse ancienne numérisée HIPE et NEWSEYE. Chaque ensemble propose deux corpus en français et en allemand. L'ensemble de données HIPE a été créé par le défi HIPE du laboratoire d'évaluation CLEF 2020 (Ehrmann *et al.*, 2020a). Il est composé d'articles de plusieurs journaux historiques suisses, luxembourgeois et américains publiés de 1790 à 2010 (Ehrmann *et al.*, 2020b).

Nous utilisons également l'ensemble de données NEWSEYE, composé de journaux historiques en français (1814-1944) et en allemand (1845-1945). Les documents ont été collectés auprès des bibliothèques nationales de France et d'Autriche (ONB), respectivement. HIPE et NEWSEYE utilisent des guides d'annotation similaires et com-

patibles entre eux. A l'exception de l'entité *TIME* qui est utilisé uniquement dans HIPE, toutes les autres classes sont identiques dans les deux jeux de données.

3. Modèle

D'abord, nous utilisons un modèle BERT pré-entraîné, et nous ajoutons ensuite n blocs *Transformer* par-dessus, finalisés avec une couche de prédiction CRF. Nous appelons ce modèle BERT $+n\times$Transf où n est un hyper-paramètre faisant référence au nombre de couches de *Transformer*.

Néanmoins, malgré l'impact majeur de BERT, les chercheurs s'interrogent sur la capacité de ce modèle à traiter des contenus bruités (Sun *et al.*, 2020) à moins que des techniques complémentaires ne soient utilisées (Muller *et al.*, 2019 ; Pruthi *et al.*, 2019). En plus de BERT, nous ajoutons ainsi une pile de blocs *Transformer* (Vaswani *et al.*, 2017) (encodeurs). Nous supposons que les couches *Transformer* complémentaires permettent d'atténuer la sensibilité du lemmatiseur intégré de BERT aux erreurs OCR tels que les mots hors vocabulaire (OOV) ou les fautes d'orthographe, et contribuer à l'apprentissage et à la reconnaissance du contexte des entités.

4. Éxpériences

Nous avons choisi comme base le modèle proposé par (Ma et Hovy, 2016), un modèle end-to-end combinant un encodage de caractères BiLSTM et CNN, afin de profiter des fonctionnalités de mots et de caractères[1]. L'analyse au niveau caractère est connue comme permettant de capturer des informations morphologiques et de forme (Kanaris *et al.*, 2007 ; Santos et Zadrozny, 2014 ; dos Santos et Guimarães, 2015).

L'évaluation de la tâche NER se fait avec le niveau entité comme unité de référence (Makhoul *et al.*, 1999). Nous calculons la précision (P), le rappel (R) et la mesure F1 (F1) au niveau micro, c'est-à-dire que les types d'erreur sont considérés sur tous les documents. Deux scénarios d'évaluation ont été considérés : *micro-strict*, qui recherche une correspondance exacte des entités, et *micro-fuzzy*, où une prédiction est correcte lorsqu'il y a au moins un chevauchement de tokens (Ehrmann *et al.*, 2020a). En outre, la significativité statistique est mesurée par un test t bilatéral, avec une valeur p estimée entre 0,01 et 0,05 (* dénote une amélioration significative par rapport au modèle d'avant à $p \leq 0{,}05$, ** dénote $p \leq 0{,}01$).

À partir des résultats de la table 1, nous pouvons voir la preuve que les modèles basés sur BERT avec $n\times$ Transf atteignent, pour les ensembles de données et les langues, des textit micro-fuzzy et textit micro-strict valeurs de performance que le modèle BERT autonome et les modèles de base. Tous les modèles ont une signification

1. Une description détaillée du modèle et des hyperparamètres peut être trouvée dans (Ma et Hovy, 2016).

	HIPE						NEWSEYE					
	DE			FR			DE			FR		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
BiLSTM-CNN												
fuzzy	83.3	70.1	76.1	89.9	83.9	86.8	81.2	42.4	55.7	82.2	77.2	79.6
strict	69.4	58.4	63.4	77.7	72.5	75.0	54.8	28.6	37.6	65.5	61.4	63.4
BERT												
fuzzy	83.4	88.3	85.8**	89.5	91.9	90.7*	60.1	67.0	63.4**	86.1	81.8	83.9**
strict	74.1	78.5	76.2**	81.1	83.3	82.1*	46.8	52.2	49.4**	70.1	66.6	68.3**
BERT+1×Transf												
fuzzy	85.8	87.3	86.5**	91.3	92.9	**92.1****	82.3	66.4	**73.5****	88.7	82.1	**85.3****
strict	77.2	78.6	77.9**	83.5	84.9	**84.2****	62.7	50.6	56.0**	74.4	68.9	**71.5****
BERT+2×Transf												
fuzzy	87.0	87.2	**87.1****	91.5	92.4	91.9**	83.3	64.4	72.6**	89.7	80.1	84.7 **
strict	78.6	78.7	**78.7****	83.4	84.2	83.8**	64.9	50.2	**56.6****	75.0	67.0	70.8**

Tableau 1 : Résultats sur les ensembles de données HIPE et NEWSEYE en français et en allemand.

statistique $< 0,01$, ainsi, l'ajout de $n\times$ Transf peut améliorer la généralisabilité du modèle pour le NER sur les documents historiques.

De plus, ils parviennent généralement à maintenir un équilibre entre rappel et précision, alors que les modèles de référence varient selon la langue. On remarque également que, si en général les deux modèles obtiennent un équilibre entre rappel et précision, il existe un déséquilibre important dans le cas du jeu de données allemand NEWSEYE. BERT $+n\times$ Transf réduit la différence à 20 points, là où les méthodes de référence souffrent d'une différence de 40%.

5. Conclusions et perspectives

Nous avons présenté une architecture d'apprentissage profond pour le NER basé sur un encodeur BERT affiné et plusieurs blocs *Transformer*. Les résultats sur les deux jeux de données historiques en français et en allemand ont montré la capacité de l'approche proposée à traiter des corpus de textes numérisés bruités dans des langues distinctes. Si les améliorations apportées par le modèle NER proposé sont claires, notre analyse des résultats a mis en évidence plusieurs facteurs susceptibles d'influencer les résultats. Une analyse plus approfondie reste à mener. Nous comptons ainsi étudier les variations détaillées de notre architecture de manière plus approfondie.

Remerciements

Ce travail a été soutenu par le programme de recherche et d'innovation Horizon 2020 de l'Union européenne au titre des subventions 770299 (NewsEye) et 825153 (Embeddia).

6. Bibliographie

Akbik A., Blythe D., Vollgraf R., « Contextual string embeddings for sequence labeling », *Proceedings of the 27th International Conference on Computational Linguistics*, p. 1638-1649, 2018.

Boroş E., Hamdi A., Pontes E. L., Cabrera-Diego L.-A., Moreno J. G., Sidere N., Doucet A., « Alleviating Digitization Errors in Named Entity Recognition for Historical Documents », *Proceedings of the 24th Conference on Computational Natural Language Learning*, p. 431-441, 2020.

Devlin J., Chang M.-W., Lee K., Toutanova K., « BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding », *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)*, Association for Computational Linguistics, Minneapolis, Minnesota, p. 4171-4186, June, 2019.

dos Santos C., Guimarães V., « Boosting Named Entity Recognition with Neural Character Embeddings », *Proceedings of the Fifth Named Entity Workshop*, Association for Computational Linguistics, Beijing, China, p. 25-33, July, 2015.

Ehrmann M., Romanello M., Bircher S., Clematide S., « Introducing the CLEF 2020 HIPE Shared Task : Named Entity Recognition and Linking on Historical Newspapers », *European Conference on Information Retrieval*, Springer, p. 524-532, 2020a.

Ehrmann M., Romanello M., Clematide S., Ströbel P. B., Barman R., « Language Resources for Historical Newspapers : the Impresso Collection », *Proceedings of The 12th Language Resources and Evaluation Conference*, p. 958-968, 2020b.

Kanaris I., Kanaris K., Houvardas I., Stamatatos E., « Words versus character n-grams for anti-spam filtering », *International Journal on Artificial Intelligence Tools*, vol. 16, n^{o} 06, p. 1047-1067, 2007.

Ma X., Hovy E., « End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF », *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, Association for Computational Linguistics, Berlin, Germany, p. 1064-1074, August, 2016.

Makhoul J., Kubala F., Schwartz R., Weischedel R. *et al.*, « Performance measures for information extraction », *Proceedings of DARPA broadcast news workshop*, Herndon, VA, p. 249-252, 1999.

Muller B., Sagot B., Seddah D., « Enhancing BERT for Lexical Normalization », *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, p. 297-306, 2019.

Pruthi D., Dhingra B., Lipton Z. C., « Combating Adversarial Misspellings with Robust Word Recognition », *57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, Florence, Italy, p. 5582-5591, 2019.

Santos C. d., Zadrozny B., « Learning character-level representations for part-of-speech tagging », *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, p. 1818-1826, 2014.

Sun L., Hashimoto K., Yin W., Asai A., Li J., Yu P., Xiong C., « Adv-BERT : BERT is not robust on misspellings ! Generating nature adversarial samples on BERT », *arXiv preprint arXiv :2003.04985*, 2020.

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser Ł., Polosukhin I., « Attention is all you need », *Advances in neural information processing systems*, p. 5998-6008, 2017.