







Ar-Q-Former: Historical Newspaper Article Separation Based on Multimodal Transformer Structure

Wenjun Sun¹ , Nancy Girdhar² , Hanh Thi Hong Tran³ ,
Carlos-Emiliano González-Gallardo⁴ , Mickaël Coustaty¹ ,
and Antoine Doucet^{1,5} 

¹ L3i, University of La Rochelle, La Rochelle, France
{wenjun.sun,mickael.coustaty,antoine.doucet}@univ-lr.fr

² University of Plymouth, Plymouth, UK
nancy.girdhar@plymouth.ac.uk

³ Arkhn, Paris, France
hanh.tran@arkhn.com

⁴ LIFAT, University of Tours, Tours, France
gonzalezgallardo@univ-tours.fr

⁵ Faculty of Computer and Information Science,
University of Ljubljana, Ljubljana, Slovenia
antoine.doucet@fri.uni-lj.si

Abstract. Article separation for historical newspapers is an important task in the analysis of historical documents. But so far, it remains an under-researched area. Since historical newspapers themselves have two modalities, text, and image, this requires multimodal processing for information extraction. To tackle this task while accounting for the unique characteristics of historical newspapers, we propose an article separation model using a multimodal transformer structure. This model processes the newspaper image along with the bounding boxes and text content of its text blocks, linking them based on a predefined rule to reconstruct the overall page structure. The text and image information of the connected text blocks are then fed into a cross-modal transformer, and the classifier determines whether the connections between the text blocks need to be removed or not. The text blocks that remain connected are recognized as forming an article. A mask method is used to allow the image to reflect the positional relationships of the text blocks. We evaluated our architecture on two datasets, Newey's NLF and BNF. The results demonstrate that *Ar-Q-former* significantly outperforms similar structural modeling methods, achieving up to 19% points higher AR_{F1} on NLF and 22% points on BNF. It also reaches a performance level comparable to the reading order simulation method. However, there remains an approximate 10% point gap in *mACS mPPA* compared to rule-based methods, which are specifically tailored to these datasets. Nevertheless, *Ar-Q-former* exhibits greater generalizability. Additionally, this approach introduces multimodal text-image analysis and interaction compared to previous studies by innovatively incorporating the mask-image method to capture visual and positional information between text blocks.

Keywords: Article Separation · Historical Newspaper · Multimodal Analysis · Layout Analysis · Information Retrieval

1 Introduction

Historical newspapers are invaluable resources that record social, political, and economic events across different periods of history. Extracting information from these documents helps preserve cultural heritage and prevents the loss of important historical knowledge. With the advancement of information technology, an increasing number of historical newspapers are being digitized, as seen in initiatives like *Europeana* [18] and *NewsEye* [8]. However, the digitization process presents significant challenges, particularly in efficiently and accurately processing the diverse content of historical newspapers.

A central challenge in this context is *article separation* (AS), also known as *article segmentation*, which involves segmenting a newspaper page into individual articles. Figure 1 is an example of article separation in a historical newspaper. This task is essential for organizing and indexing the content for tasks such as automated content retrieval, information extraction, and historical analysis. Article separation can be broadly categorized into two approaches: *visual-cue-based* methods, which rely on the structural features of the newspaper layout, and *textual-cue-based* methods, which utilize textual features such as text similarity and linguistic patterns to identify potential article segments [30].

Visual-cue-based methods are widely applied, but they often fail when dealing with historical newspapers, where degraded quality, diverse formatting, and nonstandard layouts complicate layout-based analysis. In contrast, textual cue-based methods offer a more flexible approach but may not always capture the complex visual relationships between text blocks [12]. Furthermore, historical newspapers often exhibit issues such as skewed, rotated, or blurred text, which further hinder the effectiveness of traditional visual cue-based methods [13]. The variability in layout patterns across different periods, regions, and publishers further complicates the task of article separation.

To address these challenges, we introduce a novel multimodal approach for article separation in historical newspapers, leveraging both textual and visual information to improve segmentation accuracy. Our method, *Ar-Q-former*, integrates both textual and visual information to identify the boundaries of the article effectively. Specifically, we use a cross-modal transformer to process linked text blocks, combining textual content and visual cues to form coherent article segments. Additionally, our approach introduces the mask-image technique, which preserves the positional relationships between text blocks, further enhancing the segmentation accuracy. Our method operates under the assumption that both the image of the page and the position/content of each text block are known. We connect each text block to its neighbouring units below and right, effectively modeling the page structure. For each connection (link), we use the text backbone to obtain the text semantic vectors of the text blocks at both ends of the connection. At the same time, we construct the mask-image of these

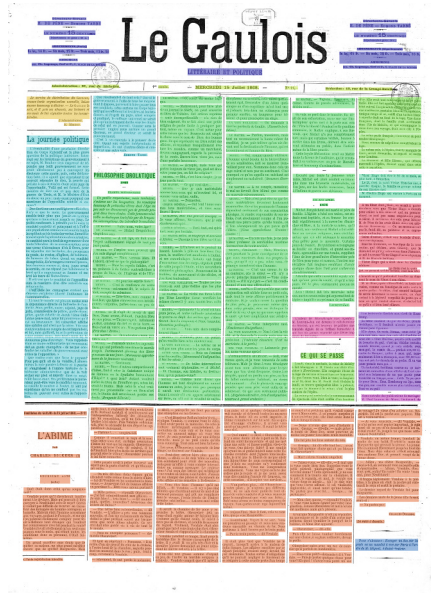


Fig. 1. A demonstration of article separation on a newspaper page from *Le Gaulois*, part of the *BNF* collection within the *NewsEye* project. Different articles are visually distinguished using unique colors, illustrating the segmentation process.

two blocks, and use the image backbone to obtain the visual semantic vectors of the mask-image. Finally, we input the text and visual semantic vectors into the cross-modal transformer (*Ar-Q-former*) to capture the semantics of the connection and determine whether the connection should be preserved by a classifier. Ultimately, the text blocks that still retain the connection form an article.

The main contributions of this work are as follows.

1. We propose a novel multimodal article separation method, namely *Ar-Q-former*, that combines visual and textual cues for accurate article segmentation in historical newspapers. It is the first approach that uses both visual and textual modal interaction ideas in the article separation task and extends the structure of *Q-Former* to enable it to obtain semantic information from both modalities by setting up an additional text query while preserving the vision query.
2. We introduce the mask-image method to effectively model the positional relationships between text blocks, improving the integration of image and text information.
3. We demonstrate the efficacy of our approach through extensive experiments on historical newspaper datasets, achieving competitive performance against existing methods.

The remainder of the paper is organized as follows. Section 2 reviews related work on article separation techniques. In Sect. 3, we detail our proposed method,

including the multimodal transformer architecture and the mask-image technique. Section 4 describes the datasets used for the evaluation, while Sect. 5 outlines the experimental setup and presents the results. Finally, we summarize our findings and explore potential directions for future research in Sect. 6.

2 Related Work

Article separation in historical newspapers remains a crucial yet underexplored problem in document analysis. Existing approaches to document segmentation can be broadly categorized into rule-based methods, machine learning techniques, and deep learning-driven strategies [13]. However, historical newspapers pose unique challenges, including complex layouts, poor print quality, and the need for multimodal processing, which requires more advanced methodologies.

Earlier approaches to document segmentation relied primarily on rule-based methods, which leverage explicit layout cues such as column structures, whitespace, and ruling lines (*horizontal* and *vertical* separators) to delineate articles [11]. Connected component analysis (CCA) and geometric heuristics have been used to group text blocks into coherent structures [16]. While these methods are effective for structured layouts, they struggle with historical newspapers due to non-uniform formatting, irregular text alignments, and ink degradation. Moreover, rule-based approaches often require dataset-specific parameter tuning, limiting their adaptability across diverse newspaper archives [2].

To address the limitations of heuristic approaches, machine learning models have been introduced for article segmentation [20]. These models utilize hand-crafted features such as font size, embeddings, and spatial relationships to classify text blocks [1, 9, 12]. Probabilistic graphical models, including conditional random fields (CRFs), have been explored to model contextual dependencies between adjacent text blocks [24]. While these approaches improve generalizability compared to heuristic methods, they require extensive feature engineering and fail to jointly model textual and visual modalities [28], which are critical for historical newspaper analysis.

With the advent of deep learning, convolutional neural networks (CNN) and transfer learning architectures have demonstrated significant advances in document analysis tasks [14]. Fully convolutional networks (FCNs) and U-Net variants have been applied for semantic segmentation of historical newspaper pages, achieving promising results in text and non-text region classification [23]. However, these pixel-based approaches often struggle with precise article boundary identification, especially when textual and visual elements are intertwined.

Recently, multimodal deep learning models have gained traction in document analysis. The combination of CNNs for visual feature extraction and recurrent neural networks (RNNs) for textual understanding has been explored in layout analysis tasks [3]. The rise of vision transformers (ViTs) has further enabled joint modeling of document images and textual structures [6, 29]. Although some transformer-based models have been employed for document layout recognition [19, 25], their application to article separation in historical newspapers remains

largely unexplored. The integration of multimodal learning for article separation is a relatively nascent area. Existing works incorporating both textual and visual cues focus mainly on modern document layouts, overlooking the complexities of historical documents [5]. The necessity of modeling cross-modal interactions between text and images has led to the development of graph-based approaches, where relationships between text blocks are explicitly modeled as graph nodes and edges [10]. However, these methods often assume high-quality OCR results, which are often not feasible for degraded historical prints.

Table 1. Summary of approaches for article segmentation in historical newspapers.

| Refs | Method | Visual | Textual | Limitation |
|---------------------------|----------------------------|--------|---------|---|
| [11, 16] | Rule-based | ✓ | ✗ | Struggles with non-uniform formatting, irregular text alignment, ink degradation. Requires dataset-specific parameter tuning. |
| [1, 9, 12] | Machine learning | ✓ | ✓ | Requires extensive feature engineering. Fails to jointly model textual and visual modalities. |
| [23] | Deep learning (CNNs) | ✓ | ✗ | Struggles with precise article boundary identification, especially when text and visual elements are intertwined. |
| [3, 6, 29] | Deep learning (Multimodal) | ✓ | ✓ | Focuses on modern documents; challenges in dealing with historical newspaper complexities (degraded prints). |
| [10] | Graph-based | ✓ | ✓ | Assumes high-quality OCR results, which is often not feasible for historical documents with degraded prints. |
| Ar-Q-former (ours) | Deep learning (Multimodal) | ✓ | ✓ | Robust to varying layouts in historical newspapers; dynamic learning through end-to-end multimodal interaction. |

Our proposed *Ar-Q-former* builds upon prior multimodal segmentation research while addressing the distinct challenges of historical newspaper analysis. This is the first time that the Q-Former structure has been utilized for multimodal semantic extraction and combined with structural for the analysis of historical newspapers in the article separation task. By leveraging a transformer-based cross-modal encoder, our approach integrates textual content and spatial relationships in a unified manner. Unlike previous methods that rely solely on CNN-based feature extraction or rule-based heuristics, our model dynamically learns article structures through end-to-end multimodal interaction, ensuring robust performance across varying newspaper layouts. Furthermore, the introduction of the mask-image method provides explicit positional awareness, which improves segmentation accuracy in complex historical newspapers. Table 1 gives a general introduction to the relevant methods.

3 Method

Our proposed method comprises four key steps, as shown in Fig. 2, the (i) *structural modeling* phase establishes connections between text blocks based on spatial arrangement, the (ii) *textual and visual semantic extraction* phase encodes meaningful representations, the (iii) semantic acquisition of connections is done with *Ar-Q-former* to capture cross-modal relationships, and the (iv) *link classification and article prediction* retains or remove connections. After these steps, text blocks that remain connected are considered to form an article.

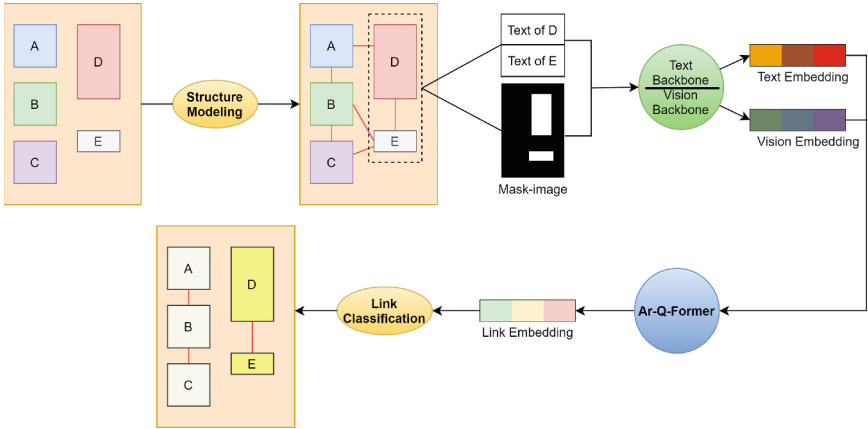


Fig. 2. Pipeline of *Ar-Q-former*.

Ar-Q-former extends the structure of Q-Former [21] to transform it from an image understanding Q&A model to a document multimodal semantic extraction model, a change that is achieved by adding text query. In the original structure, only vision query is used to extract visual semantics, and the goal is to align the visual semantics with the textual semantics and generate vectors that are suitable for text generation, rather than reflecting the semantic differences between text blocks. And *Ar-Q-former* adds text query on top of vision query, which is used to allow the model to extract text semantic features, instead of simply accepting only the raw output of the text backbone. For the vision unit, mask-image is also used to reflect the relationship between the position and layout of the two text blocks, which makes *Ar-Q-former* more suitable for the semantic extraction task compared to Q-Former. To explore the effect of the number of text and vision queries on semantics extraction, we also tested the model’s performance when the number of queries was expanded from 64 to 512.

3.1 Structure Modeling

In this step, we model the structure of historical newspapers to extract articles. As the text blocks that form an article are typically adjacent to each other

in the reading order, we connect text blocks based on proximity. Due to the absence of precise annotation of the reading order of historical newspaper data, we model the layout of the newspaper in a top-to-bottom and left-to-right order. For instance, connect the text block below and the text block right, which has the smallest Euclidean distance. When finding the lower block B in block A , the distance between the lower block B in block A , the distance between the midpoint of the lower line of the border box of A and the midpoint of the upper line of the border box of B is computed. In contrast, when finding the right block R , the midpoint of the right line of A and the midpoint of the left line segment of R are computed. Figure 3 shows examples of text block linking, with a relatively simple layout on the left and a relatively complex one on the right.

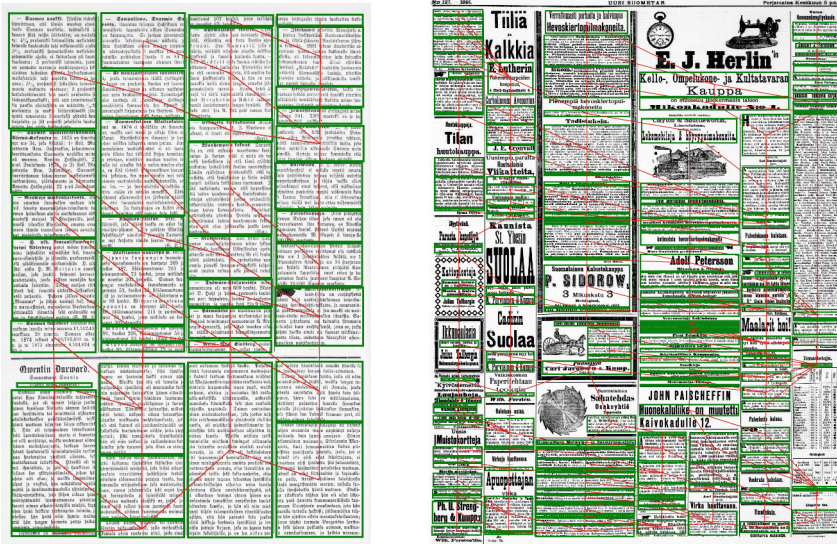


Fig. 3. Examples of text block linking.

3.2 Textual and Visual Semantic Extraction

Following the structural modeling, we get multiple connections, where each connection consists of two linked text blocks. To assess whether a connection should be retained, we analyze the semantic relationship between the two text blocks.

For textual semantics, we use a pre-trained language model based on *hmBERT* [26], which is specifically trained on historical newspapers. While *hmBERT* was pre-trained for named entity recognition, we adapted it to represent paragraph-level semantics by overlaying a transformer encoder layer on top of the original language model. This approach helps mitigate errors that could arise from using the model for historical text semantic extraction [4]. We use the

embedding of the [cls] token as the semantic representation for the entire text block.

Newspaper text blocks contain not only the textual information modality but also the layout and visual modalities, so it is necessary to analyze the semantic information of both modalities simultaneously. Most of the previous work is to obtain the semantics of the layout by embedding the positional information of the text directly into vectors [17, 22], but the pre-training data of these works are structured text, such as receipts and invoices. However, the layout of the newspaper is more complex, coupled with the insufficient volume of training data. Some related works [27, 28] show that it is difficult to enhance the semantic differentiation of text blocks by using only simple position vectors. Facing this problem, we propose the method of mask-image to unite layout and visual information: only retain the images of the two text blocks connecting the two ends and mask the other regions. We selected Vision Transformer [7] as the visual backbone and processed the mask-image to obtain semantic vectors for vision and layout modalities. Also, in order to explore the effect of separator lines on layout understanding, we binarized the newspaper image in addition to masking all the remaining areas; pixel dilation and erosion operations were performed in order to filter out pixels such as text and images, retaining only linear objects. Figure 4 shows an example of mask-image with and without separator lines.

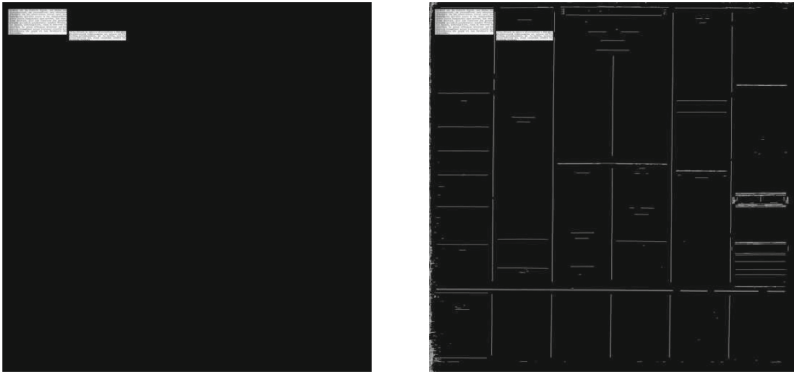


Fig. 4. Examples of mask-images. The left and right sides are without and with separator lines, respectively.

3.3 Ar-Q-Former

Ar-Q-former is a cross-modal semantic transformer model inspired by the BLIP-2 [21]. This general structure is shown as Fig. 5. It consists of two main parts: a visual unit and a textual unit. Each unit is composed of an encoder and decoder with a transformer structure. The encoder consists of self-attention layers, while the decoder employs learnable query vectors to capture the semantics

of the corresponding modality. Text and image information are first transformed into modal semantic vectors through the corresponding backbone. These vectors are then fed into the encoder of the corresponding modality unit for individual processing and then accompanied by the query for cross-attention operation to achieve the interaction between the text and visual modalities. The output of the entire *Ar-Q-former* is then used to describe the common semantics of the two text blocks, i.e., to describe the connection semantic information. The *Ar-Q-former* uses the same structure as *BERT*, and the query vectors are randomly initialized by a layer of learnable parameters.

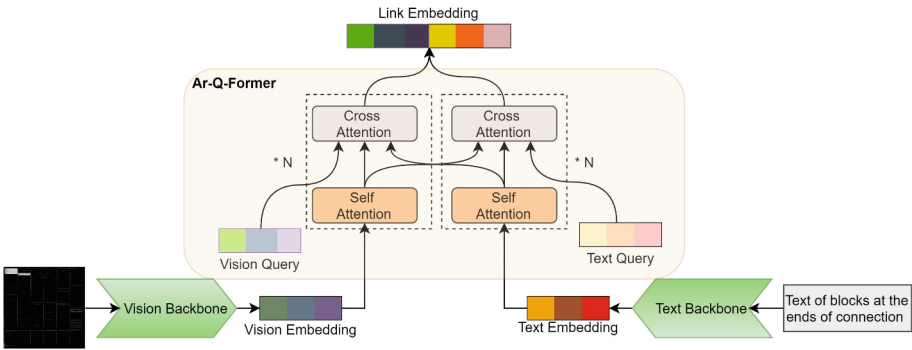


Fig. 5. Structure of Ar-Q-former.

3.4 Link Classification and Article Prediction

After obtaining the semantic vectors from *Ar-Q-former*, we classify the connections between text blocks. These vectors are fed into a classifier to determine whether the connection should be retained or removed. If the connection is retained, it implies that the two text blocks are semantically related and belong to the same article. If the connection is removed, the text blocks are considered independent and should be placed in separate articles.

Once all connections have been classified, article separation is complete. Text blocks that remain connected will form a complete article. To build an efficient classifier, we use a multi-layer fully connected network. The semantic vectors from *Ar-Q-former* serve as input, and the final classification result is generated through a single output layer. ReLU activation and dropout are applied after each fully connected layer to improve the generalization ability and prevent overfitting of the model. This architecture enables the classifier to accurately identify the relationships between text blocks, ensuring precise article segmentation.

4 Experiments

We used two subsets, BNF and NLF, from the publicly accessible NewsEye¹ dataset [30]. The evaluation metrics used were mean article coverage score ($mACS$), mean proper predicted articles ($mPPA$) [12], precision (AR_P), recall (AR_R), and F1-score (AR_{F1}) [15]. We illustrate the detailed parameters we used in our experiments and compare the results of *Ar-Q-former* against other benchmarks on these datasets.

4.1 Datasets

The NewsEye dataset [30], a collection of newspapers from the 19th to the 20th century, was used for the experiments and labelled by the annotator at the granularity of paragraphs and articles, which includes the border-box of each text block (paragraph), the textual content, and the index of the article they belong to. For data accessibility reasons, we used two subsets of the dataset: NLF and BNF, sourced from the national libraries of Finland and France, in Finnish and French, respectively. The dataset statistics are summarized in Table 2.

Table 2. Statistics of the NLF and BNF subsets of the NewsEye dataset.

| Dataset | Pages | Sentences | Paragraphs | Articles |
|---------|-------|-----------|------------|----------|
| NLF | 200 | 22,042 | 6,348 | 3,282 |
| BNF | 182 | 50,698 | 6,792 | 3,061 |

4.2 Evaluation Metrics

As article separation is a complex task, some specific evaluation metrics have been proposed, including mean article coverage score ($mACS$), mean proper predicted articles ($mPPA$) [12], precision (AR_P), recall (AR_R), and F1-score (AR_{F1}) [15].

$mACS$ and $mPPA$ evaluate the effect of article separation from the perspective of a single article and the overall newspaper page, respectively. $mACS$ is defined as

$$mACS = 1 - \sum_{x=1}^n \frac{AER_x}{n}, \quad (1)$$

where AER_x describes the error rate for a single article, which in turn generalises to the entire test set. Assuming that the whole test set contains n articles, PTP_x and GTP_x denote the prediction and truth value for the x^{th} article, respectively, which are the set of indexes of the text block. AER_x is defined as

¹ <https://www.NewsEye.eu/>.

$$AER_x = \frac{|PTP_x \oplus GTP_x|}{|PTP_x \cup GTP_x|}. \quad (2)$$

In evaluating article separation from the page perspective, we assume that there are P pages in the test set, with m articles per page, and the model successfully predicts n articles. The $mPPA$ is then defined as

$$mPPA = \frac{\sum_{p=1}^P \frac{n}{m}}{P}. \quad (3)$$

AR_P , AR_R , and AR_{F1} were initially designed for text recognition tasks but have been adapted for article separation [15]. The ground truth labels are denoted as GT , containing M articles, each represented as a set of text block indices. Predictions are denoted as PT , containing N articles. We construct an evaluation matrix Eva , where the element $item_{ij}$ corresponds to the precision value between GT_i and PT_j . The precision is calculated using Algorithm 1.

Algorithm 1. Greedy Algorithm

Form matrix $Eva \in R^{M \times N}$

$Result \leftarrow []$

while Eva is not empty **do**

$max \leftarrow$ one of the maximal elements in Eva

 Add max into $Result$

$Eva \leftarrow$ take Eva and delete corresponding row and column of max

end while

return $Avg(Result)$

AR_R is also computed in the same manner. The difference is that the items in Eva are recall values. After AR_P and AR_R are obtained, the AR_{F1} is calculated as

$$AR_{F1} = \frac{2 * AR_P * AR_R}{AR_P + AR_R}. \quad (4)$$

4.3 Benchmarks

In order to propose a fair comparison of our approach to the state-of-the-art, we used three methods, namely NewsEye [8], STRAS [12] and LIAS [28], which were developed on the selected dataset. NewsEye employs a graph neural network (GNN) to model the semantics of each text block after structural modeling, followed by clustering to group blocks into articles. STRAS is a fully rule-based method that relies on semantic similarity and other metrics to predict articles. LIAS first establishes the reading order of the text blocks and then analyzes adjacent blocks to segment the reading order and predict articles.

4.4 Experiment Details

The kernel size for the morphological operations used to obtain separator lines in *Ar-Q-former* is $(\sqrt{width} \times 1.2, 1)$ and $(1, \sqrt{height} \times 1.2)$. For the text backbone, we used pre-trained *hmBert*² and kept it frozen during the experiment. For the vision backbone, we used ViT for weight initialization, which was integrated into the training process. The input image size was set to (512, 512). *Ar-Q-former* uses a BERT structure with randomly initialized weights at the beginning of training, and the maximum position was set to 1,024. The entire model was trained using the AdamW optimizer with a learning rate of 5e-5, a batch size of 64, and a dropout rate of 0.1. The training was conducted on 2 Nvidia H100 GPUs.

5 Results and Analysis

We report in Table 3 a comparison of performance between our proposed architecture and three groups of benchmarks, including structural modeling, rule-based, and reading order-based methods. Additionally, we explore the effect of the model’s parameters on the results.

Table 3. Comparison between *Ar-Q-former* and other benchmarks on NLF and BNF datasets.

| Methods | NLF | | | | | BNF | | | | |
|---------------------------------------|-------------|-------------|-----------------------|-----------------------|------------------------|-------------|-------------|-----------------------|-----------------------|------------------------|
| | <i>mACS</i> | <i>mPPA</i> | <i>AR_P</i> | <i>AR_R</i> | <i>AR_{F1}</i> | <i>mACS</i> | <i>mPPA</i> | <i>AR_P</i> | <i>AR_R</i> | <i>AR_{F1}</i> |
| Structural modeling | | | | | | | | | | |
| <i>NewsEye_{dbscan}</i> | 0.375 | 0.066 | — | — | 0.754 | 0.447 | 0.105 | — | — | 0.697 |
| <i>NewsEye_{greedy}</i> | 0.327 | 0.054 | — | — | 0.757 | 0.356 | 0.094 | — | — | 0.652 |
| <i>NewsEye_{hierarchical}</i> | 0.382 | 0.061 | — | — | 0.757 | 0.538 | 0.139 | — | — | 0.690 |
| Rule-based | | | | | | | | | | |
| <i>STRAS_{pre}</i> | 0.790 | 0.606 | — | — | — | 0.800 | 0.634 | — | — | — |
| <i>STRAS_{sg}</i> | 0.861 | 0.785 | — | — | — | 0.834 | 0.700 | — | — | — |
| <i>STRAS_{cbow}</i> | 0.855 | 0.792 | — | — | — | 0.806 | 0.631 | — | — | — |
| <i>STRAS_{ft}</i> | 0.827 | 0.700 | — | — | — | 0.798 | 0.612 | — | — | — |
| Reading order-based | | | | | | | | | | |
| <i>LIAS</i> | 0.907 | 0.719 | 0.961 | 0.955 | 0.957 | 0.870 | 0.588 | 0.897 | 0.943 | 0.919 |
| <i>LIAS+bbox</i> | 0.886 | 0.688 | 0.949 | 0.958 | 0.952 | 0.804 | 0.470 | 0.888 | 0.916 | 0.901 |
| Multimodal | | | | | | | | | | |
| <i>Ar-Q-former₅₁₂</i> | 0.850 | 0.683 | 0.909 | 0.970 | 0.948 | 0.709 | 0.624 | 0.879 | 0.962 | 0.919 |

² [bert-base-historic-multilingual-64k-td-cased](#).

5.1 Ar-Q-Former vs. Structural Modeling

$NewsEye_x$ represents the NewsEye method using different clustering algorithms, and the results show that our method greatly outperforms the benchmark in all metrics. The NewsEye method is the most similar to our proposed method, i.e., firstly, the structure of the newspaper’s page is modeled to obtain the semantic vectors of the text block, and then article aggregation is performed. The difference between the two methods is that after NewsEye uses the text backbone to obtain the initial semantic vectors, GNN is used to process the graph structure composed of text blocks. The feature vectors of the nodes in the graph are the initial semantic vectors. *Ar-Q-former*, on the other hand, uses a stacked transformer encoder to mitigate the semantic error of the backbone. NewsEye only uses the semantics of one text modality, while *Ar-Q-former* uses mask-image on top of it to introduce visual and layout semantics, which makes the model learn more dimensional features thus improving the results of article separation.

5.2 Ar-Q-Former vs. Rule-Based Approaches

$STRAS_X$ represents the semantic vectors of text blocks obtained using different text backbones. For *mACS*, *Ar-Q-former* achieves similar results to the *STRAS* method on the NLF dataset but worse on the BNF dataset. This phenomenon is due to the fact that the structure of the news pages in the BNF dataset is more complex, and the adaptability of our text block connection method to this dataset is weakened. For *mPPA*, *STRAS* outperforms *Ar-Q-former* on both datasets, which is a result of low *mACS*, and for the entire test set, *Ar-Q-former*’s performance at the article level affects the page level as well. This comparison of results shows that our method is quite a bit farther away from a rule set that is specialized for the dataset. However, one of the major drawbacks of the *STRAS* approach is the lack of generalizability. Since all the rules are specific to the current dataset, they need to be reformulated when facing a new newspaper dataset, which makes it difficult to quickly migrate it to other types of newspapers. However, *Ar-Q-former* can use the same architecture when facing the same problem and only needs to retrain the weights of the semantic extractors for each modality. But one thing we need to think about more deeply is how to improve the performance of the method while ensuring its generalizability.

5.3 Ar-Q-Former vs. Reading Order-Based Approaches

$LIAS_{bb_{ox}}$ denotes the introduction of position embedding on top of the original model. *Ar-Q-former* achieves the best *mPPA* on BNF and performs similarly to $LIAS_{bb_{ox}}$ on other metrics. *LIAS* reconstructs the reading order through separator lines. Whereas *Ar-Q-former*’s structural modeling approach ignores the precise logical order between text blocks. This comparison shows that the reading order is a piece of crucial information for article separation when analyzing the structure by layout. The comparison also shows that simply using position embedding to reflect the layout information of a page does not produce semantic

enhancement in historical newspaper analysis, which is the motivation for *Ar-Q-former* to use mask-image. The results also show that this method enables the model to achieve similar results when key information such as reading order is missing, which also proves the feasibility of mask-image.

5.4 Impact of Query Count and Separator Lines

Table 4. Performance Comparison of *Ar-Q-former* Variants on NLF and BNF datasets

| Params | NLF | | | | | | BNF | | | | | |
|-----------------------------------|--------------|--------------|-----------------------|-----------------------|------------------------|--------------|--------------|--------------|-----------------------|-----------------------|------------------------|--------------|
| | <i>mACS</i> | <i>mPPA</i> | <i>AR_P</i> | <i>AR_R</i> | <i>AR_{F1}</i> | Avg. | <i>mACS</i> | <i>mPPA</i> | <i>AR_P</i> | <i>AR_R</i> | <i>AR_{F1}</i> | Avg. |
| Ar-Q-former ₆₄ | 0.796 | 0.659 | 0.871 | 0.980 | 0.921 | 0.845 | 0.679 | 0.584 | 0.819 | 0.979 | 0.888 | 0.790 |
| Ar-Q-former _{64,no_sep} | 0.725 | 0.595 | 0.823 | 0.974 | 0.898 | 0.803 | 0.466 | 0.449 | 0.679 | 0.780 | 0.726 | 0.620 |
| Ar-Q-former ₁₂₈ | 0.798 | 0.661 | 0.815 | 0.981 | 0.892 | 0.829 | 0.680 | 0.608 | 0.857 | 0.970 | 0.910 | 0.805 |
| Ar-Q-former _{128,no_sep} | 0.772 | 0.582 | 0.829 | 0.977 | 0.892 | 0.810 | 0.669 | 0.580 | 0.675 | 0.946 | 0.788 | 0.732 |
| Ar-Q-former ₂₅₆ | 0.815 | 0.668 | 0.884 | 0.975 | 0.927 | 0.854 | 0.688 | 0.624 | 0.879 | 0.959 | 0.918 | 0.814 |
| Ar-Q-former _{256,no_sep} | 0.803 | 0.639 | 0.853 | 0.969 | 0.907 | 0.834 | 0.665 | 0.604 | 0.850 | 0.953 | 0.898 | 0.794 |
| Ar-Q-former ₅₁₂ | 0.850 | 0.683 | 0.909 | 0.970 | 0.948 | 0.872 | 0.709 | 0.624 | 0.879 | 0.962 | 0.919 | 0.819 |
| Ar-Q-former _{512,no_sep} | 0.825 | 0.640 | 0.879 | 0.903 | 0.890 | 0.827 | 0.681 | 0.620 | 0.857 | 0.955 | 0.903 | 0.793 |

As shown in Table 4, *Ar-Q-former_{x,no_sep}* represents x queries in *Ar-Q-former* for text and visual modalities, respectively, with separator lines removed from mask-image. The results show that across all datasets and metrics, the performance of the model improves accordingly with the increase in query, except for the NLF dataset where there is a decrease in F1 value when increasing the number of queries from 64 to 128, which may be related to the training parameters. Nonetheless, the overall trend remains consistent. An increase in the number of queries implies that more vectors are available to capture the semantic information of both visual and textual modalities, providing more semantically discriminative input vectors for the connection classifier. However, in addition to increasing the number of queries, it may also be beneficial to explore expanding the dimensionality of individual query vectors.

Furthermore, by comparing the results for the same number of queries with and without separator lines in the mask-image, it is evident that incorporating separator lines in the visual modality yields better results. Since the data analyzed are historical newspapers, separator lines can assist the model in understanding the layout information, especially since separator lines act as text segmentation in the newspaper itself.

6 Conclusion

We propose a new multimodal historical newspaper article separation method, *Ar-Q-former*. This approach establishes corresponding query vectors for both

visual and textual modalities and introduces a cross-attention mechanism to facilitate interaction between the two modalities during semantic vector computation, ultimately generating a unified embedding for each text block. It is the first model to use textual and visual multimodality in the historical newspaper article separation task. Based on Q-Former, we added a text query to transform it from an image query model to a multimodal semantic extraction model.

A key feature of this method is the integration of the mask-image, which incorporates the layout semantics of the text block into the visual modality. The method involves connecting text blocks on a newspaper page according to a predefined rule set, followed by structural modeling of the page. A mask-image is created for each connection, and the embeddings of the text and mask-image at both ends of the connection are extracted through the text and vision backbones. The semantic vectors of both modalities are then processed through *Ar-Q-former* to obtain the connection embedding. Finally, a classifier is applied to determine whether the connection should be retained.

We evaluated the method on NewsEye’s NLF and BNF for testing, and the results demonstrate that our method outperforms similar structural modeling techniques and achieves performance comparable to reading-order simulation methods. However, a gap remains compared to rule-based methods specifically tailored for the datasets used. Nonetheless, *Ar-Q-former* exhibits greater generalizability. The results also highlight the importance of query vectors and separator lines, with experiments confirming that increasing the number of query vectors improves the model’s performance. Combining the experimental results of each model, we can see that the symbolic approaches specifically formulated for the dataset are still better than the neuronal approaches proposed for the pursuit of universality, which gives us a direction worth exploring: the symbolic-neuronal approach, i.e., to first formulate a rule set based on the data, e.g., the connection of text blocks or the reading order, and then use neural network models to further analyze it.

Acknowledgments. Co-funded by the European Union HORIZON-WIDERA-2023-TALENTS-01-01 grant 101186647—AI4DH. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

References

1. Aiello, M., Pegoretti, A.: Textual article clustering in newspaper pages. *Appl. Artif. Intell.* **20**(9), 767–796 (2006)
2. Bansal, A., Chaudhury, S., Roy, S.D., Srivastava, J.: Newspaper article extraction using hierarchical fixed point model. In: 2014 11th IAPR International Workshop on Document Analysis Systems, pp. 257–261. IEEE (2014)
3. Barman, R., Ehrmann, M., Clematide, S., Oliveira, S.A., Kaplan, F.: Combining visual and textual features for semantic segmentation of historical newspapers. *J. Data Mining Digit. Humanit. (HistoInformatics)* (2021)

4. Boros, E., et al.: Alleviating digitization errors in named entity recognition for historical documents. In: Proceedings of the 24th Conference on Computational Natural Language Learning, pp. 431–441 (2020)
5. Canhui, X., Yuteng, L., Cao, S., Honghong, Z., Hengyue, B., Yinong, C.: Him: hierarchical multimodal network for document layout analysis. *Appl. Intell.* **53**(20), 24314–24326 (2023)
6. Da, C., Luo, C., Zheng, Q., Yao, C.: Vision grid transformer for document layout analysis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 19462–19472 (2023)
7. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. In: International Conference on Learning Representations (2020)
8. Doucet, A., et al.: Newseye: a digital investigator for historical newspapers. In: 15th Annual International Conference of the Alliance of Digital Humanities Organizations, DH 2020 (2020)
9. Furmaniak, R.: Unsupervised newspaper segmentation using language context. In: Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), vol. 2, pp. 1263–1267. IEEE (2007)
10. Garz, A., Seuret, M., Fischer, A., Ingold, R.: A user-centered segmentation method for complex historical manuscripts based on document graphs. *IEEE Trans. Hum.-Mach. Syst.* **47**(2), 181–193 (2016)
11. Gatos, B., Mantzaris, S., Chandrinou, K., Tsigris, A., Perantonis, S.J.: Integrated algorithms for newspaper page decomposition and article tracking. In: Proceedings of the Fifth International Conference on Document Analysis and Recognition. ICDAR 1999 (Cat. No. PR00318), pp. 559–562. IEEE (1999)
12. Girdhar, N., Coustaty, M., Doucet, A.: STRAS: a semantic textual-cues leveraged rule-based approach for article separation in historical newspapers. In: International Conference on Asian Digital Libraries, pp. 89–105. Springer (2023)
13. Girdhar, N., Coustaty, M., Doucet, A.: Digitizing history: transitioning historical paper documents to digital content for information retrieval and mining—a comprehensive survey. *IEEE Trans. Comput. Soc. Syst.* (2024)
14. Girdhar, N., Sharma, D., Coustaty, M., Doucet, A.: Leveraging transfer learning for article segmentation in historical newspapers. In: International Conference on Theory and Practice of Digital Libraries, pp. 222–238. Springer (2024)
15. Grüning, T., Labahn, R., Diem, M., Kleber, F., Fiel, S.: Read-bad: a new dataset and evaluation scheme for baseline detection in archival documents. In: 2018 13th IAPR International Workshop on Document Analysis Systems (DAS), pp. 351–356. IEEE (2018)
16. Hadjar, K., Hitz, O., Ingold, R.: Newspaper page decomposition using a split and merge approach. In: Proceedings of Sixth International Conference on Document Analysis and Recognition, pp. 1186–1189. IEEE (2001)
17. Huang, Y., Lv, T., Cui, L., Lu, Y., Wei, F.: Layoutlmv3: pre-training for document AI with unified text and image masking. In: Proceedings of the 30th ACM International Conference on Multimedia, pp. 4083–4091 (2022)
18. Isaac, A., Haslhofer, B.: Europeana linked open data—data. *Europeana. eu. Semantic Web* **4**(3), 291–297 (2013)
19. Kastanas, S., Tan, S., He, Y.: Document AI: a comparative study of transformer-based, graph-based models, and convolutional neural networks for document layout analysis. arXiv preprint [arXiv:2308.15517](https://arxiv.org/abs/2308.15517) (2023)

20. Kettunen, K., Ruokolainen, T., Liukkonen, E., Tranouez, P., Antelme, D., Paquet, T.: Detecting articles in a digitized finnish historical newspaper collection 1771-1929: early results using the pivaj software. In: Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage, pp. 59–64 (2019)
21. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: International Conference on Machine Learning, pp. 19730–19742. PMLR (2023)
22. Luo, C., Cheng, C., Zheng, Q., Yao, C.: Geolayoutlm: geometric pre-training for visual information extraction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7092–7101 (2023)
23. Meier, B., Stadelmann, T., Stampfli, J., Arnold, M., Cieliebak, M.: Fully convolutional neural networks for newspaper article segmentation. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 1, pp. 414–419. IEEE (2017)
24. Palfray, T., Hebert, D., Nicolas, S., Tranouez, P., Paquet, T.: Logical segmentation for article extraction in digitized old newspapers. In: Proceedings of the 2012 ACM Symposium on Document Engineering, pp. 129–132 (2012)
25. Pillai, P., Mangsuli, P.: Document layout analysis using detection transformers. In: Abu Dhabi International Petroleum Exhibition and Conference, p. D031S102R001. SPE (2021)
26. Schweter, S., März, L., Schmid, K., Çano, E.: hmbert: historical multilingual language models for named entity recognition. arXiv preprint [arXiv:2205.15575](https://arxiv.org/abs/2205.15575) (2022)
27. Sun, W., Tran, H.T.H., González-Gallardo, C.E., Coustaty, M., Doucet, A.: Global-seg: text semantic segmentation based on global semantic pair relations. In: International Conference on Document Analysis and Recognition, pp. 253–269. Springer (2024)
28. Sun, W., Tran, H.T.H., González-Gallardo, C.E., Coustaty, M., Doucet, A.: Lias: layout information-based article separation in historical newspapers. In: International Conference on Theory and Practice of Digital Libraries, pp. 256–272. Springer (2024)
29. Tang, Z., et al.: Unifying vision, text, and layout for universal document processing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19254–19264 (2023)
30. Girdhar, N., Coustaty, M., Doucet, A.: Benchmarking NAS for article separation in historical newspapers. In: International Conference on Asian Digital Libraries, pp. 76–88. Springer (2023)