Actes de CORIA-TALN-RJCRI-RECITAL 2025
Marseille, France, 30 juin au 4 juillet 2025

Génération augmentée de récupération pour les journaux historiques

The Trung Tran[1] Carlos-Emiliano González-Gallardo[2] Antoine Doucet[3]
(1) ICTLab, Université des sciences et des technologies de Hanoï, Vietnam
(2) LIFAT, Université de Tours, France
(3) L3i, Université de la Rochelle, France
trungtt.bi12-452@st.usth.edu.vn, gonzalezgallardo@univ-tours.fr,
antoine.doucet@univ-lr.fr

RÉSUMÉ

La numérisation des archives historiques permet d'améliorer leur accessibilité et leur préservation à long terme, ouvrant ainsi de nouvelles perspectives de recherche interdisciplinaire. Cependant, l'ampleur des données disponibles pose des défis considérables. Diverses tâches de traitement automatique du langage naturel, telles que la reconnaissance d'entités nommées et la segmentation en articles, ont permis de faciliter l'accès du public en extrayant et structurant l'information. Néanmoins, l'agrégation des articles de presse historiques demeure largement inexplorée. Ce travail met en évidence le potentiel d'un cadre de génération augmentée de récupération (RAG), combinant des grands modèles de langage, un module de recherche sémantique et des bases de connaissances, pour agréger des articles de journaux historiques. Nous proposons également des métriques d'évaluation des systèmes génératifs ne nécessitant pas de vérité de terrain. Les premiers résultats de notre chaîne de traitement RAG sont prometteurs, démontrant que la récupération sémantique, renforcée par le reranking et la reconnaissance d'entités nommées, peut atténuer les erreurs d'océrisation et les fautes de frappe dans les requêtes.

ABSTRACT

Retrieval Augmented Generation for Historical Newspapers

The digitization of historical records significantly enhances their accessibility and long-term preservation, opening new research opportunities across disciplines. However, the vast amount of data poses challenges for effective analysis. Various natural language processing tasks, such as named entity recognition and article separation, have been developed to facilitate public access by extracting and structuring information. Yet, historical newspaper article aggregation remains largely unexplored. This work showcases the potential of a retrieval-augmented generation framework (RAG) that combines large language models, a semantic retrieval module, and knowledge bases to aggregate historical newspaper articles. We also propose metrics for evaluating generative systems without requiring ground truth. Initial results from our RAG pipeline are promising, demonstrating that semantic retrieval, aided by reranking and NER, can mitigate OCR errors and query misspellings.

MOTS-CLÉS : Humanités numériques, Génération augmentée de récupération, Grands modèles de langage.

KEYWORDS: Digital Humanities, Retrieval-Augmented Generation, Large Language Models.

ARTICLE : **Accepté à** JCDL'24. *Tran, T. T., González-Gallardo, C. E., & Doucet, A. (2024, De-*

cember). Retrieval Augmented Generation for Historical Newspapers. In Proceedings of the 24th ACM/IEEE Joint Conference on Digital Libraries (pp. 1-5) (LIEN) (OPEN ACCESS).

---

1 Résumé de l'article

La numérisation massive des archives historiques a créé de nouvelles opportunités pour les chercheur·e·s et les étudiant·e·s, en particulier dans le domaine des humanités numériques. Cependant, le travail avec les collections de journaux historiques présente plusieurs défis, notamment des données non structurées, des erreurs d'océrisation et des variations linguistiques. Ces obstacles rendent difficile la récupération et l'analyse efficace des informations.

Des projets récents tels que NewsEye [1] (Doucet *et al.*, 2020) ou Impresso [2] ont abordé l'analyse des journaux historiques en utilisant la reconnaissance d'entités nommées (REN) avec des bases de connaissances (Boros *et al.*, 2022; González-Gallardo *et al.*, 2023a). Cette première étape permet d'obtenir des informations structurées, mais elle reste limitée en termes d'agrégation des informations. La capacité des grands modèles de langage (GML) à générer du texte ouvre la voie à une simplification de l'analyse des journaux historiques ; cependant, ils présentent des problèmes d'hallucination (González-Gallardo *et al.*, 2023b, 2024). La génération augmentée de récupération (RAG), qui combine récupération et génération, atténue ces limitations (Lewis *et al.*, 2020). Les recherches existantes sur le RAG ont optimisé les processus d'indexation, de récupération et de classement, mais leur application aux journaux historiques restait inexplorée.

Le système RAG proposé et détaillé dans l'article d'origine (Tran *et al.*, 2025) se compose des éléments clés suivants :

- **Construction de la base de données :** Les articles de journaux historiques sont d'abord découpés et convertis en espace vectoriel à l'aide du modèle multilingue E5. Les données sont indexées dans la base de données vectorielle Chroma, et une base de données supplémentaire pour les métadonnées titre/résumé est créée.
- **Routage des requêtes :** Un module pré-RAG détermine s'il est possible de répondre à la requête en utilisant la base de données. Si la récupération échoue en raison de faibles scores de similarité, la requête est transmise à un module de recherche web pour la récupération d'informations externes. Sinon, le système passe au module de récupération principal.
- **Récupération et reclassement :** Les documents récupérés sont d'abord traités par une étape de récupération de documents standard, suivie d'une reclassement. Le module de reclassement intègre le module de reclassement multilingue Cohere ainsi qu'un score de reclassement basé sur la REN pour gérer les variations des entités nommées et réduire les erreurs d'océrisation.
- **Génération de réponses :** Les informations récupérées sont ensuite agrégées en une instruction générative (prompt) et envoyées à LLaMA3 afin de générer la réponse finale.

Nous avons d'abord testé le système sur les sous-ensembles français et finnois de Miracl (Zhang *et al.*, 2023), un ensemble de données multilingue destiné à l'évaluation de la recherche d'information. Les métriques Recall@100 et NDCG@10 ont été utilisées pour l'évaluation. Les résultats montrent que l'encodage dense dépasse nettement les méthodes traditionnelles basées sur des représentations éparses. De plus, l'intégration du reclassement comme filtre final améliore les résultats finaux.

1. https://www.newseye.eu/
2. https://impresso.github.io/

Cependant, dans ce cas, le module de reclassement classique surpasse légèrement la version intégrant le score REN. Cela peut s'expliquer par le fait que les données testées ne sont pas orientées vers les entités nommées, ce qui limite l'efficacité de cette approche.

L'évaluation suivante a été menée sur le module de génération de réponses. Dans cette partie, nous proposons quatre métriques permettant d'évaluer le modèle sans recourir à une vérité terrain. Ces métriques reposent sur l'hypothèse que la réponse finale doit être une agrégation de toutes les informations récupérées à l'étape précédente. En nous basant sur cette hypothèse, nous avons défini quatre métriques distinctes : LLM score, BERTscore, similarité cosinus et indice de qualité. Les résultats montrent que le modèle LLaMA3 obtient des performances nettement supérieures pour les langues bien dotées en ressources, comme l'anglais et le français, mais qu'il rencontre des difficultés avec les langues à faibles ressources. Par ailleurs, l'analyse de la corrélation entre ces métriques révèle que BERTscore et la similarité cosinus sont fortement corrélées, ce qui pourrait constituer une piste d'expérimentation pour de futurs travaux. LLMscore montre une corrélation moyenne, tandis que l'indice de qualité présente la corrélation la plus faible dans toutes les langues testées.

Enfin, l'évaluation qualitative a été réalisée sur des exemples spécifiques issus des données de NewsEye (Girdhar *et al.*, 2023). Cette analyse montre que, pour les données où les entités nommées jouent un rôle central, l'intégration de la REN a amélioré de manière significative les scores finaux de reclassement.

En conclusion, cette étude présente un système RAG conçu pour l'agrégation et l'analyse de la presse ancienne. L'intégration d'encodages denses, de modules de reclassement et de modèles génératifs a montré des résultats prometteurs en améliorant l'accès aux archives journalistiques historiques. Toutefois, certains défis persistent, notamment les erreurs d'océrisation et la performance des GMLs sur les langues à faibles ressources. Les travaux futurs viseront à affiner chaque module, notamment en ajustant les GMLs pour mieux gérer les erreurs d'océrisation, en renforçant la robustesse des modèles de récupération et en améliorant les métriques d'évaluation pour mieux capter la qualité des réponses générées.

Remerciements

Ce travail a été soutenu par les projets ANNA (2019-1R40226), TERMITRAD (2020-2019-8510010), Pypa (AAPR2021-2021-12263410) et Actuadata (AAPR2022-2021-17014610), financés par la Région Nouvelle-Aquitaine (France).

Références

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. In *CLEF (Working Notes)*, p. 1064–1078.

DOUCET A., GASTEINER M., GRANROTH-WILDING M., KAISER M., KAUKONEN M., LABAHN R., MOREUX J., MÜHLBERGER G., PFANZELTER E., THERENTY M., TOIVONEN H. & TOLONEN M. (2020). Newseye : A digital investigator for historical newspapers. In L. ESTILL & J. GUILIANO,

Éds., *15th Annual International Conference of the Alliance of Digital Humanities Organizations, DH 2020, Ottawa, Canada, July 20-25, 2020, Conference Abstracts*.

GIRDHAR N., COUSTATY M. & DOUCET A. (2023). Benchmarking nas for article separation in historical newspapers. In *International Conference on Asian Digital Libraries*, p. 76–88 : Springer.

GONZÁLEZ-GALLARDO C.-E., BOROS E., GIAMPHY E., HAMDI A., MORENO J. G. & DOUCET A. (2023a). Injecting temporal-aware knowledge in historical named entity recognition. In *European Conference on Information Retrieval*, p. 377–393 : Springer.

GONZÁLEZ-GALLARDO C.-E., BOROS E., GIRDHAR N., HAMDI A., MORENO J. G. & DOUCET A. (2023b). Yes but.. can chatgpt identify entities in historical documents ? In *2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, p. 184–189 : IEEE.

GONZÁLEZ-GALLARDO C.-E., HANH T. T. H., HAMDI A. & DOUCET A. (2024). Leveraging Open Large Language Models for Historical Named Entity Recognition. In *The 28th International Conference on Theory and Practice of Digital Libraries*, Ljubljana, Slovenia. HAL : hal-04662000.

LEWIS P., PEREZ E., PIKTUS A., PETRONI F., KARPUKHIN V., GOYAL N., KÜTTLER H., LEWIS M., YIH W.-T., ROCKTÄSCHEL T. *et al.* (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, **33**, 9459–9474.

TRAN T. T., GONZÁLEZ-GALLARDO C.-E. & DOUCET A. (2025). Retrieval augmented generation for historical newspapers. In *Proceedings of the 24th ACM/IEEE Joint Conference on Digital Libraries*, JCDL '24, Hong Kong, China : Association for Computing Machinery. DOI : 10.1145/3677389.3702542.

ZHANG X., THAKUR N., OGUNDEPO O., KAMALLOO E., ALFONSO-HERMELO D., LI X., LIU Q., REZAGHOLIZADEH M. & LIN J. (2023). Miracl : A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics*, **11**, 1114–1131.