Proceedings of the 3rd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/150/2025.22321

Multi-rule Graph Construction Network for Entity Linking in Visually Rich Documents

Yinhuan Zheng[1], Jiabao Chen[1], Weiru Zhang[2*]

[1]*School of Mechanical and Electronic Engineering, Wuhan University of Technology, Wuhan, China*
[2]*Birmingham Institute of Fashion and Creative Art, Wuhan Textile University (NANHU CAMPUS), Wuhan, China*
**Corresponding Author. Email: wrzhang@wtu.edu.cn*

Abstract: Entity linking in visually rich documents (VRDs) is critical for industrial automation but faces challenges from complex layouts and computational inefficiency in existing models. Traditional approaches relying on pre-trained transformers or graph networks struggle with noisy OCR outputs, a high number of parameters, and invalid edge predictions in industrial VRDs. To Vision Information Extraction (VIE) for Industrial VRDs, we introduce a lightweight multi-rule graph construction network for entity linking, integrating text and layout embeddings as graph nodes. A multi-rule filtering method reduces invalid edges using node distance, link interference, and standardization rules inspired by document production logic and reading habit. A node relation enhancement module with Graph Attention Networks (GAT) enhances nodes through multi-rule edges and attention scores, enabling robust reasoning for noisy and complex layouts. Evaluated on FUNSD and SIBR datasets, our model achieves F1 scores of 58.96% and 65.08% with only 18M parameters, outperforming non-pretrained baselines while maintaining deployability for resource-constrained environments.

Keywords: Vision Information Extraction, Entity Linking, Graph Construction, Relation Extraction

1. Introduction

Entity linking or entity relation extraction in visually rich documents (VRDs) has evolved from template-based approaches to sophisticated deep learning frameworks, driven by the increasing demand for automated information extraction in industrial applications. Entity linking is a downstream task for Vision Information Extraction (VIE) that generally predicts links between any two semantic entities considering both multimodal information.

Early methods for VIE relied on rule-based systems [1] to identify key-value pairs in structured documents, but these methods struggled with layout diversity and noise. With the development of deep learning-based approaches, there have been many studies on document analysis. Especially as the success of BERT [2] in Natural Language Processing (NLP), the pre-trained transformer models have become the focus of modern research and achieved state-of-the-art (SOTA) results. LayoutLM [3] pioneered the integration of text and layout features, achieving SOTA results on semantic entity recognition (SER) tasks. Subsequent models like LayoutLMv2 [4] and LayoutXLM [5] incorporated

DOI: 10.54254/2755-2721/150/2025.22321

visual information and multilingual capabilities, further advancing SER performance. Although they achieved these achievements, relation extraction (RE) remained underdeveloped, often treated as a secondary task with simplistic classifiers.

The models based on graph neural network also have achieved good performance besides these pre-trained models. Liu [6] utilized graph neural networks (GNNs) that enabled contextual modeling of entities and their spatial relationships, improving performance on semi-structured data. However, related researches [7, 8] faced challenges in handling large-scale datasets and complex document layouts.

A large number of VRDs are generated during the factory production process, making information extraction from VRDs, particularly entity linking, an essential task in industrial informatization. These industrial VRDs impose unique challenges. Industrial documents frequently contain nested tables, overlapping entities, and multimodal elements, requiring robust geometric reasoning and multimodal fusion. As shown in Figure 1, we used PaddleOCR and LayoutLMv2 to perform entity linking on industrial visually rich documents. The complex layout resulted in many errors in linking, including violating reading and table making rules to link two entities at an oblique angle, crossing multiple entity regions to link entities, and linking entities that are relatively far away in the document. Besides, the recent models based on pre-trained Transformer like LayoutLMv3 [9] and GeoLayoutLM [10] lack lightweight architectures, making them unsuitable for edge deployment. Additionally, these models often fail to capture dependencies between adjacent entities linked to the same key, leading to errors in complex layouts [11].

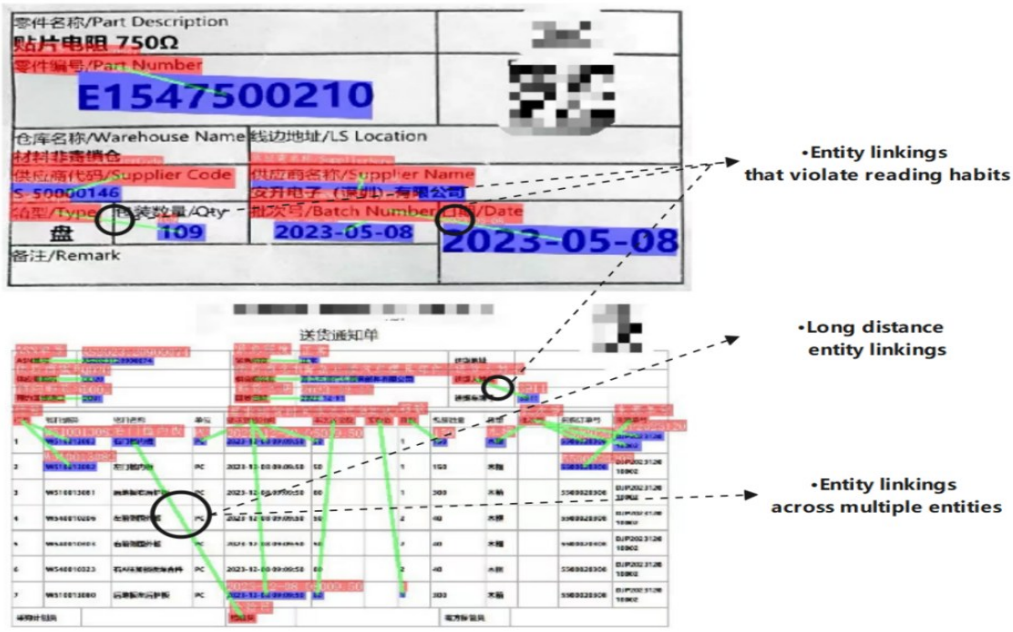

Figure 1: Some errors of entity linking in existing models for visually rich documents

To address this, we propose a lightweight multi-rule graph construction network for entity linking after entity label (or semantic entity recognition) in VIE. This paper can be summarized as follows:

1) A graph construction approach using multi-rule filtering method to reduce weight of invalid links in complex layouts of documents like nested tables. The multi-rule filtering method is based on document production logic and reading habits, including node distance, link interference, and link standardization.

2) A node relation enhancement module that based on Graph Attention Networks (GAT) and multi-rule edge feature to enhance the relation of relevant nodes. Specifically, multi-rule edge features

DOI: 10.54254/2755-2721/150/2025.22321

are fed into a multi-head GAT. Each rule edge independently balances the correlation scores between nodes in each head to obtain rule-influenced attention scores. While extracting the semantics and layout between nodes through attention, rules are used to reduce attention from invalid links.

3) We report competitive results on the public datasets FUNSD and SIBR, with F1 scores of 58.96% and 65.08%. Respectively, while our model is lightweight.

2. Related work

Early visual information extraction (VIE) systems relied on template matching and handcrafted features [12] to parse structured documents. These methods are interpretable, but labor-intensive and lack generalization to unstructured layouts.

The development of Transformers [13] revolutionized VIE. LayoutLM combined text and layout features, achieving SOTA in semantic entity recognition (SER). Subsequent related works [4, 9, 5, 14] have continuously advanced SER tasks and achieved excellent entity extraction capabilities. However, these models treated RE as a secondary task, often using simple bilinear layers that underperformed on complex relationships.

GeolayoutLM propose a relation heads that pre-training by a geometric tasks and fine-tuned for RE. The relation heads enhance spatial reasoning, boosting RE F1 scores on FUNSD from 80.35% to 89.45%. Yang [15] proposed a unified framework for labeling and linking entities by a Entities-As-Points (EAP) module. EntityLayout [16] proposed an entity-level representation by fusing token-level features after pre-training the Transformer Encoder, and utilizes Graph Pruning Module (GPM) to effectively prune invalid links. Although these works achieve excellent performance in entity Relation Extraction or Eneity Linking task, they are still pre-trained models that have high computational costs and lack light-weight design.

LiLT [17] proposed shared parameters for language-agnostic layout modeling, reducing size while maintaining performance. However, it still have high costs for pre-training. GraphRevisedIE [18] used dynamic graph revision without heavy pre-training, but it underperformed on complex documents layouts in Entity Linking task. Although DAMGCN [11] modeled Entity linking using graph pruning and attention mechanisms, it still a Bert-based model like [19-21] and have numerous parameter in model.

3. Methodology

We propose a multi-rule graph construction network for entity linking in Vision Information Extraction (VIE). It is designed to address the challenges of relationship extraction from Visually rich documents (VRDs). Like many previous works of information extraction in VRDs, our entity linking is the downstream task of OCR. It means that text segment contents, bounding boxes, and the corresponding semantic entity classes in the document, are already available. The framework as shown in Figure 2, it consists of two key components: Multimodal graph construction, and graph relation enhancement.

3.1. Multimodal Graph Construction

In our model, the graph structure is determined from entities in terms of spatial and layout information, where the semantic entities are represented as graph nodes, and the potential directional relations between pairs of nodes are represented as graph edges.

DOI: 10.54254/2755-2721/150/2025.22321

3.1.1. Multimodal Node Representation

Textual content is the primary feature of semantic information in VRDs. The text segments are converted to Character-level embedding to capture semantics. Unlike those methods that use BERT to extract text semantics, our approach is more lightweight. Meanwhile, the Character-level embedding is crucial for low-resource languages or noisy OCR outputs.

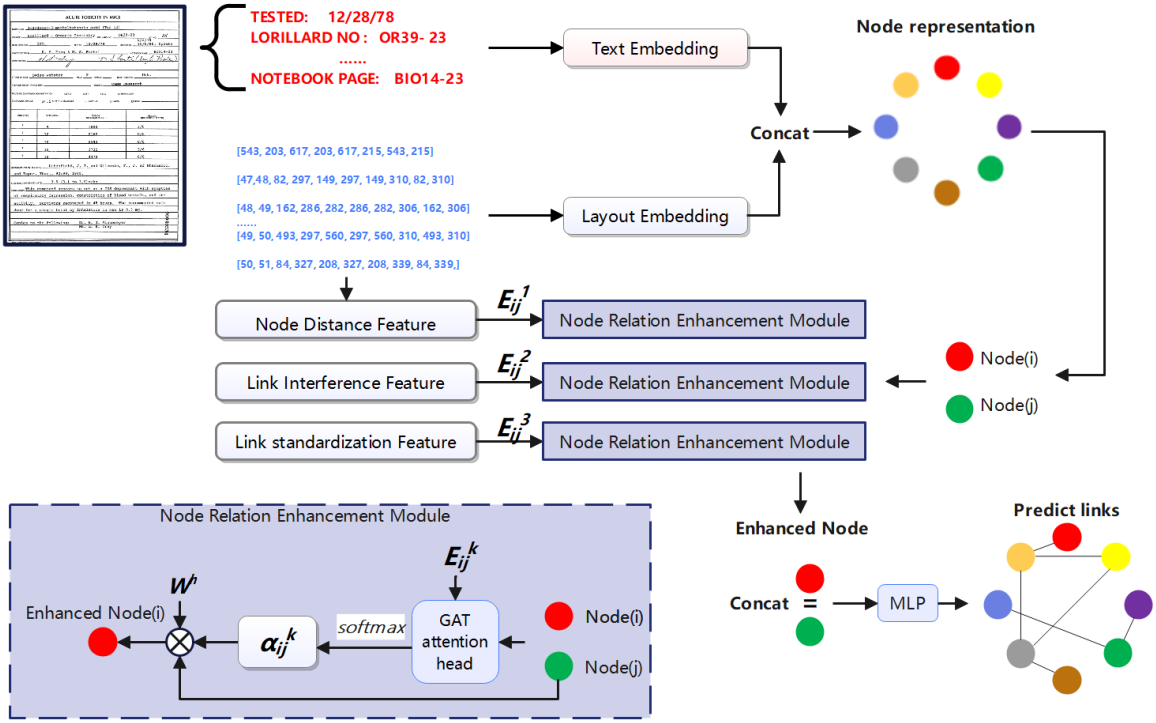

Figure 2: Proposed model architecture.

Textual content is the primary feature of semantic information in VRDs. The text segments are converted to Character-level embedding to capture semantics. Unlike those methods that use BERT to extract text semantics, our approach is more lightweight. Meanwhile, the Character-level embedding is crucial for low-resource languages or noisy OCR outputs.

Given a text segment $S^i = (c_1^i, c_2^i, c_3^i, \ldots, c_j^i)$, it is convert to a number vector based on the vocabulary, and then embedded to the character-level embedding e_t^i :

$$e_t^i = Concat\left(embed(c_1^i) + \cdots + embed(c_j^i)\right) \in \mathbb{R}^{L \times \frac{D}{2}} \quad (1)$$

$$E_{text} = Concat(e_t^1, e_t^2, e_t^3, \ldots, e_t^{i-1}, e_t^i,) \in \mathbb{R}^{N \times L \times \frac{D}{2}} \quad (2)$$

where embed c_j^i is the jth character embedding of the ith text segment. E_{text} concatenate every text segment embedding to a whole embedding. N is the number of segments, L is the length of segment, D is the embedding dimensionality.

In addition to the text modality, the rich layout information in documents is the key feature for visual information extraction, especially entity linking. Layout features encode spatial relationships between text segments in documents critical for entity linking. For text segment i with bounding box center $(\bar{x}_i, \bar{y}_i)$ and size (w_i, h_i), calculate layout features:

DOI: 10.54254/2755-2721/150/2025.22321

$$E_{layout} = Concat\left(f^{\text{sinu}}\left(\text{nor}(\bar{x}_i)\,, \text{nor}(\,\bar{y}_i)\,,\ \frac{w_i}{w}\ ,\frac{h_i}{h}\right)\right)\cdot W_p\ \in\ \mathbb{R}^{N\times\frac{D}{2}} \tag{3}$$

where $\text{nor}(\bar{x}_i)\,, \text{nor}(\,\bar{y}_i)$ is normalize coordinates for $(\bar{x}_i, \bar{y}_i)$ as BROS [19]. $f^{\text{sinu}} \in \mathbb{R} \to \mathbb{R}^{d_k}$ is the sinusoidal embedding. $W_P \in \mathbb{R}^{d_k} \to \mathbb{R}^{\frac{D}{2}}$ is the linear projection matrix that maps from the sinusoidal embedding dimension to the model dimension. Since tokens in the same segment share the same bounding box coordinates, we resize and get the final layout embedding $E_{layout} \in \mathbb{R}^{N\times L\times\frac{D}{2}}$.

The feature of each text segment in the documents is represented as Node(i), which is determined by concatenating the text embedding and layout embedding:

$$N_i = Concat\left(E_{text}, E_{layout}\right) \in \mathbb{R}^{N\times L\times D} \tag{4}$$

3.1.2. Multi-rule Edge Feature

For constructing the graph structure, we propose a multi-rule filtering method to build the graph edges considering as three components.

Node Distance: For key node $\boldsymbol{N_i}$, its corresponding value node $\boldsymbol{N_j}$ is an entity that is close in distance within the document image space, rather than a distant entity. The node distance feature $f_{\text{dist}}(i,j)$ is therefore defined as:

$$f_{\text{dist}}(i,j) = 1 - \frac{d_{ij}}{d_{max}} \in \mathbb{R}^{\text{N}\times\text{N}} \tag{5}$$

where d_{ij} is Euclidean distance between box center $(\bar{x}_i, \bar{y}_i)$ and center $\left(\bar{x}_j, \bar{y}_j\right)$. d_{max} is Diagonal of document.

Link Interference: For linking between $\boldsymbol{N_i}$ and $\boldsymbol{N_j}$ cross the area of other text segment, the likelihood of them being key-value pairs will decrease. The linking interference conforms to document reading habits and the logic of document production, this $f_{inter}(i,j)$ is defined as:

$$f_{inter}(i,j) = \begin{cases} 0, & \text{if } n_{ij} \geq \mu \\ e^{-\lambda n_{ij}}, & \text{otherwise} \end{cases} \in \mathbb{R}^{\text{N}\times\text{N}} \tag{6}$$

where λ is Decay rate controlling interference penalty. n_{ij} is the number of other text segment cross by linking. μ is the threshold which is set as 10.

Link standardization: The VRDs structure used in the factory is more standardized, and in order to better read the information in the document, the linking of key-value pairs is often horizontal or vertical. We design a link standardization feature $f_{standar}(i,j)$:

$$\theta_{ij} = \arctan\left(\frac{|\bar{y}_j - \bar{y}_i|}{|\bar{x}_j - \bar{x}_i| + \epsilon}\right) \tag{7}$$

$$\text{f}_{\text{standar}}(\text{i},\text{j}) = \exp\left(-\frac{\Delta\theta_{\text{ij}}{}^2}{2\sigma^2}\right) \in \mathbb{R}^{\text{N}\times\text{N}} \tag{8}$$

where $\epsilon = 1\text{e} - 5$ avoids division by zero, $\Delta\theta_{ij} = \frac{\min(\theta_{ij}, 90° - \theta_{ij})}{45°}$, and σ is Hyper parameter controlling the spread of the Gaussian kernel.

Combine these three linking features, $\text{E}_\text{r}(i,j)$ is the multi-rule linking features:

$$\text{E}_\text{r}(i,j)\ = [f_{\text{dist}}(i,j)\,,\ f_{inter}(i,j),\ f_{\text{standar}}(i,j)] \tag{9}$$

where $E_r^k(i,j)$ is the $k-th$ linking feature in $\text{E}_\text{r}(i,j), k \in \{1,2,3\}$ is the number of rules.

Furthermore, we utilize Entity Class and Node distance Filtering to filter edges between same-type entities and only preserve multi-rule linking feature under K-Nearest Neighbors topk node distance:

DOI: 10.54254/2755-2721/150/2025.22321

$$E_r^k(i,j) = \begin{cases} 0 & \text{if } c_i = c_j \text{ or } d_{ij} \text{ is not topk,} \\ E_r^k(i,j) & \text{otherwise,} \end{cases} \tag{10}$$

where c_i denotes the entity type of the N_i, d_{ij} is Euclidean distance between N_i and N_j.

Follow the above multimodal node entities and multi-rule linking features, the document graph is described as G = (N, E), where N = Node(i), i ∈ [1, N] and E = $E_r(i,j)$ is edge for represents the relationship between each two segment nodes.

As shown in Figure 3, '***sample No.***' is a known key entity and there are several different value entities '***6030***', '***85 mm Filter***', and '***Dr.A.W.Spears***' around in document. Through the multi-rule filtering method, the link standardization score of ***'85 mm Filter'*** in the document graph is lower, resulting in a low weight of edges. The link Interference and node distance scores of ***node*** '***Dr.A.W.Spears***' reduces the possibility of links between '***sample No.***' and '***Dr.A.W.Spears***'.

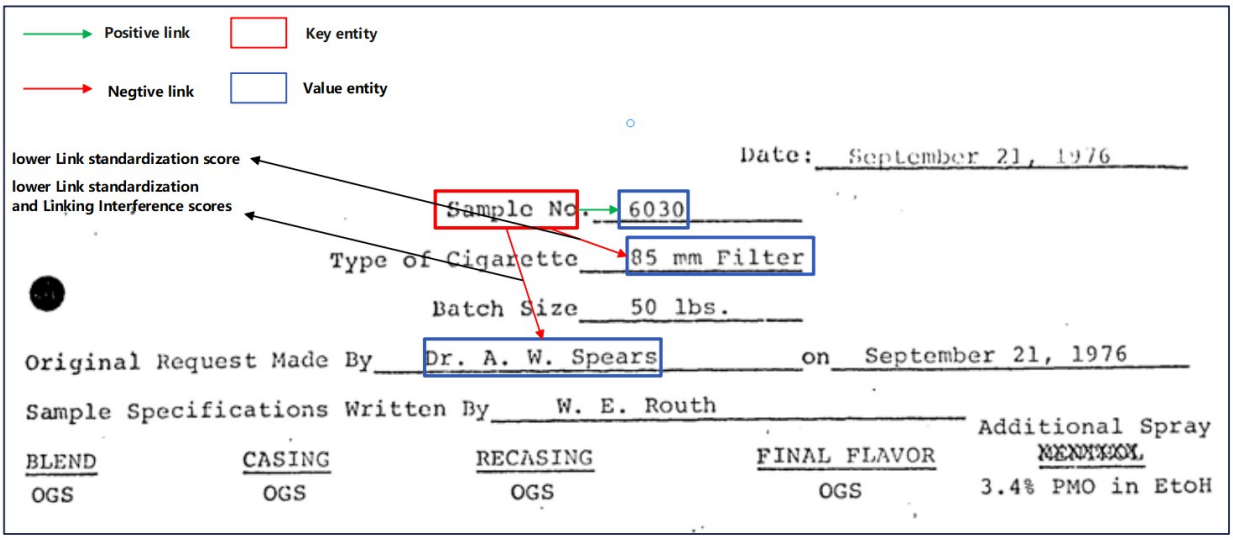

Figure 3: An example demonstrating the multi-rule filtering method to reduce the likelihood of edges that do not conform to reading habits and logic.

3.2. Node Relation Enhancement Module

Although we have constructed edge features between nodes using multi-rule filtering method, the rule relationships of these edges are more based on the construction logic and reading habits of the document. Then these structured information are not comprehensive enough. In the case where the document is not structured enough, it is necessary to determine the link correlation relationship between nodes through text segment semantic information.

In section 3.1.1, we have constructed the multimodal text-layout nodes in document. To dynamically refine the multi-rule edge features through semantic and geometric interactions, inspired by works [22, 23], we incorporate Graph Attention Networks (GAT) [24] for node feature enhancement, which better captures semantic dependencies. Unlike conventional graph neural networks that treat edges as static connections or rely on attention-only mechanisms, our module explicitly models the interplay between predefined rules and textual semantics, enabling robust reasoning for noisy or complex layouts.

Given the multimodal node features $N_i \in \mathbb{R}^{N\times L\times D}$ and multi-rule edge feature $E_r^k(i,j) \in \mathbb{R}^{N\times N\times 3}$, we first aggregate the character-level nodel embeddings in the segments to produce the segment-level node embedding $N_i \in \mathbb{R}^{N\times D}$.Furthermore, we utilize a multi-layer GAT to learn semantic layout

DOI: 10.54254/2755-2721/150/2025.22321

dependencies between nodes, and the last layer is designed with a multi-head GAT to compute each attention scores $\alpha_{ij}^{k} \in \mathbb{R}^{N\times N}$ for each multi-rule edge $E_r^k(i,j)$.

$$e_{i,j}^{k} = LeakyReLU\left((\mathbf{a}^{k})^{\mathbf{T}}\left[W^{k}h_i^{(l)}||W^{k}h_j^{(l)}\right]\right) + \beta^{k}\cdot E_r^k(i,j) \tag{11}$$

$$\alpha_{ij}^{k} = \frac{\exp\left(e_{i,j}^{k}\right)}{\sum_{i,j\in\mathcal{N}_i}\exp\left(e_{i,j}^{k}\right)} \tag{12}$$

where W^k are learnable weight matrix for head k, β^k are the learnable weight for $E_r^k(i,j), k \in \{1,2,3\}$ is the number of rules. $\mathbf{a}^k$ is attention parameter vector. $||$ is concatenation operator. *LeakyReLU* is the activation function. This score $\alpha_{ij}^{k} \in [0,1]$ reflects the node correlation between nodes N_i and N_j based on their text and layout embeddings under multi-rules.

In order to integrate the multi-rule information of edge features in document into the text-layout information of nodes through attention scores α_{ij}, the enhanced node features N_i^h are:

$$N_i^h = \sigma\left(\sum_{k=1}^{3}(\gamma^k\cdot\sum_{j\in N_i}\alpha_{ij}^{k}\cdot W^k\cdot N_j)\right) \tag{13}$$

where W^k is a learnable weight matrix. And γ^k is the learnable weight of each attention head after aggregation. σ is activate function.

Finally, the edge features and enhanced nodes are jointly used for link prediction, and apply binary cross-entropy loss over all candidate edges:

$$P_{ij} = \sigma\left(MLP\left(N_i^h||N_j^h\right)\right) \tag{14}$$

$$\mathscr{L}_{\text{link}} = -\sum_{(i,j)\in\mathcal{E}}\left[y_{ij}\log P_{ij} + (1-y_{ij})\log(1-P_{ij})\right] \tag{15}$$

where the mean value of positive samples and negative samples is calculated separately to offset the gradient deviation. $||$ is concatenation operator. σ is activation function. $\mathcal{E}$ denotes all candidate edges, and $y_{ij} \in \{0,1\}$ is the ground-truth matrix indicating whether a link exists between N_i and N_j.

4. Experiments

In this experiment section, we aim to evaluate the performance of our proposed method against existing approaches. The evaluation of the results was conducted through a series of experiments on a diverse set of datasets, using the F1 score as a metric to demonstrate the effectiveness of our method.

4.1. Dataset Description

Our model is evaluated on multiple real-world public datasets FUNSD [25] and SIBR [15]. FUNSD is one of the most well-known form-based datasets which consists of 199 forms annotated with 4 entity types: question, answer, header, and other. The training set has 149 forms, and the test set has 50 forms.It not only supports entity extraction task for Semantic Entity Recognition task, but also support entity linking task for Relation Extraction task.

SIBR is a structurally rich dataset that represents complexity in real-world scenarios. It contains 1000 documents, including 600 Chinese invoices, 300 English bills of entry, and 100 bilingual receipts. The dataset is divided into a training set of 600 and a test set of 400. The entity labels are categorized in the same way as in FUNSD: question, answer, header, and others. However, the granularity of text boxes differs from that of FUNSD.

DOI: 10.54254/2755-2721/150/2025.22321

4.2. Implementation Details

Our model is implemented in an environment using PyTorch 1.8.0, an NVIDIA GTX 3060 GPU with 12GB memory, and Windows system. It is trained for 100 epochs and checkpoints where saved every 10 epochs using an Adam optimizer with a decaying learning rate. The learning rate is initially set to 1e-4. For datasets where validation is available, Early Stopping is utilized to save the best checkpoints. For set of the model, in the embedding module, the embedding dimension of text embedding and layout embedding is 512, the embedding dimension D of multimodal embedding is 1024. The sinusoidal embedding dimension d_k is 1024. We use dropout with a ratio of 0.1 on NREM. The k for K-Nearest Neighbors is 80 in FUNSD as work [11].

4.3. Experiment Results

Table 1: Performance of entity linking task on the FUNSD dataset.

Model	Pre-trained	Params	F1
LayoutLM [3]	yes	343 M	42.81
StrucTexT [21]	yes	107 M	44.10
BROS [19]	yes	138 M	66.96
LayoutLMv2 [4]	yes	426 M	70.57
LayoutLMv3 [9]	yes	368 M	80.35
GeoLayoutLM [10]	yes	399 M	89.45
DAMGCN [11]	no	250 M+	80.63
GeoContrastNet [23] + GT	no	14 M	32.45
Doc2Graph [26]+ GT	no	6.2 M	53.36
ours	no	18 M	58.96

Table 1 shows the entity linking performance with comparison with other methods. It is shown that our proposed multi-rule graph construction network achieve 58.96% F1 for entity linking task on FUNSD dataset. It is a relatively great performance in models without pre-trained. Compared to other lightweight model such as GeoContrastNet [23] and Doc2Graph [26], our proposed model achieve higher performance in an approximate number of parameters. However, it should be noted that these two models focus on entity annotation and entity linking tasks, and GeoContrastNet does not use text modality, while our model does not use image modality

As shown in Table 2, we then performed our model on the SIBR dataset to evaluate the performance on Multilingual dataset. The SIBR dataset contains text content in both Chinese and English. When ESP [15] processes multilingual datasets, utilize an entity-image text matching (EITM) task, aligns vision-language embedding F1 with text embeddings given by BERT by a contrastive learning strategy, achieve a 85.96% F1. Although our model only achieved 65.08% F1, our method has fewer parameters and does not require processing of image modalities.

Table 2: Performance of entity linking task on the SIBR dataset, where T:Text, V:Visual, L:Layout.

Model	Pre-trained	Modalities	Params	F1
LayoutXLM [12]	yes	T + V + L		83.99
ESP [15]	no	T + V + L	50 M	85.96
Our	no	T + L	18 M	65.08

The ablation study on the FUNSD dataset evaluate the importance of multi-rule filtering method are shown in Table 3. The node distance feature, especially by minimizing the occurrence of invalid

DOI: 10.54254/2755-2721/150/2025.22321

edges through K-Nearest Neighbors (KNN), is crucial for eliminating invalid links between nodes in document. Its F1 Score decreases by approximately 4.02%. Among these three edge linking features, Link standardization feature has the smallest impact, but there are also 0.97%.

Table 3: Evaluation ablation study on FUNSD.

Components of multi-rule filtering method	F1 Score
Full model	58.96
Node Distance	↓ 4.02
Link Interference	↓ 3.28
Link standardization	↓ 0.97

5. Conclusion

In this paper, we present a lightweight multi-rule graph construction network for entity linking in Vision Information Extraction. This lightweight model is suitable for environments with insufficient computing resources and limited deployment conditions, such as factories that require local deployment. Specifically, We introduce a multimodal graph construction module model text and layout information as node in document graph, and propose a multi-rule filtering method to reduce the occurrence of invalid edges. Furthermore, we utilize a node relation enhancement module to extract linking information. In addition, we validated the performance of our model on the FUNSD and SIBR datasets. Although we did not achieve the performance of the state-of-the-art models, our model also has the advantage that is lightweight.

References

[1] *Jung K, Kim KI, Jain AK. Text information extraction in images and video: A survey[J]. Pattern Recognition, 2004, 37(5):977-997. DOI:10.1016/j.patcog.2003.10.012.*

[2] *Devlin J, Chang W, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint: arXiv:1810.04805 (2018).*

[3] *Xu Y, Li M, Cui L, et al. LayoutLM: Pre-training of text and layout for document image understanding[C]//KDD ' 20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. ACM, 2020. DOI:10.1145/3394486.3403172.*

[4] *Xu Y, Xu Y, Lv T, et al. LayoutLMv2: Multi-modal pre-training for visually-rich document understanding[J]. arXiv preprint: arXiv:2012.14740 (2020).*

[5] *Xu Y, Lv T, Cui L, et al. LayoutXLM: Multimodal pre-training for multilingual visually-rich document understanding[C]//Proceedings of the 2021 Conference on Neural Information Processing Systems (NeurIPS). 2021. DOI:10.48550/arXiv.2104.08836.*

[6] *Liu X, Gao F, Zhang Q, et al. Graph convolution for multimodal information extraction from visually rich documents[C]//Proceedings of the 2019 Conference on North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). 2019. DOI:10.18653/v1/N19-2005.*

[7] *Tang G, Xie L, Jin L, et al. MatchVIE: Exploiting match relevancy between entities for visual information extraction[C]//IJCAI-21: The 30th International Joint Conference on Artificial Intelligence. 2021. DOI:10.24963/ijcai.2021/144.*

[8] *Yu W, Lu N, Qi X, et al. PICK: Processing key information extraction from documents using improved graph learning-convolutional networks[C]//2021 International Conference on Pattern Recognition (ICPR). IEEE, 2021. DOI:10.1109/ICPR48806.2021.9412927.*

[9] *Huang Y, Lv T, Cui L, et al. LayoutLMv3: Pre-training for document AI with unified text and image masking[J]. arXiv preprint: arXiv:2204.08387 (2022).*

[10] *Luo CW, Cheng CX, Zheng Q, Yao C. GeoLayoutLM: Geometric pre-training for visual information extraction[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023:7092-7101. DOI:10.1109/CVPR52729.2023.00685.*

[11] *Chen YM, Hou XT, Lou DF, et al. DAMGCN: Entity linking in visually rich documents with dependency-aware multimodal graph convolutional network[C]//Document Analysis and Recognition - ICDAR 2023, Part III. Lecture Notes in Computer Science, vol 14189. Springer, Cham, 2023:33-47. DOI:10.1007/978-3-031-41682-8_3.*

DOI: 10.54254/2755-2721/150/2025.22321

[12] Schuster D, Muthmann K, Esser D, et al. Intellix - End-User trained information extraction for document archiving[C]//International Conference on Document Analysis and Recognition (ICDAR). IEEE Computer Society, 2013. DOI:10.1109/ICDAR.2013.28.
[13] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. arXiv preprint: arXiv:1706.03762 (2017).
[14] Li P, Gu J, Kuen J, et al. SelfDoc: Self-supervised document representation learning[J]. arXiv preprint: arXiv:2106.03331 (2021).
[15] Yang ZB, Long RJ, Wang PF, et al. Modeling entities as semantic points for visual information extraction in the wild[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2023. DOI:10.1109/CVPR52729.2023.01474.
[16] Xu CB, Chen YM, Liu CL. EntityLayout: Entity-level pre-training language model for semantic entity recognition and relation extraction[C]//Document Analysis and Recognition - ICDAR 2024, Part I. Lecture Notes in Computer Science, vol 14804. Springer, Cham, 2024:262-279. DOI:10.1007/978-3-031-70533-5_16.
[17] Wang J, Jin L, Ding K. LiLT: A simple yet effective language-independent layout transformer for structured document understanding[J]. arXiv preprint: arXiv:2202.13669 (2022).
[18] Cao P, Wu J. GraphRevisedIE: Multimodal information extraction with graph-revised network[C]//International Conference on Multimodal Interaction (ICMI). 2024.
[19] Hong T, Kim D, Ji M, et al. BROS: a pre-trained language model focusing on text and layout for better key information extraction from documents[C]//Proceedings of the AAAI Conference on Artificial Intelligence, vol 36. AAAI Press, 2022:10767-10775.
[20] Chi Z, et al. InfoXLM: an information-theoretic framework for cross-lingual language pre-training[J]. arXiv preprint: arXiv:2007.07834 (2020).
[21] Li Y, et al. StrucText: structured text understanding with multi-modal transformers[C]//29th ACM International Conference on Multimedia (MM). ACM, 2021:1912–1920.
[22] Chen G, Chen P, Wang Q, et al. EMGE: Entities and mentions gradual enhancement with semantics and connection modelling for document-level relation extraction[J]. Knowledge-Based Systems, 2025, 309. DOI:10.1016/j.knosys.2024.112777.
[23] Biescas N, Boned C, Lladós J, et al. GeoContrastNet: Contrastive key-value edge learning for language-agnostic document understanding[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2024.
[24] Veličković P, Cucurull G, Casanova A, et al. Graph attention networks[J]. arXiv preprint: arXiv:1710.10903 (2017).
[25] Jaume G, Ekenel HK, Thiran JP. FUNSD: A dataset for form understanding in noisy scanned documents[C]//2019 International Conference on Document Analysis and Recognition Workshops (ICDARW). IEEE, 2019. DOI:10.1109/ICDARW.2019.10029.