Retrieval Augmented Generation for Historical Newspapers

The Trung Tran, Carlos-Emiliano González-Gallardo, Antoine Doucet

Retrieval Augmented Generation for Historical Newspapers

The Trung Tran
trungtt.bi12-452@st.usth.edu.vn
University of Science and Technology of Hanoi, ICTLab
Hanoi, Vietnam

Carlos-Emiliano González-Gallardo*
gonzalezgallardo@univ-tours.fr
University of Tours, LIFAT / CESR
Tours, France

Antoine Doucet
antoine.doucet@univ-lr.fr
University of La Rochelle, L3i
La Rochelle, France

Abstract

Nowadays, the accessibility and long-term preservation of historical records are significantly impacted by the sharp increase in the digitization of these archives. This shift creates new opportunities for researchers and students in multiple disciplines to broaden their knowledge or conduct multidisciplinary research. However, given the vast amount of data that needs to be analyzed, using this knowledge is not easy. Different natural language processing tasks such as named entity recognition, entity linking, and article separation have been developed to make this accessibility easier for the public by extracting information and structuring data. However, historical newspaper article aggregation is still unexplored. In this work, we demonstrate the potential of the retrieval-augmented generation framework that integrates large language models (LLMs), a semantic retrieval module, and knowledge bases to create a system capable of aggregating historical newspaper articles. In addition, we propose a set of metrics that permit evaluating these generative systems without requiring any ground truth. The results of our proposed RAG pipeline are promising at this early stage of the system. They show that semantic retrieval with the help of reranking and additional information (NER) reduces the impact of OCR errors and query misspellings.

CCS Concepts

• **Information systems** → **Digital libraries and archives**; **Retrieval models and ranking**; • **Applied computing** → **Arts and humanities**; **Digital libraries and archives**; • **Computing methodologies** → **Natural language generation**; **Natural language processing**.

Keywords

Digital Humanities, Retrieval-Augmented Generation, Large Language Models, Historical Newspapers

ACM Reference Format:

The Trung Tran, Carlos-Emiliano González-Gallardo, and Antoine Doucet. 2018. Retrieval Augmented Generation for Historical Newspapers. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

*This work was done while at the University of La Rochelle.

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

With the ever-increasing pace of technology, the digitization of historical archives has experienced a drastic upward trend that profoundly affects the accessibility and long-term preservation of these documents. Billions of images have been recorded in manuscripts, newspapers, and old journals during this revolution in digitization. With the support of optical character recognition (OCR) methods, handwritten text recognition (HTR) algorithms, or simply by manual annotation, people have been able to extract these data into large corpora [3, 16]. By converting physical materials into their digital form, these corpora not only act as a memory bank that safeguards all fragile documents but also provide accessibility to all users worldwide, regardless of physical or geographical constraints. This massive change also opens up new opportunities for researchers and students in multiple disciplines, including history, literature, linguistics, and archaeology. These disciplines can leverage the stored archives to enhance their knowledge, gain new insights, or even conduct multidisciplinary collaboration to acquire multidimensional understanding, strengthening research activities in the digital humanities field and beyond.

However, given the vast amount of data they have to examine, it is nearly impossible for researchers to gain deeper insights into these documents composed of plain texts and images. To alleviate or further resolve the problem, multiple projects have recently focused on using advanced techniques in natural language processing (NLP), which allows researchers to easily benefit from the extracted and structured data. Several projects have recently been led on the topic of newspaper analysis, such as NewsEye[1] [4] and Impresso[2]. These projects aim to provide tools for analyzing ancient and multilingual articles for researchers in digital humanities and the general public.

Given the multiple challenges that historical data pose, such tools cannot be easily developed. OCR errors, the unstructured nature of the data, and the differences in languages or names between different periods are some of the significant challenges that limit their performance. Some works develop methods to extract and index text data from a large corpus into a knowledge graph for later retrieval [1]. Others have also tried to solve the problem of time-changing content by injecting temporal information into the graph [7]. With the rise of large language models (LLMs) such as GPT, Llama, and Mistral, researchers have drastically improved several NLP tasks and achieved state-of-the-art results. Regarding historical documents, researchers have started to integrate and test closed

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

and open LLMs into the named entity recognition information extraction task [8, 9]. Nevertheless, both lines of work come with certain limitations. One of the principal drawbacks of knowledge bases is their limitation in automatically generating a full answer for the user. LLMs, despite their human-like and plausible-sounding answer, can easily produce hallucinations, thus providing false information. Hence, in this work, we make use of retrieval-augmented generation (RAG) [11], an emerging and promising method that enables the combination of both LLMs and knowledge bases to enhance the result of language model generation by combining it with external knowledge. Using a multilingual approach, we implemented a complete RAG system for the historical newspaper articles of the NewsEye data [6]. We also implement finetuning and additional vector representation injection to handle some of the existing problems from the data related to their historical nature. Lastly, we evaluate the system in both quantitative and qualitative approaches. For the former, we evaluate the retrieval phase using standard information retrieval metrics. For the latter, we create sample queries and discuss the system's results[3].

2 Related Work

RAG has shown promising advancements in information retrieval methods, attracting many researchers and leading to rapid advancement. The most basic form of RAG follows a pipeline by first indexing the documents into a database, proceeding with a retrieval process on those embedded documents, and finally creating a prompt based on the original query and the retrieved information to generate an answer. This establishes a baseline for more advanced RAG architectures, mitigating some drawbacks, such as low accuracy in the retrieval process or hallucinations in the generation module. With that in mind, much work has been done to maximize the potential of RAG systems. Some involve optimizing the indexing strategy and the original query, usually called the pre-retrieval step [10, 12]. Afterward, a post-retrieval step can be developed to better select the retrieved documents by choosing the best quality set of information while removing some of the redundancy existing in the first retrieval step.

A complete RAG pipeline does not need to be constrained using only these modules, and it has shown flexibility depending on the task it needs to work on. Fan et al. [5] proposed a RAG system by training a model that can be used for hashtag ranking, offering the ability to work on the recommendation of the mainstream hashtags. Soman et al. [15] developed a particular RAG system for knowledge graphs that can work in knowledge-intensive domains such as the biomedical one. However, to our knowledge, no work has been done using RAG for historical newspaper articles.

3 Methodology

Our RAG system pipeline can be seen in Figure 1. First, we create a database using the small multilingual model E5[4] to embed all documents into vector representations. We also add metadata to the text during this step, such as the article's title. We then index the E5 embeddings in the open-source vector database Chroma[5]. The document's similarity calculation is calculated using cosine similarity and the retrieval process will employ maximal marginal relevance (MMR) [2]. We use the same procedure on the title or summary of each article to create a second database which is in the role of a semantic router in our pipeline. After the creation of the necessary databases, the main system make use of these and comprises of four main components: a query router where we redirect the behavior of the system based on the user's question, a base retrieval model where some of the best documents related to the question are retrieved and they will be ranked using a rerank module to filter out some irrelevant ones. And lastly, all of these retrieved documents are aggregated and forward to the prompt of an LLM to generate the final answer given a user's query. The base system uses a query routing procedure to adapt the system to whether or not to go directly to a web search if none of the retrieved documents is relevant to the query. We provide an article-level retrieval mode based on the article database created earlier that can assess whether the system can continue the normal retrieving path. In other words, we first use a retrieval module to retrieve text at the title or summary level, which helps to determine the path to take. The system goes directly to the web search module if we cannot retrieve an article with a similarity score that exceeds a certain threshold. Otherwise, it continues on the traditional path. Then, we retrieve the documents based on the article title we have retrieved before employing reranking to select some of the best candidates to be further forwarded to the prompt. In this step, we set a threshold to remove irrelevant information. The reranker consists of two different paths presenting two ways of calculating the similarity score for a query-document pair; one uses Cohere Multilingual Reranker[6] which outputs a Cohere score; the other works on the named entities and generates a NER score, which is not affected by the OCR errors, thus injecting more stable information into the module. First, we extract the named entities with hmBERT [7] [14] and inject this information into our module. The task involves first creating a string that encompasses the named entities:

$$NE\_str = \texttt{concat}(ne_1, ne_2, \ldots, \text{“}O\text{”})$$

It can be seen that the string is simply the concatenation of multiple detected entities. The last "O", used for non-entity tokens, is added to avoid the problem of an empty string. We pass this string as metadata for the text embedded using the TF-IDF module during the reranking process to receive another vector representation apart from the dense embedding we already have. The NER score is then calculated on the basis of these vector representations using cosine similarity. The final score is the weighted sum of the NER and the Cohere scores. In our experiments, we use LLaMA3 in the answer generation module.

4 Experiments and Evaluation

In this section, we present the experiments performed and evaluate the system's performance in two ways: quantitative evaluation using benchmark data and qualitative evaluation using historical data as the primary target.

[3]The code is available in https://anonymous.4open.science/r/RAG-project-8C13/
[4]https://huggingface.co/intfloat/multilingual-e5-small
[5]https://docs.trychroma.com/
[6]https://cohere.com/blog/rerank-3
[7]https://huggingface.co/hmbert

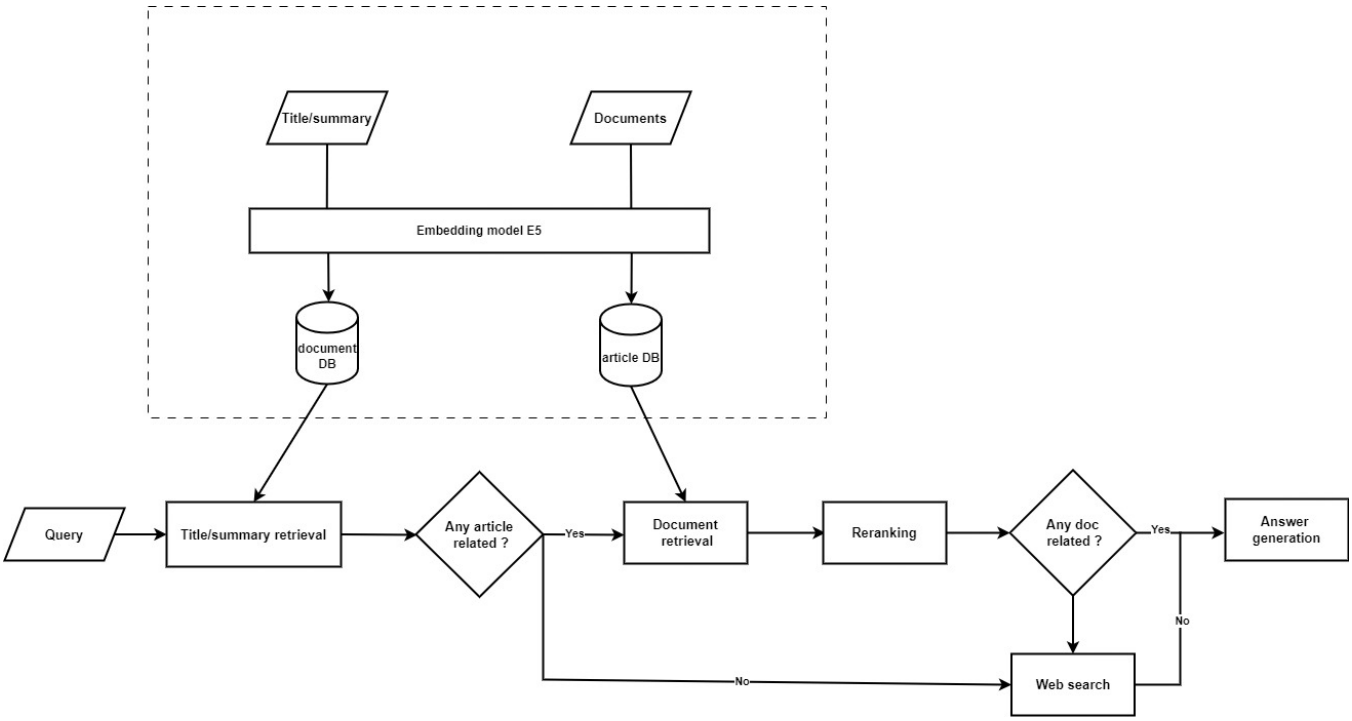

Figure 1: Modules of the RAG pipeline

4.1 Quantitative Evaluation

4.1.1 Setup. For this evaluation, we used the French and Finnish subsets of Miracl [20], a multilingual dataset used for the evaluation of information retrieval. In addition to French and Finnish, we added English to evaluate the answer generation module. Since this dataset is already split into passages, we directly indexed each chunk in the database. In addition, Miracl provides a title as metadata for each of these passages, which we indexed and used for title retrieval. Lastly, we set the temperature of LLaMA3 to 0.3 for the final answer generation module.

We evaluated the system in two tasks: information retrieval and answer generation. For information retrieval, we used Recall@100 and NDCG@10. The generation of answers is evaluated based on the retrieved information. We used four metrics and computed their linear correlation. These metrics compute the similarity or distance between the answer and the passages that were previously retrieved. We used the BERT score [19] (F1 score) on the concatenation of all retrieved passages and the output (the answer), an average cosine similarity score, a LLM score where we directly prompted LLaMA3 to output a number between 1 and 10 based on the quality of the answer, and a Quality Index score[13], which we applied to TF-IDF vectors using the formula:

$$Q = \sum_{i=1}^{N} \frac{\max_{p \in P} p_i}{\sum_{j=1}^{N} \max_{p' \in P} p'_j} \sum_{p \in P} \frac{p_i}{\sum_{p' \in P} p'_i} |r_i - p_i|,$$

where N is the number of features in the vector representation, P is all the retrieved paragraphs, r is the answer and p_i is the value of the vector representation of p at index i.

Table 1: Information retrieval results on the Miracl dataset

		BM25 [20]	E5	E5 + Cohere	E5 + Cohere + NER
French	Recall@100	0.653	0.814	0.870	0.869
	NDCG@10	0.183	0.425	0.528	0.482
Finnish	Recall@100	0.891	0.923	0.930	0.936
	NDCG@10	0.551	0.688	0.750	0.714

4.1.2 Results. Table 1 shows all the results from the different methods: sparse retrieval (BM25), dense retrieval with cosine similarity (E5), the combination of dense retrieval with a cross-encoder model (E5+Cohere), and the reranker after injecting information (E5+Cohere+NER). As expected, dense retrieval produces a superior result compared to sparse retrieval thanks to its ability to capture the semantic and context meaning of a sentence. On the other hand, BM25 is highly dependent on vocabulary matching. In addition, we can see that with the normal reranking method (E5+Cohere), the system can increase its performance by nearly 0.1 points for NDCG. However, injecting NER seems to worsen the final result. This could be due in large part to the fact that using NER could cause misleading information in certain situations where named entities do not directly contribute to the key features of a sentence. Another factor might come from the errors of the NER extraction model itself.

Regarding the results of the answer generation module with E5 retrieval shown in Table 2, in the three languages, Finnish shows the lowest results among the three languages. This might show the inability of LLaMA3 to handle some languages with fewer resources. Meanwhile, the difference created by the BERTscore

Table 2: Answer Generation results on the Miracl dataset

	LLM score	BERTscore	Cosine Similarity	Quality Index
English	5.95	0.84	0.58	0.18
French	5.14	0.66	0.59	0.17
Finnish	2.9	0.58	0.37	0.27

between English and French is significantly higher than that of other metrics where the gaps are unnoticeable. We computed the linear correlations between these scores to analyze how well each metric relates to each other across languages.

Table 3: Metrics correlation on Miracl's English data

	LLM score	BERTscore	Cosine Similarity	Quality Index
LLM score	1	0.36	0.30	0.02
BERTscore	–	1	0.81	-0.41
Cosine Similarity	–	–	1	-0.56
Quality Index	–	–	–	1

Table 4: Metrics correlation on Miracl's French data

	LLM score	BERTscore	Cosine Similarity	Quality Index
LLM score	1	0.8	0.78	-0.14
BERTscore	–	1	0.86	-0.15
Cosine Similarity	–	–	1	-0.13
Quality Index	–	–	–	1

Table 5: Metrics correlation on Miracl's Finnish data

	LLM score	BERTscore	Cosine Similarity	Quality Index
LLM score	1	0.56	0.51	-0.01
BERTscore	–	1	0.71	0.00
Cosine Similarity	–	–	1	0.02
Quality Index	–	–	–	1

All three languages share a similar result, where the BERTscore and cosine similarity show the highest correlation with each other, greater than 0.7 between each pair. The reason behind this might involve the fact that both scores work on dense embedding vectors, making them similar. In addition to this, the quality index is slightly more correlated with other scores compared to the LLM score in the English data (Table 3). However, the reverse trend can be observed for French and Finnish, in Tables 4 and 5, respectively. The quality index can be seen to have a very low correlation in these languages. In contrast, the LLM score has a much higher correlation, with roughly 0.8 in relation to the BERTscore and cosine similarity in the French data.

4.2 Qualitative Evaluation

Sinceno historical newspaper dataset with ground truth for RAG is available to allow quantitative evaluation, we performed a qualitative evaluation to analyze how the system performs using query samples.

4.2.1 Setup. We used the NAS NewsEye dataset [6], a multilingual dataset of historical newspapers. It comprises articles pages from the 19th to early 20th century in three languages: French, provided by the National Library of France, Finnish, by the National Library of Finland, and German, provided by the Austrian National Library. For each language, 150 pages are supplied, containing 4,836 distinct articles. The dataset is separated by paragraphs, which mitigates the work of chunking. However, many of these paragraphs contain only a few words, which could be the case that a title is incorrectly annotated as a paragraph. To handle this problem, during the extraction process, we concatenated all of these short passages into one if they belong to the same article. We also removed all unwanted characters or symbols. For title/summary retrieval, we observed that most of the models could not adapt to the noise of the documents, such as OCR errors. This, in turn, worsens the final system performance, and thus, the module will not be used for these data. For the retrieval module, we used the large version of E5[8] which can convert text in a 1024-embedding space. Finally, we finetune the retrieval model using the combination of two approaches, GPL [18] and TSDAE [17] to improve the final performance.

Query: Qui est Carlos Sandov ?
Retrieved: Cependant survient dans le village don Caros Sandoval, un brillant capitaine de dragons,e fils de l'alcade, cherchant aventure amoureuse,qui roucoule une sérénade sous le balcon de labrune Inès, et se prend de querelle avec, Salvador,qui le provoque et le tue. Or, la loi est formelle,emeurtrier, fût-ce en duel, est puni de mort, etl'alcade, qui veut venger son fils, fait arrêter Salvador dans la chambre même d'Inès, où il s'étaitéfugié.
Cohere score: 0.64
NER score: 1
Final score: 0.71

Figure 2: Reranking example 1

Query: Qui est Raffael?
Retrieved: M. Raffaelli est le peintre des pauvres diables.Il a rêvé d'écrire à sa façon sa Chanson desueux, et celle-là vaudra l'autre. Ses petitsbourgeois et ses chiffonniers disent bien tousl'inutilité, le vide et la misère de leur existencenorne ou terrible, misérable ou sans joie. Chacune ses tableaux est un chapitre de leur vie admirablement conté.
Cohere score: 0.88
NER score: 1
Final score: 0.91

Figure 3: Reranking example 2

4.2.2 Results. In Figure 2, we show an example in which we deliberately damage the query by modifying the name of the exposition from *Caros Sandoval* to *Carlos Sandov*. Similarly with Figure 3 and Figure 4, the named entities are being added noises.This process is to experiment with whether the system can produce robust results

[8]https://huggingface.co/intfloat/multilingual-e5-large

Query: Qu'est-ce qu'il c'est passé à Bulwaro le 14 janvier 1900 ?

Retrieved: TRISTE CONSTATATION
Buluwaro, 14 janvier. — Le capitaineLlevelyn a fait une reconnaissance avecle train blindé, à quatre milles au sudde Gaberones. Il a trouvé le pont, quiavait été si laborieusement reconstruitpar les Anglais, à l'état de ruines fumantes.

Cohere score: 0.97
NER score: 0.93
Final score: 0.96

Figure 4: Reranking example 3

despite OCR errors or user misspellings when querying. It can be seen that, though the Cohere model is obtaining a good answer, its score is relatively low for Figure 2. However, with the injection of NER into the model, the score is much higher, indicating that more precise information can boost the system's final overall performance. In other cases, Cohere model could still show some robust performances despite the noises we have added and provide high scores without the help of NER information.

Query: Qui est Raffaelli?
Answer: M. Raffaelli is a painter who depicts the lives of poor people and outcasts in his work. He is known for his sincere and original talent, which brings truth to life through his art. He is both a figurative artist of the first order and an excellent landscape painter.

Query: Qu'est-ce qu'il c'est passé à Buluwaro le 14 janvier 1900 ?
Answer: According to the context, it appears that something significant happened at Buluwaro on January 14, 1900. The text mentions a reconnaissance mission by Captain Llevelyn with a train and finds a bridge in ruins, suggesting some kind of conflict or destruction had occurred.

Query: Qui est Caros Sandoval?
Answer: Based on the provided context, Caros Sandoval is a brilliant dragon captain and the son of the alcade (mayor) who seeks adventure and romance. He is killed in a duel with Salvador, which leads to his father seeking revenge.

Figure 5: Final answer examples

The final result (Answer) of LLaMA3 in Figure 5 can be seen to be reasonable given the data. However, it can be seen that the answer is returned in different languages, whereas the queries are made only in French. This could create inconsistent quality and should be addressed in the future.

5 Conclusion and Future Work

In this article, we presented a fully functional RAG system for historical newspaper articles. To our knowledge, this is the first RAG system focused on this type of document, opening a new path for analyzing digital heritage. This can facilitate the way scholars, professionals, and the general public access this rich source of information. We have also introduced some improvements using metadata to inject more information, thus helping the RAG system to better aggregate the historical newspaper articles. In addition, we have proposed four evaluation metrics for the answer generation module for which no ground truth is needed. The first results and correlation analysis show the system's potential to produce aggregated articles despite the noisy nature of historical documents. However, the proposed metrics are currently highly constrained, as they are best suited to the more specific task of information aggregation. The next steps will be to independently improve each module in the RAG pipeline by injecting more information and finetuning the LLM with historical newspaper data to increase its capacity to deal with OCR errors and query misspellings. Finally, we will develop more comprehensive metrics that better fit the task.

Acknowledgments

This work has been supported by the ANNA (2019-1R40226), TERMITRAD (2020-2019-8510010), Pypa (AAPR2021-2021-12263410), and Actuadata (AAPR2022-2021-17014610) projects funded by the Nouvelle-Aquitaine Region (France).

References

[1] Emanuela Boros, Carlos-Emiliano González-Gallardo, Edward Giamphy, Ahmed Hamdi, José G Moreno, and Antoine Doucet. 2022. Knowledge-based Contexts for Historical Named Entity Recognition & Linking.. In *CLEF (Working Notes)*. 1064–1078.

[2] Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. 335–336.

[3] Tim Causer and Melissa Terras. 2016. 'Many Hands Make Light Work. Many Hands Together Make Merry Work'1: Transcribe Bentham and Crowdsourcing Manuscript Collections. In *Crowdsourcing our cultural heritage*. Routledge, 57–88.

[4] Antoine Doucet, Martin Gasteiner, Mark Granroth-Wilding, Max Kaiser, Minna Kaukonen, Roger Labahn, Jean-Philippe Moreux, Günter Mühlberger, Eva Pfanzelter, Marie-Eve Therenty, Hannu Toivonen, and Mikko Tolonen. 2020. NewsEye: A digital investigator for historical newspapers. In *15th Annual International Conference of the Alliance of Digital Humanities Organizations, DH 2020, Ottawa, Canada, July 20-25, 2020, Conference Abstracts*, Laura Estill and Jennifer Guiliano (Eds.). https://dh2020.adho.org/wp-content/uploads/2020/07/721_NewsEyeAdigitalinvestigatorforhistoricalnewspapers.html

[5] Run-Ze Fan, Yixing Fan, Jiangui Chen, Jiafeng Guo, Ruqing Zhang, and Xueqi Cheng. 2024. RIGHT: Retrieval-Augmented Generation for Mainstream Hashtag Recommendation. In *European Conference on Information Retrieval*. Springer, 39–55.

[6] Nancy Girdhar, Mickaël Coustaty, and Antoine Doucet. 2023. Benchmarking nas for article separation in historical newspapers. In *International Conference on Asian Digital Libraries*. Springer, 76–88.

[7] Carlos-Emiliano González-Gallardo, Emanuela Boros, Edward Giamphy, Ahmed Hamdi, José G Moreno, and Antoine Doucet. 2023. Injecting temporal-aware knowledge in historical named entity recognition. In *European Conference on Information Retrieval*. Springer, 377–393.

[8] Carlos-Emiliano González-Gallardo, Emanuela Boros, Nancy Girdhar, Ahmed Hamdi, Jose G Moreno, and Antoine Doucet. 2023. Yes but.. can chatgpt identify entities in historical documents?. In *2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE, 184–189.

[9] Carlos-Emiliano González-Gallardo, Tran Thi Hong Hanh, Ahmed Hamdi, and Antoine Doucet. 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. https://univ-rochelle.hal.science/hal-04662000

[10] Rolf Jagerman, Honglei Zhuang, Zhen Qin, Xuanhui Wang, and Michael Bendersky. 2023. Query expansion by prompting large language models. *arXiv preprint arXiv:2305.03653* (2023).

[11] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.

[12] Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting for retrieval-augmented large language models. *arXiv preprint arXiv:2305.14283* (2023).

[13] Gemma Piella and Henk Heijmans. 2003. A new quality metric for image fusion. In *Proceedings 2003 international conference on image processing (Cat. No. 03CH37429)*, Vol. 3. IEEE, III–173.

[14] Stefan Schweter, Luisa März, Katharina Schmid, and Erion Çano. 2022. hmbert: Historical multilingual language models for named entity recognition. *arXiv preprint arXiv:2205.15575* (2022).

[15] Karthik Soman, Peter W Rose, John H Morris, Rabia E Akbas, Brett Smith, Braian Peetoom, Catalina Villouta-Reyes, Gabriel Cerono, Yongmei Shi, Angela Rizk-Jackson, Sharat Israni, Charlotte A Nelson, Sui Huang, and Sergio E Baranzini. 2024. Biomedical knowledge graph-optimized prompt generation for large language models. arXiv:2311.17330 [cs.CL] https://arxiv.org/abs/2311.17330

[16] Melissa M. Terras. 2011. *The Rise of Digitization.* SensePublishers, Rotterdam, 3–20. https://doi.org/10.1007/978-94-6091-299-3_1

[17] Kexin Wang, Nils Reimers, and Iryna Gurevych. 2021. Tsdae: Using transformer-based sequential denoising auto-encoder for unsupervised sentence embedding learning. *arXiv preprint arXiv:2104.06979* (2021).

[18] Kexin Wang, Nandan Thakur, Nils Reimers, and Iryna Gurevych. 2021. GPL: Generative pseudo labeling for unsupervised domain adaptation of dense retrieval. *arXiv preprint arXiv:2112.07577* (2021).

[19] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* (2019).

[20] Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. Miracl: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics* 11 (2023), 1114–1131.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009