STRAS : Une approche à base de règles sémantiques et d'indices textuels pour la séparation des articles dans les journaux historiques ⋆

Nancy Girdhar, Mickaël Coustaty, Antoine Doucet

▶ **To cite this version:**

Nancy Girdhar, Mickaël Coustaty, Antoine Doucet. STRAS : Une approche à base de règles sémantiques et d'indices textuels pour la séparation des articles dans les journaux historiques ⋆. CORIA-RJCRI 2024 (COnférence en Recherche d'Information et Applications), Apr 2024, La Rochelle, France. hal-04676745

HAL Id: hal-04676745
https://hal.science/hal-04676745v1

Submitted on 23 Aug 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

STRAS : Une approche à base de règles sémantiques et d'indices textuels pour la séparation des articles dans les journaux historiques*

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

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

Résumé

Cet article présente STRAS, une approche à base de règles qui s'appuie sur des indices textuels sémantiques pour la séparation des articles dans les journaux historiques. En utilisant des encastrements de régions de texte, notre approche catégorise et sépare avec succès les articles dans les journaux français et finlandais des 19ème et 20ème siècles. Parmi les modèles évalués (sgSTRAS, cbowSTRAS, ftSTRAS, preSTRAS), sgSTRAS démontre une performance supérieure sur les deux ensembles de données, soulignant l'importance des caractéristiques sémantiques du texte. Dans l'ensemble, STRAS représente une avancée prometteuse dans l'analyse des journaux historiques, en relevant les défis de la mise en page et en suggérant des pistes d'amélioration pour la tâche AS. Cette soumission est le résumé traduit d'un article publié à la conférence ICADL 2023 qui y a obtenu le prix du meilleur article [1].

Mots-Clés

analyse sémantique, segmentation logique, séparation des articles, presse ancienne, text embedding

1. Introduction

Les journaux historiques sont des ressources inestimables pour analyser le passé, mais leur numérisation pose des défis, en particulier en ce qui concerne la séparation des articles (AS). Les méthodes existantes, principalement basées sur des indices visuels, se heurtent à la diversité des mises en page et à la qualité des supports. Cet article propose une nouvelle approche, indépendante de la mise en page, axée sur les caractéristiques textuelles, appelée ***S***emantic ***T***extual-cues leveraged ***R***ule-based ***A***rticle ***S***eparation (STRAS) [1]. Elle utilise des indices sémantiques et des règles dérivées des structures syntaxiques, et démontre son efficacité dans la gestion de formats et mises en page variés, offrant une solution flexible pour la tâche d'AS.

2. Méthodologie Proposée

Utilisant des règles de grammaire syntaxique de base sur les caractéristiques textuelles, l'algorithme de STRAS traite les données d'entrée, y compris les fichiers PAGE au format ALTO et les images de journaux scannées dans différents formats (*.jpg*, *.png*, ou *.tif*). La figure 1

. 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)

. © 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

présente l'approche en trois étapes : (i) la génération de l'intégration des régions de texte, (ii) le calcul de la similarité des régions de texte et la catégorisation en catégories dépendantes et indépendantes, avec une sous-classification supplémentaire en régions de texte source ou non source, et (iii) l'extraction d'articles basée sur des règles.

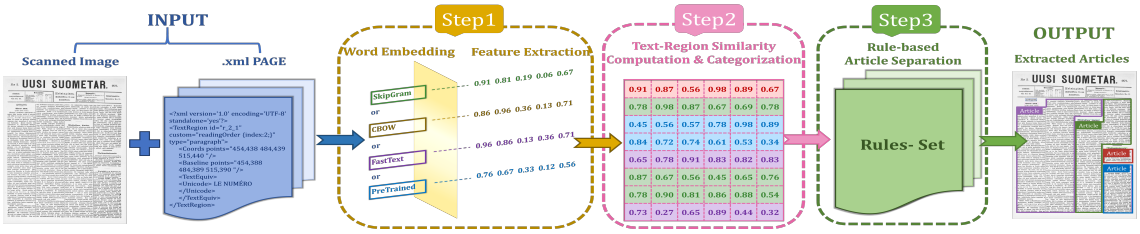

Figure 1 : Cadre de la méthodologie STRAS proposée.

3. Expériences et résultats

Pour l'expérimentation, des ensembles de données de journaux historiques français (BNF) et finlandais (NLF) [1] ont été choisis dans le cadre du projet NewsEye [2]. Outre l'évaluation de l'intersection sur l'union (IoU), trois nouvelles mesures d'évaluation sont introduites : *taux d'erreur des articles* (AER), *score de couverture des articles* (ACS) et *taux d'article prédit correct* (PPA) [3]. Les embeddings de texte ont été extraits à l'aide de modèles *SpaCy* [2] pré-entraînés et de nos propres modèles entraînés *skip-gram* [4] (sgSTRAS), *sac de mots continu* [4] (cbowSTRAS), et *FastText* [5] (ftSTRAS). Un seuil de $0,9$ a été appliqué pour la similarité en cosinus, et la comparaison a été effectuée avec la baseline [3] de NewsEye pour la tâche d'AS.

Les résultats, présentés dans le tableau 1, montrent la supériorité de STRAS par rapport à la référence, en particulier en ce qui concerne le score PPA moyen. Le modèle *sgSTRAS* surpasse systématiquement les autres, soulignant son efficacité à capturer les informations contextuelles et les relations sémantiques. L'analyse comparative met en évidence l'importance du choix des méthodes d'intégration appropriées, le skip-gram et le CBoW apparaissant comme de bons candidats. Les résultats soulignent que l'entraînement des modèles sur des données spécifiques à un domaine améliore la performance par rapport à l'utilisation de modèles pré-entraînés.

4. Conclusion

Nous avons présenté STRAS, une nouvelle approche basée sur des règles de similarité textuelle sémantique pour la séparation d'articles de journaux historiques. En s'appuyant sur des modèles entraînés et pré-entraînés, notre approche indépendante de la mise en page a démontré une segmentation efficace sur les ensembles de données français et finlandais. Elle a largement surpassé la référence, soulignant l'importance des données spécifiques au domaine pour améliorer la segmentation, tout en reconnaissant l'efficacité des modèles pré-entraînés.

1. BNF : https://bnf.fr ; NLF : https://kansalliskirjasto.fi
2. https://spacy.io/models/fr#fr_core_news_lg et https://spacy.io/models/fi#fi_core_news_lg
3. https://github.com/CITlabRostock/citlab-article-separation-new

Table 1
Résultats sur des ensembles de données de journaux historiques pour la tâche de séparation des articles (*pre* : pré-entraîné ; *sg* : skip-gram ; *cbow* : sac de mots continu ; *ft* : texte rapide). Dans la colonne *Modèle*, l'exposant * indique la baseline AS.

Dataset	Modèle	mACS	mPPA	mIoU
BNF [6]	preSTRAS	0.8006	0.6347	0.7878
	ftSTRAS	0.7984	0.6122	0.7867
	cbowSTRAS	0.8067	0.6310	0.7997
	sgSTRAS	**0.8343**	**0.7003**	**0.8238**
	dbscanAS*	0.4470	0.1057	-
	greedyAS*	0.3568	0.0940	-
	hierarchicalAS*	0.5382	0.1393	-
NLF [7]	preSTRAS	0.7905	0.6065	0.8247
	ftSTRAS	0.8271	0.7001	0.8664
	cbowSTRAS	0.8552	**0.7924**	0.8687
	sgSTRAS	**0.8611**	0.7857	**0.8774**
	dbscanAS*	0.37511	0.0662	-
	greedyAS*	0.3275	0.0548	-
	hierarchicalAS*	0.3823	0.0615	-

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, Stras : A semantic textual-cues leveraged rule-based approach for article separation in historical newspapers, in : International Conference on Asian Digital Libraries, Springer, 2023, pp. 89–105.

[2] 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.

[3] 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.

[4] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781 (2013).

[5] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Transactions of the association for computational linguistics 5 (2017) 135–146.

[6] G. Muehlberger, G. Hackl, NewsEye / READ AS training dataset from French Newspapers (19th, early 20th C.), 2021. doi :10.5281/zenodo.4600636.

[7] G. Muehlberger, G. Hackl, NewsEye / READ AS training dataset from Finnish Newspapers (19th C.), 2021. doi :10.5281/zenodo.4600746.