Querying the Past:
Automatic Source Attribution with Language Models

Ryan Muther[1], Mathew Barber[2] and David Smith[1]

[1]*Northeastern University, 360 Huntington Ave, Boston, MA 02115, USA*

[2]*Aga Khan University Institute for the Study of Muslim Civilisations, 10 Handyside St, London, GVQG 23, UK*

Abstract
This paper explores new methods for locating the sources used to write a text by fine-tuning a variety of language models to rerank candidate sources. These methods promise to shed new light on traditions with complex citational practices, such as in medieval Arabic where citations are ambiguous and boundaries of quotation are poorly defined. After retrieving candidates sources using a baseline BM25 retrieval model, a variety of reranking methods are tested to see how effective they are at the task of source attribution. We conduct experiments on two datasets—English Wikipedia and medieval Arabic historical writing—and employ a variety of retrieval- and generation-based reranking models. In particular, we seek to understand how the degree of supervision required affects the performance of various reranking models. We find that semi-supervised methods can be nearly as effective as fully supervised methods while avoiding potentially costly span-level annotation of the target and source documents.

Keywords
information retrieval, citation modeling, source attribution

1. Introduction

When reading a text, it is often useful to know which sources were used to write it. Knowledge of the sources used to write a news article, for example, can inform a reader of bias in how information in the article is reported. In historical domains, the sources used to write a document can both provide insight into how the author worked and what materials they had access to. We define the problem of determining the sources used to write a piece of text as that of source attribution.

Researchers in natural language processing most often study source attribution in scientific papers, inferring links to referenced articles based on citations. Part of why this can be done so well is that modern citations follow a standardized format—often generated by required typesetting packages—that can be parsed by regular expressions or other simple methods. This comparative ease of data creation in turn allows the creation of large data sets for training fully supervised models for source attribution using the bibliographic information recovered from the citations. These models tend to work best when there is a 1:1 correspondence between first printings of a work and papers. In more ambiguous domains, where potential sources

can be redundant, these bibliographic models can often fail to disambiguate which of multiple possible editions of a work is the correct source.

Citations in medieval Arabic historiography are a particularly complex case. From the seventh until around the twelfth century, the predominant form of citation was the *isnad*, a chain of authorities linking the author of a text back to the original (often presumed oral) source for a narrative [15]. *Isnad*s gradually gave way to the citation of individual authors, occasionally with reference to the title of an author's text. All of these forms of citation are highly ambiguous. While they can often be precisely traced to source texts, the process of manually doing so is laborious and subject to error.

For example, the historian al-Maqrizi (d. 1442CE/845AH) wrote multiple large texts, which are full of citations both to lost and extant works. As Bauden has shown, these citations reveal how al-Maqrizi worked, the sources to which he had access, and his attitudes towards source attribution [3, 4, 2]. Al-Maqrizi is not exceptional: for examples of other large, richly-cited, texts see [10, 13, 5]. If we are to understand how the Arabic historiographical tradition approaches quotation and citation, we need methods that can disambiguate vast numbers of citations across a range of texts and citation styles.

In order to retrieve sources in settings where the citations are harder to locate and more ambiguous, creating fully annotated data in large amounts can be time consuming and require significant domain expertise. To circumvent this, we experiment with different levels of supervision in the models we use to retrieve potential sources. As we will see, semi-supervised methods can perform comparably to more annotation-intensive fully supervised models.

Looking only at the text itself, there are two principle forms of information about the sources: **text reuse** and **citation**. Text reuse is when an author directly copies material from their source, possibly involving more complex transformations. This is common practice in highly intertexual domains like historical Arabic writing. Citation involves the author explicitly telling the reader which source is being used, as one often sees in modern scientific writing or Wikipedia entries. Citations can have varying degrees of specificity, ranging from simply the author(s) and year, as one sees in some fields of scientific literature, to a more full-fledged citation including a title and page number, as is more common in many fields in the humanities. In some cases, the citation may take the form of a unique identifier, such as a URL or Wikipedia headword.

Each of these forms of citation and reuse can be viewed as part of a broader spectrum of the form of relationship between a text and its sources. At one extreme, we have Wikipedia, where the simplest form of citations to other Wikipedia articles is a link to the cited article by only headword. At the opposite extreme, there is the highly intertextual classical Arabic domain, where source attribution is more easily performed by recognizing the source text rather than any attribution on the part of the author. We view both text reuse and citations as species of *queries*, that allow a reader to combine information in the text, domain knowledge, and bibliographic knowledge to track down the intended source.

Different architectures can be used to model the process of source attribution from the perspective of the author or that of the reader. From the perspective of the author, one could imagine a process in which they select a source (retrieval) and then use the text of that source as the basis for their own writing (generation). This process is similar to the process used by recently-proposed retrieval-augmented generation models in the work by Lewis et al. [8] From the perspective of the reader, the source retrieval problem is more like one of retrieval alone,

as the reader doesn't need to create the target themselves, but can use it in the construction of a query to find sources.

We operationalize the problem by turning it into a two-stage retrieval and reranking problem. We first use a baseline retrieval model to retrieve candidate sources for a given target document. A second model is then used to rerank the possible sources.

This paper is organized as follows: §2 covers related work; §3 provides an overview of the datasets we experiment with; §4 covers the forms of models we use; §5 describes our experiments and results; and §6 provides a discussion of the results as well as potential avenues for future work.

2. Related Work

This paper is closely related to work by others on combined retrieval and generation methods for question answering, citation suggestion, and literary evidence retrieval, similar to that done by Lewis, Perez, Piktus, Petroni, Karpukhin, Goyal, Küttler, Lewis, Yih, Rocktäschel, Riedel, and Kiela [8]. In contrast their work, however, we are more focused on improving the model's retrieval performance rather than the generation performance. Also of interest here is the work of Mao, He, Liu, Shen, Gao, Han, and Chen [11] on generation-augmented retrieval. They focus more on generating better queries by applying generative models than on using the generative model as a reranker.

The problem of source attribution is similar to that of citation recommendation, which is usually thought of as a tool for writers to find relevant citations. Zhang and Zhu [20] evaluate various forms of citation prediction model based on the similarity between the citation context and the citing paper to predict citations in PLOS ONE. While this work is valuable for helping authors of scientific publications, it is limited in scope due to the way scientific papers tend to engage with their sources at the coarse paper level. In more humanistic domains, authors often engage with the text of their sources directly and may reference multiple parts of a source, making our source attribution problem a more granular version of the commonly studied citation recommendation problem.

The most similar problem studied elsewhere is literary evidence retrieval as proposed by Thai, Chang, Krishna, and Iyyer [19]. In RELIC, the goal is to retrieve the correct quoted passage from a known text based on its context in a work of literary analysis. The objective here is similar, but the relationship between the citing and cited texts may be more complex than direct quotation and the source text is not necessarily known, complicating source retrieval.

3. Datasets and Tasks

We work with two datasets in this paper: Wikipedia links to other Wikipedia pages (Wiki-Link) and two classical Arabic texts taken from the OpenITI corpus of digitized Arabic texts [16]. These datasets datasets in particular were chosen as the represent different points in the spectrum of relationships between texts and sources. For the Wikipedia link task, the use of the source requires very little modification and could be reduced to copying the headword of the source article. With Maqrizi, the relationship between the text and its sources are more

Table 1
Sizes (in source documents) of datasets used in our experiments

Dataset	Train	Test
Wiki-Link	116,038	12,880
Maqrizi	162	19

complicated, with the source text often uncited or cited in ways that are difficult to recognize automatically, lacking any sort of standardized form like that found in modern genres. Additionally, the source may often be heavily edited or paraphrased, further complicating the source-target relationship.

Table 1 shows the sizes of training and test sets. The texts used in the classical Arabic experiments are al-Maqrizi (d. 1442CE/845AH)'s *al-Mawa'iz wa-l-Itti'bar bi-Dhikr al-Khitat wa-l-Athar* - a topographical history of Egypt, often referred to as the *Khitat* - and one of Maqrizi's sources, Ibn ʿAbd al-Hakam (d. 871CE/257AH)'s *Futuh Misr wa-l-Maghrib* - a history of the Muslim conquest of Egypt. Based on passim analysis of the OpenITI Corpus (passim release 2022.2.7), the *Futuh Misr* is the second-most reused text by al-Maqrizi in his *Khitat*, with around 20,000 word tokens shared between the two books.

These reused passages of the *Futuh Misr* are often cited in conjunction with other sources. The following excerpt is from the *Khitat*, found in a section entitled, 'The canals that intersect the Nile' (citations have been underlined):

> It is known that once the Nile has finished rising, canals and channels are cut from it... [summary of the canals and their names]
>
> Ibn ʿAbd al-Hakam said, [quoting] from Abu Riham al-Sama'i: Old Cairo [Misr] had stone and other kinds of bridges, [built] by decree and design, such that water passed under its houses and [through] its courtyards...[a description of some of the canals dug in Egypt]. The Sakha canal was dug by Tudarus b. Sa b. Qubtim b. Misrayim b. Baysar b. Ham b. Nuh. He was the first of the ancient Coptic kings, who ruled Egypt in the first age.
>
> Ibn Wasif Shah said: King Tudarus was the first King to rule the full extent of it [Egypt] after his father Sa... [there follows a short biography of Tudarus]
>
> The Sardus Canal: Haman dug it. Ibn Wasif Shah said: King Talma b. Qumas sat on the on the throne... [Brief biography of Talma]
>
> Others mention: ... When he [Talma] became King he spent money freely, brought close those who were loyal to him, and killed those who opposed him. His rule was moderate. He appointed Haman as a successor...[more description of Talma's rule, including the canals that he commissioned]
>
> Ibn ʿAbd al-Hakam said, [quoting] from ʿAbdullah b. ʿAmru b. al-ʿAs (God be pleased with them both): The Pharoah tasked Haman with digging the Sardus canal... [12, pp. 186-188]

This quotation is characteristic of the kind of writing found in the *Khitat*, where multiple sources are threaded together to describe topographical features or landmarks and outline their

history, including significant tangents. With no quotation markers, it can be difficult to separate quoted sources from the author's commentary. Al-Maqrizi introduces the section, giving an overview of the canals. He then quotes Ibn ʿAbd al-Hakam for a general historical introduction, at the conclusion of which the King Tudarus is mentioned, connected to the Sakha canal. This leads him to quote Tudarus' biography from another author, Ibn Wasif Shah. Al-Maqrizi then moves on to the next canal, the Sardus Canal, which he states was dug by Haman. Having introduced in his own words, al-Maqrizi re-cites Ibn Wasif Shah - using him once again for a biography of a pre-Islamic ruler. At the close of that biography, there is an ambiguous citation - 'Others mention' - for a description of Talma's connection to Haman. Al-Maqrizi then cites Ibn ʿAbd al-Hakam to describe the digging of the Sardus canal.

As should be clear from this example, citation and quotation in the *Khitat* can be quite ambiguous, especially as the author moves between sources. For example, 'Others mention' could be a citation used by Ibn Wasif Shah, or it could be al-Maqrizi's own citation. Retrieval of potential source texts would allow us to resolve ambiguous references such as this.

The citations themselves are references to authors - 'so-and-so said'. References to book titles occur, but much less frequently, for example:

> al-Asʿad b. Mammati said in his book *Qawanin al-Dawawin*: the Alexandria canal has a number of channels... [12, p. 189]

The use of author names, rather than book titles, increases the level of ambiguity. Authors might be referred to by different names, and - more crucially - authors often wrote more than one book. We need, therefore, to be able to resolve a citation and quotation to the original source text. Text reuse detection can only partially solve this problem, as the same passage might be quoted by multiple authors (and might be used by the same author in multiple works). Moreover, it is necessary to separate the target author's source text from the texts that are being quoted by the source text.

Addressing source attribution in the *Khitat* promises to reveal more about al-Maqrizi's sources and his use of them. Al-Maqrizi produced a large oeuvre, including 9 works (in the OpenITI Corpus) that exceed 100,000 tokens in length, much of which he copied from early source texts (both extant and lost). Of these works, the *Khitat* contains the largest number of citations; in other works, he more frequently quotes from sources without citation. As al-Maqrizi shares so many of his sources between his works, identifying citations within the *Khitat* and their corresponding sources promises to unlock the identity of sources quoted in his other works. For a small-scale case study, see Barber's examination of al-Maqrizi's quotations from the lost Fatimid biography, the *Sirat al-Yazuri* [1]. Given the size of al-Maqrizi's works and the breadth of his source usage, computational methods are essential if we are to move beyond small case studies. This examination of the *Khitat* and its use of the *Futuh Misr* is, therefore, an essential preliminary step in understand these kinds of complex citation and source attribution questions.

The texts have been annotated by a domain expert on al-Maqrizi to create a dataset of 181 regions of shared text between the two works, some with and some without direct attribution in the form of citations by al-Maqrizi. When the other text is directly cited, the citation is separately marked in the annotations. The annotations are created based on the output of the text reuse detection algorithm Passim [18], which operates by aligning sections of texts with a

high number of shared character n-grams to find regions of shared material, which the domain expert refined to create the dataset of source-target pairs that we use in our experiments. Since the works in the corpus are so long, rather than aligning full texts, we cut the works up into 300-token chunks and align those. The alignments created by our annotator are at the chunk level, where chunk X of *Futuh Misr* is a source for chunk Y of *Khitat*. The goal of the experiments with this dataset is to retrieve the proper source chunk for a given target chunk.

To create the dataset of Wikipedia citations, we collected 150,000 links from Wikipedia articles to other Wiki articles where the link's anchor text was the name of the cited page from Singh, West, and Colavizza [17]. To better handle long articlesretrieval and reranking is done at the section level, and any retrieved section from the correct source page is counted as relevant for evaluation purposes. For this dataset, the goal is to be able to retrieve a section from the cited page using the sentence with the link from the citing page

While one could accomplish the source attribution task in this simplified Wikipedia setting by using a simple lookup table of headwords from other Wiki pages, this task should not be discounted as uninteresting as a form of citation. The goal with the experiments on this data is to demonstrate that, as a form of citation, the kinds of models used in contexts where the relationship between the source and the target is more complicated than simple copying are also usable for this simplified domain.

4. Models

For our experiments we compare several kinds of models: a baseline retrieval model, as well as several forms of reranking models applied to the results of the baseline retrieval model; embedding similarity and a generative target text model. We use each form of model to rerank candidate sources retrieved by a baseline BM25 retrieval model. Each of these models is meant to test the usefulness of different architectures in solving the source attribution task. The embedding similarity model is a baseline for how well untuned embedding models can solve this task. The generative models are used to examine how effectively generative models can learn to copy material from the source to the target as well as how the text is transformed in moving from the source to the target.

4.1. Baseline Retrieval Model

The starting point of all of our experiments is a BM25 retrieval model used to retrieve possible sources, for which we use pyserini's implementation.[9] To allow the retrieval model to leverage information present in citations, we augment the source documents with bibliographic data, which is often otherwise not present in the source documents. In the case of the Wikipedia link, this takes the form of the article title. For the al-Maqrizi dataset, we augment sources with the author's name and the title of the source text, some combination of which is frequently be used by al-Maqrizi to indicate his sources when citations are present.

4.2. Embedding Similarity

The simplest form reranks documents by descending order of cosine similarity of the representations of the source and target under a BERT embedding model trained on English and Arabic by Lan et al. [6] As one might infer, the similarity is calculated as in Equation 1 where t_{BERT} and s_{BERT} are the BERT embeddings of the target and source respectively.

$$sim = CosineSimilarity(t_{BERT}, s_{BERT}) \tag{1}$$

This places source documents with more similar embeddings to the target document higher in the ranked list. The intuition behind this approach is that the sources used to write a text will be topically similar, meaning that the source documents may be nearby in embedding space. In practice, however, this intuition may not lead to better source retrieval performance in all domains, as experiments in Section 5 will show.

4.3. Generator-Only

The generator-only models rerank sources using the likelihood from a BART-based generation model [7] of some portion of the target (citing) document conditioned on the source (cited) document and the unmasked sections of the target. Unlike the embedding similarity method described in the previous section that only requires supervision at the document level, this method also requires the span of interest in the target to be annotated. The span of interest is the section of the target that we are interested in attributing to a particular source. At training time, this section of the target is replaced with <MASK> tokens and the BART model is trained to predict the masked span conditioned on concatenation of the masked target and source by minimizing the log likelihood of the masked span as in Equation 2.

$$L = -\log p(t_{mask}|t_{obs}, s) \tag{2}$$

where t_{mask} is the masked span in the target, t_{obs} is the observed portion of the target, and s is the source document. At inference time, the retrieved sources are reranked using the same loss, moving source documents that are more useful for generating the target text up in the ranked list of sources. For the al-Maqrizi dataset, since some annotated spans are often quite long, with the longest being 300 words (an entire input document), we truncate the masked sections to be at most 100 word pieces in length, leaving some of the target document to condition the generator.

Additionally, to test the feasibility of less annotation-heavy semi-supervised models, we also experiment with a semi-supervised version of BART where rather than using the known-correct source s we substitute the top-ranking retrieved source from the baseline retrieval model s', giving us

$$L = -\log p(t_{mask}|t_{obs}, s') \tag{3}$$

as a loss function. In theory, a sufficiently strong baseline retrieval model will give BART enough correct documents to learn from, while learning to ignore erroneously retrieved irrelevant sources. This form of model, of course, still requires annotation at the span level in the

Table 2

Results for Reranking Experiments with various models and datasets

Dataset	Model	Supervision	R@10	MRR
Wiki	Baseline	N/A	.64	.478
Wiki	+BERT	N/A	.14	.068
Wiki	+BART	Pair- and Span-level	.97	.927
Wiki	+BART-Semi	Span-Level Only	.94	.895
al-Maqrizi	Baseline	N/A	.84	.680
al-Maqrizi	+BERT	N/A	.36	.300
al-Maqrizi	+BART	Pair- and Span-level	.95	.948
al-Maqrizi	+BART-Semi	Span-Level Only	.95	.947
al-Maqrizi	+BART-Passim	Semi-supervised	.89	.897

target, but frees the annotator from locating the correct source for the target span of interest, which is often the more time-intensive portion of the task.

We can, of course, also relax this assumption that the human annotator needs to mark the span of interest as well by selecting a span of interest using some automatic method. For Wikipedia, this is done by construction as the source links are themselves chosen automatically when constructing the dataset. For al-Maqrizi, such a dataset can be constructed using the raw passim alignments that the annotator used as the basis to construct the dataset in the first place. Rather than using the human-annotated target spans and verified source documents, the model can be trained on model-retrieved target spans and source documents.

5. Experiments

We divide our experiments into two sections; retrieval-oriented and generation-oriented. The retrieval experiments are meant to explore how effective different forms of models with varying degrees of supervision are at solving the problem of source attribution, and are evaluated using Recall@10 and Mean Reciprocal Rank. The generation experiments aim to understand the degree to which generative models are capable of learning to copy from source documents, rather than relying on the ability to fill in masked text using only the surrounding context using the understanding of language gained during pretraining. To this end, we evaluate the accuracy of the generator when tasked with predicting the masked text, rather than employing the generator as a reranker as in the retrieval experiments. This is of particular interest in the domain of Wikipedia, as it was part of the model's pretraining dataset, so data leakage from the pretraining data may have occurred.

5.1. Retrieval Experiments

We will now describe our experiments on the Wikipedia and al-Maqrizi datasets. As a baseline, we use a simple bag-of-words BM25 retrieval model as implemented in pyserini.[9] For Wikipedia, the baseline retrieval model attains a recall at 10 of .64 and MRR of .478. If we then use these retrieval results as input to a BART-based (i.e. generation only) reranking model, which has been trained to generate the link text conditioned on the masked target text and

Table 3
Results for Generation Experiments

Training Data	Generation Accuracy
None	0
Target-Only	.170
Target and Source	.714
Target and Semi-Supervised Source	.709

complete source text, the recall at 10 increases to .97 and the MRR to .927. The purely retrieval-based model, untrained BERT, display vastly worse performance than the baseline model Training with partial supervision performs almost as well as a fully supervised model with .94 recall at 10 and .895 MRR, despite the worse performance of the baseline retrieval model.

On the al-Maqrizi dataset, the baseline retrieval model attains a recall at 10 of .84 and an MRR of .680. Reranking with BERT without any further pretraining for the task actually degrades performance, decreasing recall at 10 to .36 and MRR to .30. Again, similar to Wikipedia, reranking with a purely generative approach significantly improves performance, reaching .95 recall at 10 and an MRR of .948. Interestingly, as with the Wikipedia dataset, most of this performance increase is maintained if we switch the training data from fully supervised training to a semi-supervised setup with only span-level supervision in the target documents, with a very minor decrease in MRR to .947. This continues to hold true even when one relaxes the constraint that the spans to generate also be human-annotated, as one can see from the BART-Passim model, where the model outperforms the retrieval baseline both in terms of MRR and Recall@10. Due to the extreme length of the masked sections in this data set, a similar evaluation for the predictive accuracy of BART without conditioning on the source document would likely be uninformative as the odds of correctly predicting an entire span of 100 subword tokens exactly correctly would be much lower, making the exact match evaluation as performed on the Wikipedia data in the next section much less informative.

The downside of this approach is that, in addition to telling the model to generate some text which may be derived from the source, both these approaches also force the model to generate either some or all of the unrelated content in the sentence not derived from the source. This has two main effects on performance. Firstly, all pages that are not useful for explaining part of a text become less likely to be sources according to the trained reranker. In contrast, source candidates that do happen to explain part of the unrelated text are erroneously considered more likely to be sources. One way to get around this would be to use a text reuse model detection to attempt to identify passages of direct reuse from the potential source set, and bias them asking to make tokens marked as reuse more often than those not marked as reuse, potentially making the reconstruction loss more useful as a means of determining whether a potential source is useful or not.

5.2. Generation Experiments

The improved performance on Wikipedia link prediction may come from the generative ability imparted by BART's pretraining rather than the learned ability of the fine-tuned model to

copy from sources. Furthermore, for the Wikipedia experiments, the training and test sets are themselves a part of what BART was originally pretrained on, making data leakage between the pretraining dataset and this downstream task possible. To test that this isn't artificially inflating the models' source reranking performance, we measure the predictive accuracy of three forms of BART-only models at the task of filling in the masked link text on the Wikipedia dataset. The results for these experiments can be seen in Table 3. First, we test completely untuned BART-base using the text of the target only. Secondly, we use the same test input, but fine tune the model on the link prediction task. Finally, we do the full conditioning on the masked target and source section as described above. The untuned model was completely unable to predict the link text (0% accuracy), while the target-only model achieved 17% accuracy. The combined source-target model, which is the fully supervised setup from the previous section, achieved 71.4% accuracy. This shows that this training process is indeed teaching the model to copy from the source document, rather the ability to rerank being a side effect of the presence of Wikipedia in BART's pretraining data. This predictive performance is largely maintained when we use the highest-ranked BM25 retrieved source in place of the known-correct source, with a very minor decrease to 70.9% accuracy.

6. Discussion and Future Work

The experiments discussed in the previous section provide insight into how well different forms of large language model can be used to solve the problem of source attribution, as well as the importance of task-specific fine tuning. It is clear that generative models like BART are capable of learning more complex source-target relationships than simple copying as evidenced by the strong performance of BART on both data sets. However, the required annotation to train such a model makes it unattractive as a general solution to the problem, though they may be useful for small specialized domains like that of al-Maqrizi in this case. In contrast, the poor performance of the untuned BERT model indicates that some additional training is necessary.

There are several interesting avenues for potential future research on this topic. Further work on unsupervised methods like retrieval augmented generation may be appealing with access to better hardware capable of running a more complete version of the model. The experiments also show the potential benefits of applying semi-supervised methods to this problem to avoid costly annotation for, as it seems, a small loss in performance. It would be worth evaluating this approach on larger datasets to see if the conclusions we draw from the al-Maqrizi test set generalize. Indeed, one benefit of moving to a semi-supervised approach where both the spans and document pairs don't require human annotation is that all the human annotated data can be used for evaluation. We avoid doing so here in the interest of fair comparison between the various forms of model that do require human annotated training data. It would also be worthwhile to examine the performance of fine-tuned BERT trained to embed sources and targets more closely as a potential reranker.

Additionally, as al-Maqrizi and Wikipedia represent very extreme notions of what source-target relationships look like, it would be valuable to find another dataset where citation is more formally structured than al-Maqrizi, but less than in Wikipedia as a third use case, ideally without the OCR errors present in the Internet Archive dataset. For instance, once could

imagine looking the the work of 19th century philosopher J. S. Mill and his sources, which not only have a citations more along the lines of what one sees in modern writing, but would also allow one to examine the utility of these methods in a cross-lingual setting, as he often cites sources in languages other than English that are also digitized [14]. One could also attempt to look at Wikipedia's citations to other sources such as Google Books or the Internet Archive. As the human record becomes more tractable to computation, models of source attribution promise not only to improve search but also to help us understand the reading and writing methods of the past.

Acknowledgments

References

[1]   M. Barber. "Fatimid historiography and its survival. A case study of the vizierate of al-Yāzūrī (r. 442-450/1050-1058)". PhD thesis. University of Edinburgh, 2021.

[2]   F. Bauden. "Ismaili and Fatimid Studies in Honor of Paul E. Walker". In: ed. by B. D. Craig. Middle East Documentation Center, 2010. Chap. Maqriziana XII. Evaluating the Sources for the Fatimid Period: Ibn al-Maʾmūn alBaṭāʾiḥī's History and Its Use by al-Maqrīzī (with a Critical Edition of His Résumé for the years 501-515), pp. 33–85.

[3]   F. Bauden. "Maqriziana II: Discovery of an Autograph Manuscript of al-Maqrīzī: Towards a Better Understanding of His Working Method, Analysis". In: *Mamluk Studies Review* 12 (2008), pp. 51–118.

[4]   F. Bauden. "Maqriziana IX: Should al-Maqrīzī Be Thrown Out with the Bath Water? The Question of His Plagiarism of al-Awḥadī's Khiṭaṭ and the Documentary Evidence". In: *Mamluk Studies Review* 24 (2010), pp. 159–232.

[5]   F. Bora. *Writing history in the medieval Islamic world: the value of chronicles as archives.* Bloomsbury, 2019.

[6]   W. Lan, Y. Chen, W. Xu, and A. Ritter. "An Empirical Study of Pre-trained Transformers for Arabic Information Extraction". In: *arXiv:2004.14519 [cs]* (2020). URL: http://arxiv.org/abs/2004.14519.

[7]   M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension". In: *CoRR* abs/1910.13461 (2019). arXiv: 1910.13461. URL: http://arxiv.org/abs/1910.13461.

[8]   P. S. H. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks". In: *CoRR* abs/2005.11401 (2020). arXiv: 2005.11401. URL: https://arxiv.org/abs/2005.11401.

[9]   J. Lin, X. Ma, S.-C. Lin, J.-H. Yang, R. Pradeep, and R. Nogueira. "Pyserini: A Python Toolkit for Reproducible Information Retrieval Research with Sparse and Dense Representations". In: *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. 2021, pp. 2356–2362.

[10] D. Little. *An Introduction to Mamlūk Historiography: An analysis of Arabic Annalistic and Biographical sources for the Reign of al-Malik an-Nāṣir Muḥammad ibn Qalāʾūn*. F. Steiner, 1970.

[11] Y. Mao, P. He, X. Liu, Y. Shen, J. Gao, J. Han, and W. Chen. *Generation-Augmented Retrieval for Open-domain Question Answering*. 2021. arXiv: 2009.08553 [cs.CL].

[12] al-Maqrizi. *al-Mawa'iz wa-l-Itti'bar fi Dhikr al-Khitat wa-l-Athar*. Ed. by A. F. Sayyid. Vol. 1. Furqan, 2013.

[13] E. Muhanna. *The World in a Book: Al-Nuwayri and the Islamic Encyclopedic Tradition*. Princeton University Press, 2018.

[14] H. O'Neill, A. Welsh, D. A. Smith, G. Roe, and M. Terras. "Text mining Mill: Computationally detecting influence in the writings of John Stuart Mill from library records". In: *Digital Scholarship in the Humanities* (2021).

[15] C. F. Robinson. "The Oxford History of Historical Writing: Volume 2: 400-1400". In: ed. by S. Foot and C. F. Robinson. Oxford University Press, 2012. Chap. Islamic Historical Writing, Eighth through the Tenth Centuries, pp. 238–264.

[16] M. Romanov and M. Seydi. *OpenITI: a Machine-Readable Corpus of Islamicate Texts*. Version 2019.1.1. Zenodo, 2019. DOI: 10.5281/zenodo.3082464. URL: https://doi.org/10.5281/zenodo.3082464.

[17] H. Singh, R. West, and G. Colavizza. "Wikipedia Citations: A comprehensive dataset of citations with identifiers extracted from English Wikipedia". In: *CoRR* abs/2007.07022 (2020). arXiv: 2007.07022. URL: https://arxiv.org/abs/2007.07022.

[18] D. A. Smith, R. Cordel, E. M. Dillon, N. Stramp, and J. Wilkerson. "Detecting and modeling local text reuse". In: *IEEE/ACM Joint Conference on Digital Libraries*. 2014, pp. 183–192. DOI: 10.1109/jcdl.2014.6970166.

[19] K. Thai, Y. Chang, K. Krishna, and M. Iyyer. *RELIC: Retrieving Evidence for Literary Claims*. 2022. arXiv: 2203.10053 [cs.CL].

[20] J. Zhang and L. Zhu. "Citation recommendation using semantic representation of cited papers' relations and content". In: *Expert Systems with Applications* 187 (2022), p. 115826. DOI: https://doi.org/10.1016/j.eswa.2021.115826. URL: https://www.sciencedirect.com/science/article/pii/S095741742101191X.