Benchmarking du jeu de données NAS pour la séparation d'articles dans la presse ancienne*

Nancy Girdhar[1,*], Mickaël Coustaty[1] and Antoine Doucet[1]

[1]L3i, Université de La Rochelle, La Rochelle, France

Résumé

Cet article concerne l'accessibilité de collections de presse ancienne. L'un des principaux défis à relever pour rendre les contenus accessibles est l'extraction d'articles individuels à partir d'images de pages numérisées en vue d'exploiter les documents à la granularité adéquate. Nous évaluons le jeu de données *N*ewsEye ***A***rticle ***S***eparation (NAS), qui contient des pages de presse ancienne des 19e et 20e siècles en allemand, finnois et français. NAS représente un défi en raison de la diversité des mises en page et des styles de police. Nous introduisons en outre de nouvelles mesures, notamment le *taux d'erreur des articles*, le *score de couverture des articles*, le *taux d'articles correctement prédits* et la *segmentation*, afin d'évaluer les performances des modèles. Le jeu de données NAS est accessible au public. Cette soumission est le résumé traduit d'un article publié à la conférence ICADL 2023 [1].

Mots-Clés

multilinguisme, jeux de données, séparation d'articles, documents historiques, presse ancienne

1. Introduction

Cet article se concentre sur le défi de l'accessibilité des journaux historiques, en particulier l'étape cruciale de la séparation des articles au sein des pages numérisés. Les méthodes existantes, qui vont des méthodes basées sur des règles [2, 3] à l'apprentissage profond [4], utilisent principalement des jeux de données contemporains pour évaluer les modèles de données historiques, soulignant la rareté des jeux de données historiques accessibles au public [5, 4, 6, 7, 8]. Pour combler cette lacune, le jeu de données multilingues NAS (***N***ewsEye ***A***rticle ***S***eparation) est introduit [1], issu de bibliothèques nationales européennes et offrant diverses mises en page et styles de police. En outre, de nouvelles métriques d'évaluation sont proposées pour mesurer les performances des modèles pour la tâche de séparation des articles. NAS est accessible au public et vise à faire progresser l'accessibilité des collections de presse ancienne, en fournissant une référence pour les modèles de séparation d'articles [9, 10, 11].

Des directives d'annotation issues du projet Horizon 2020 NewsEye [10] et respectant le format PAGE ont été mises en œuvre à l'aide de la plateforme Transkribus [12]. Le processus d'annotation en deux étapes comprend l'annotation des éléments de mise en page (*lignes de texte et régions*) et la délimitation des frontières de l'article sur la base de ces éléments de mise en page. Les régions sont classées en *Texte*, *Image*, *Table*, *Advert* et *Separator*, avec des balises de structure

. nancy.girdhar@univ-lr.fr (N. Girdhar); mickael.coustaty@univ-lr.fr (M. Coustaty); antoine.doucet@univ-lr.fr (A. Doucet)

. 0000-0002-1009-3875 (N. Girdhar); 0000-0002-0123-439X (M. Coustaty); 0000-0001-6160-3356 (A. Doucet)

Table 1
Résultats de référence pour la tâche de séparation des articles sur le jeu de données NAS.

Dataset	Méthode	Évaluation		
		mACS	**mAER**	**mPPA**
ONB	DBSCAN	**0.5844**	**0.4156**	**0.1914**
	Greedy	0.5360	0.4640	0.1736
	Hierarchical	0.5545	0.4455	0.1659
NLF	DBSCAN	0.3751	0.6249	**0.0662**
	Greedy	0.3275	0.6725	0.0548
	Hierarchical	**0.3823**	**0.6177**	0.0615
BNF	DBSCAN	0.4470	0.5530	0.1057
	Greedy	0.3568	0.6432	0.0940
	Hierarchical	**0.5382**	**0.4618**	**0.1393**

supplémentaires pour les *TextRegions*. NAS comprend un total de 613 pages scannées du 19e et du début du 20e siècle [10] en français (BNF), en finnois (NLF) et en allemand (ONB) [1]. Un prétraitement consiste à supprimer les blocs de texte dépourvus de lignes de base ou d'étiquettes d'articles afin d'améliorer la qualité de la vérité de terrain.

2. Expériences et résultats

Les résultats des expériences utilisant le modèle de référence NewsEye pour la séparation des articles sur le jeu de données NAS [2] [13, 14, 15] sont présentés dans le tableau 1. L'évaluation s'appuie sur de nouvelles mesures : *taux d'erreur des articles* (AER), mesurant l'écart entre les régions de texte prédites et réelles ; *score de couverture des articles* (ACS), qui quantifie les régions d'articles extraites avec succès ; et *taux d'article prédit correct* (PPA), qui calcule le rapport entre les articles prédits avec précision et le nombre total d'articles de référence. Les résultats soulignent que, pour le jeu de données de l'ONB, *DBSCAN* obtient le score ACS moyen le plus élevé, tandis que pour les jeux de données NLF et BNF, *Hierarchical* est plus performant. Toutefois, il convient de noter qu'aucun des modèles ne reconstruit entièrement les articles, comme le montrent les faibles scores moyens de PPA.

3. Conclusion

Ce travail aborde le défi de la séparation des articles dans la presse ancienne en introduisant le jeu de données NAS et de nouvelles mesures d'évaluation associées. Ce jeu de données multilingues offre une ressource unique, couvrant des pages de journaux allemands, finlandais et français des XIXe siècle et XXe siècles. Les directives d'annotation sont accessibles au public et les métriques introduites permettent l'évaluation des performances des modèles. NAS est accessible aux chercheurs [3]. Nous pensons que la mise à disposition d'une telle référence facilitera les futurs travaux de recherche en séparation d'articles.

1. BNF : https://bnf.fr ; NLF : https://kansalliskirjasto.fi ; ONB : https://onb.ac.at
2. https://github.com/CITlabRostock/citlab-article-separation-new
3. BNF : https://zenodo.org/records/5654841 ; NLF : https://zenodo.org/records/5654858 ; ONB : https://zenodo.org/records/5654907

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.

Références

[1] N. Girdhar, M. Coustaty, A. Doucet, Benchmarking nas for article separation in historical newspapers, in : International Conference on Asian Digital Libraries, Springer, 2023, pp. 76–88.

[2] B. Gatos, S. Mantzaris, K. Chandrinos, A. Tsigris, S. J. Perantonis, Integrated algorithms for newspaper page decomposition and article tracking, in : Proceedings of the Fifth International Conference on Document Analysis and Recognition. ICDAR'99 (Cat. No. PR00318), IEEE, 1999, pp. 559–562.

[3] T. Palfray, D. Hebert, S. Nicolas, P. Tranouez, T. Paquet, Logical segmentation for article extraction in digitized old newspapers, in : Proceedings of the 2012 ACM symposium on Document engineering, 2012, pp. 129–132.

[4] B. Meier, T. Stadelmann, J. Stampfli, M. Arnold, M. Cieliebak, Fully convolutional neural networks for newspaper article segmentation, in : 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), volume 1, IEEE, 2017, pp. 414–419.

[5] R. Cohen, A. Asi, K. Kedem, J. El-Sana, I. Dinstein, Robust text and drawing segmentation algorithm for historical documents, in : Proceedings of the 2nd International Workshop on Historical Document Imaging and Processing, 2013, pp. 110–117.

[6] S. A. Oliveira, B. Seguin, F. Kaplan, dhsegment : A generic deep-learning approach for document segmentation, in : 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), IEEE, 2018, pp. 7–12.

[7] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr, et al., Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers, in : Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 6881–6890.

[8] W. Zhu, N. Sokhandan, G. Yang, S. Martin, S. Sathyanarayana, Docbed : A multi-stage ocr solution for documents with complex layouts, in : Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, 2022, pp. 12643–12649.

[9] A. Doucet, M. Gasteiner, M. Granroth-Wilding, M. Kaiser, M. Kaukonen, R. Labahn, J.-P. Moreux, G. Muehlberger, E. Pfanzelter, M.-È. Thérenty, et al., Newseye : A digital investigator for historical newspapers, in : 15th Annual International Conference of the Alliance of Digital Humanities Organizations, DH 2020, 2020.

[10] M. Johannes, M. Weidemann, R. Labahn, A. Doucet, Newseye : A digital investigator for historical newspapers, https://www.newseye.eu/fileadmin/deliverables/NewsEye-T23-D27-ArticleSeparation-c-final-Submitted-v6.0.pdf, 2022. (Accessed on 05/26/2023).

[11] J. Michael, M. Weidemann, B. Laasch, R. Labahn, Icpr 2020 competition on text block segmentation on a newseye dataset, in : Pattern Recognition. ICPR International Workshops and Challenges : Virtual Event, January 10-15, 2021, Proceedings, Part VIII, Springer, 2021, pp. 405–418.

[12] S. Colutto, P. Kahle, H. Guenter, G. Mühlberger, Transkribus. a platform for automated text recognition and searching of historical documents, in : 2019 15th International Conference on eScience (eScience), IEEE, 2019, pp. 463–466.

[13] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert : Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[14] G. Andrade, G. Ramos, D. Madeira, R. Sachetto, R. Ferreira, L. Rocha, G-dbscan : A gpu accelerated algorithm for density-based clustering, Procedia Computer Science 18 (2013) 369–378.

[15] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et al., A density-based algorithm for discovering clusters in large spatial databases with noise., in : kdd, volume 96, 1996, pp. 226–231.