P²Net: Parallel Pointer-based Network for Key Information Extraction with Complex Layouts

Kaiwen Wei[1], **Jie Yao**[2], **Jiang Zhong**[1*], **Yangyang Kang**[3*],
Jingyuan Zhang[4], **Changlong Sun**[5], **Xin Zhang**[5], **Fengmao Lv**[6], **Li Jin**[7]
[1]College of Computer Science, Chongqing University, China
[2]Ant Group, [3]Zhejiang University, [4]Kuaishou Technology, [5]Alibaba Group,
[6]School of Computing and Artificial Intelligence, Southwest Jiaotong University,
[7]Aerospace Information Research Institute, Chinese Academy of Sciences
{weikaiwen, zhongjiang}@cqu.edu.cn, yangyangkang@zju.edu.cn

Abstract

Key Information Extraction (KIE) is a challenging multimodal task aimed at extracting structured value entities from visually rich documents. Despite recent advancements, two major challenges remain. First, existing datasets typically feature fixed layouts and a limited set of entity categories, while current methods are based on a full-shot setting that is difficult to apply in real-world scenarios, where new entity categories frequently emerge. Secondly, current methods often treat key entities simply as parts of the OCR-parsed context, neglecting the positive impact of the relationships between key-value entities. To address the first challenge, we introduce a new large-scale, human-annotated dataset, Complex Layout document for Key Information Extraction (CLEX). Comprising 5,860 images with 1,162 entity categories, CLEX is larger and more complex than existing datasets. It also primarily focuses on the zero-shot and few-shot KIE tasks, which are more aligned with real-world applications. To tackle the second challenge, we propose the Parallel Pointer-based Network (P²Net). This model frames KIE as a pointer-based classification task and effectively leverages implicit relationships between key-value entities to enhance extraction. Its parallel extraction mechanism enables simultaneous and efficient extraction of multiple results. Experiments on widely-used datasets, including SROIE, CORD, and the newly introduced CLEX, demonstrate that P²Net outperforms existing state-of-the-art methods (including GPT-4V) while maintaining fast inference speeds.

1 Introduction

Key Information Extraction (KIE) aims to extract value entities (i.e., the text contents with their layout coordinates) from visually rich documents in the semi-structured document (Cao et al., 2022), such as forms, or scanned documents. For instance,

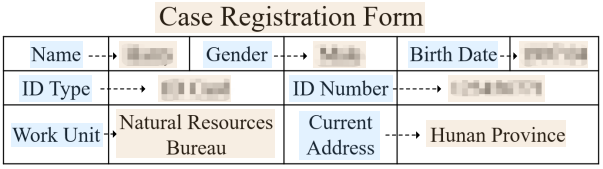

Figure 1: An example of the KIE. The key and value entities are marked in blue and orange boxes, respectively. The dotted arrows indicate implicit relations between key-value pairs. The content of the table is synthesized and some values are decorated for privacy purposes.

as shown in Fig. 1, given an input visually rich document and a query "organization", KIE aims to extract the value entity "Natural Resources Bureau". KIE plays a critical role in many downstream tasks, such as document information registration (Majumder et al., 2020) and document understanding (Cheng et al., 2020; Hamdi et al., 2021).

Many tasks and datasets (Jaume et al., 2019; Xu et al., 2021b; Guo et al., 2019; Huang et al., 2019; Park et al., 2019; Wang et al., 2021a; Kuang et al., 2023) have been proposed for KIE across various domains. In addition, a series of pre-trained multi-modal models (Huang et al., 2022; Xu et al., 2021b; Peng et al., 2022) were presented. Based on these models, some methods (Zhang et al., 2021; Hu et al., 2023; Gao et al., 2022; Kim et al., 2022; Hwang et al., 2021; Zhang et al., 2023) leveraged biaffine model, Question Answering (QA), generation, graph-based, or token path prediction methods to extract the value entities. Recently, many multimodal large language models (MLLMs) (Hu et al., 2024; Li et al., 2023a; OpenAI, 2023) were proposed, contributing to the KIE task. Despite the promising results, those datasets and methods still suffer from the following two challenges:

(1) **From the dataset and task setting view**: Existing KIE datasets are constrained by relatively fixed layouts and a limited number of entity categories. Statistically, the dataset with the most

*Corresponding Author

Findings of the Association for Computational Linguistics: ACL 2025, pages 10611–10626

categories CORD (Park et al., 2019) includes only 30 entity categories, which is insufficient to meet the needs of real-world applications. Furthermore, many real-world documents feature complex layout styles, and new entity categories continuously emerge. However, it is challenging to collect sufficient samples for these new entities. Existing KIE methods are typically designed for full-shot settings, which fail to address the dynamic and evolving nature of practical scenarios.

(2) **From the method and information utilization view**: To extract value entities, existing methods typically combine text parsed by optical character recognition (OCR) with visual features and feed them into pre-trained multi-modal models, treating key entities as mere contextual components. However, they often neglect the crucial role of implicit associations between key and value entities. For instance, as shown in Fig. 1, once the key entity "Work Unit" is identified, it becomes easier to extract the value entity "Natural Resources Bureau" due to the implicit relationships between them. These associations are not only semantic but also related to their positional coordinates.

To address the challenge of existing datasets in terms of layout variety and entity categories, we propose a new large-scale human-annotated dataset named **C**omplex **L**ayout document for key information **EX**traction (CLEX), which includes 5,860 scanned document images with 155 document types and 1,162 entity categories. In order to meet the practical needs and make the model adaptable to the scenario of emerging unseen entities, we mainly consider **zero/few-shot KIE** task. Specifically, this task leverages a set of known category samples, followed by the introduction of zero or a few number (1/5/10) of samples from previously unseen document categories for training the model. The model's performance is then evaluated on these new, unseen document types.

To tackle the second challenge, we propose the **P**arallel **P**ointer-based **Net**work (P^{2}Net). This model frames KIE as a pointer-based classification task, where key and value entities can point to each other. By doing so, it effectively captures the implicit relationships between key entities and key-value pairs, facilitating more accurate extraction of value entities. Additionally, P²Net incorporates multiple questions into the input context and utilizes token-linking operations to generate results for each question simultaneously. This parallel extraction mechanism greatly enhances the speed of P²Net, improving both efficiency and performance. We conduct extensive few/zero-shot KIE experiments on the SROIE (Huang et al., 2019), CORD (Park et al., 2019), and the newly proposed CLEX datasets. The results demonstrate that P^{2}Net outperforms existing state-of-the-art methods, and the inference speed increases 6.4 times compared to traditional QA method on CLEX. In summary, the contributions of this paper are:

1) We introduce CLEX, a large-scale human-annotated complex layout dataset for KIE, containing more images and entity categories. Additionally, to meet the requirements in practical situations, we mainly consider few/zero-shot KIE.

2) We propose a pointer-based method P^{2}Net, which parallelly extracts value entities in the KIE task while incorporating implicit clues from key entities and key-value pairs.

3) The experiment results in few/zero-shot KIE scenarios illustrate that P^{2}Net outperforms state-of-the-art methods with a fast inference speed.

2 Related Work

KIE Datasets. There are some datasets released in the KIE research area. For instance, FUNSD (Jaume et al., 2019) contains 199 noisy scanned documents in English with 4 entity categories. After that, XFUND (Xu et al., 2021b) extend it into 7 different languages. Several other datasets, such as MATEN (Guo et al., 2019), SROIE (Huang et al., 2019), CORD (Park et al., 2019), EPHOIE (Wang et al., 2021a), Kleaster NDA and Kleaster Charity (Stanislawek et al., 2021), VRDU-Registration Form and VRDU-Ad-buy Form (Wang et al., 2023), POIE (Kuang et al., 2023) were proposed based on different fields. However, the layouts of existing KIE datasets are relatively fixed, making it hard to be applied to the real scenario where new entities are constantly emerging. To alleviate this problem, we propose the few/zero-shot KIE task, and introduce a new large-scale dataset with complex layouts named CLEX with 5,860 images and 1,162 entity categories. Please note that different from traditional few/zero-shot KIE task (Wang and Shang, 2022; Wang et al., 2023), in this work, the models are first pre-trained on a set of known category samples, and then trained by zero or several unseen category samples. The motivation of this setting is to leverage the migration ability of the models from known category samples to those unknown category samples.

	Source	Image Number	Entity Type	Language
SROIE	Receipts	1000	4	English
CORD	Receipts	1000	30	English
EPHOIE	Paper head	1494	10	Chinese
FUNSD	Forms	199	4	English
XFUND	Forms	199	4	7 Languages
Kleaster NDA	EDGAR	540	4	English
Kleister Charity	UK Charity Commission	2778	8	English
VRDU-Registration Form	Foreign Agents Registration Act	1915	6	English
POIE	Product	3000	21	English
CLEX (ours)	Invoices, Certificates, etc.	5860	1162	Chinese/English

Table 1: The comparison between existing KIE datasets and the proposed CLEX dataset. CLEX is larger in terms of the number of images and offers a richer variety of entity categories, thus better reflecting real-world scenarios.

KIE Methods. With the development of deep learning (Song et al., 2022; Liu et al., 2024b; Wang et al., 2022) and information extraction (Wei et al., 2021; Liu et al., 2024a; Wei et al., 2023), KIE requires the consideration of textual, visual, and layout information. Many pre-trained multi-modal models, such as Layoutlmv3 (Huang et al., 2022), LayoutXLM (Xu et al., 2021b), and ERNIE-Layout (Peng et al., 2022) have been introduced. Based on these pre-trained models, several approaches have achieved promising results, including QA-based (Hu et al., 2023; Gao et al., 2022), generative-based (Cao et al., 2022, 2023), token path prediction (Zhang et al., 2023), graph-based (Hwang et al., 2021; Li et al., 2023b) methods. Recently, multimodal large language models (MLLMs) (Hu et al., 2024; Li et al., 2023a; OpenAI, 2023) were proposed, which perform well on general vision-language tasks. However, these methods do not leverage the implicit information brought by key entities and key-value pairs. Additionally, they fail to achieve a balance between performance and inference speed. To address these issues, we propose the P^2Net model, which parallelly extracts relevant answers based on pointers in an end-to-end manner.

3 CLEX Dataset

In this section, we introduce the data construction process of the proposed **C**omplex **L**ayout document for key information **EX**traction (CLEX) dataset, as well as its statistics.

Data Collection and Annotations. The CLEX dataset contains the following 4 components with 155 categories: 1) 18 categories of invoice data; 2) 32 categories of card certificates; 3) 82 categories of legal bills; 4) 23 categories of enterprise qualifications, software copyrights and patent certificates. All the data were first collected by crawling the Internet to obtain the initial data, and then manually refactored to avoid data privacy and copyright issues. For addressing concerns of information security, our approach involves rigorous data desensitization practices, which are implemented as follows: (1) Mask out all the content in the original crawled forms and only kept the templates of those forms. (2) Based on the collected document templates, we invited annotator who are familiar with real data in related scenarios to reconstruct the content based on some public documents (e.g., public cases of adjudication documents) combined with personal experience, while ensuring a certain logic of the content. (3) Comprehensive manual inspection, where we invited quality inspectors to check whether there are privacy leak issues. After the whole process, since all the content are refactored and only kept the template, there is little risk of leaking private information.

With the obtained refactored documents, we then designed the guidelines for annotating the KIE task, according to the frequency of occurring categories in each kind of document. Specifically, we invited 20 full-time, experienced annotators to label all the data according to our guidelines using a specially designed annotation platform. In the platform, annotators first identified the bounding boxes of key and value entities, and assigned a predefined category to the value entity. Then, they annotated the key-value relationship by using drag and drop. Finally, we invited two quality control personnel to randomly check 30% of the samples for each category. If the quality was found the inter-rater agreement Fleiss's K (Fleiss et al., 1981) less than 0.85, the annotation for that category should be revised until the required standard is met.

Dataset Splits. To simulate real complex layout in KIE, we mainly consider zero/few-shot scenarios. We divide the dataset for training and testing based on the categories of the documents. Specifically,

in the zero-shot scenario, we randomly divide the training and test sets based on the categories at the ratio of 108:47 (nearly 7:3), where the categories of documents existing in test sets do not occur in the training set. In the few-shot scenario, we move a few samples (1/5/10 shot) of each category from the test set to the training set. As a result, a portion of the same category data will be involved in the few-shot training process.

Dataset Analysis and Statistics. After the entire process of annotation, we finally obtained a total of 5,860 images of the forms. Under the zero-shot situation, 3,993 forms are set for training, and the rest 1,867 forms are for testing. Among all the forms, there exist 75,438 value entities with 1,162 entity types, where 52,420 value entities are with key entities and the rest 23,018 are without the key entities. As for the diversity of the CLEX dataset, since we annotate based on the total of 155 categories of data, where each category of the dataset may contain multiple layouts. Therefore, CLEX dataset contains more than 155 kinds of layout for training and inference. The comparison of different datasets is illustrated in Table 1. We also list the top 20 entity types of CLEX in Appendix A.7.

4 Method

4.1 Problem Formulation

The input of KIE is a form-like document image D and the context C containing S textual tokens $C = [c_1, c_2, ..., c_S]$ with their corresponding bounding box coordinates. The coordinates of the i-th bounding box could be denoted as $B_i = [x_i^1, y_i^1, x_i^2, y_i^2]$, where (x_i^1, y_i^1) and (x_i^2, y_i^2) are the top-left and bottom-right corner coordinates. Those text tokens and coordinates could be obtained from the Optical Character Recognition (OCR) tools. The goal of KIE is to extract a list of value entities $V = [V_1, V_n, ..., V_N]$ with their types $l \in L$, where V_n indicates the n-th value entity, and L is the predefined entity label set. Please note that the OCR tool parses line by line, thus the entities may be discontinuous after parsing (e.g., in Fig. 1, after OCR parsing, "Natural Resources" and "Bureau" are not continuous). P^{2}Net formulates KIE as a word-word pointer-based relation classification task, which models the relations between key-value entities and answer many questions in parallel. The framework of P^{2}Net is illustrated in Fig. 2.

4.2 Encoding

Following the encoding procedure in Xu et al. (2021b), the input embedding consists of two parts: the visual token embedding and the text token embedding. The text input embedding is composed of the questions and the input context C, where all the tokens in C are serialized into a 1D sequence from the OCR tools. Specifically, P^{2}Net first concatenates questions $[Q_1, Q_2, ...]$ to the context C and forms the text input X, containing S tokens:

$$X = \langle s \rangle\, Q_1[T]Q_2[T]... \langle /s \rangle\, C, \tag{1}$$

where Q_1, Q_2, ... are different questions. $\langle s \rangle$ and $\langle /s \rangle$ are special tokens from pre-trained models. We leverage a special token $[T]$ to separate the input questions. In the experiment, we utilize the categories of the entities that likely have in the input image D as the questions. If too many questions are concatenated to the context C, due to the maximum length requirement of the backbone (e.g., LayoutXLM), parts of the content would be truncated, resulting in loss of information. To balance the length of the questions and the context, we leverage sliding windows. Specifically, if the length of the concatenated questions and the context exceeds the maximum length of the backbone, we dynamically crop a portion of the questions to make the concatenated input X meets the input length limit, and the rest questions are placed in the next sample.

After being fed into the pre-trained multi-modal model, the input X is mapped into text token embeddings T and each token is assigned a segment embedding. The visual image is first resized into a 224×224 feature map. After a fully connected layer, it is flattened into visual embeddings A. Then the bounding boxes B are mapping to the position embeddings. The token and image embeddings are concatenated together, after summing with segment embeddings and position embeddings, the final feature output H is obtained. The encoding process could be formulated as follows:

$$H = \text{Encoder}(T, A, B) \tag{2}$$

After that, to get the relations between different tokens, we feed the hidden states matrix into two feed forward neural networks $FFN_a(.)$, $FFN_b(.)$. We could get the prediction score matrix $Y \in \mathbb{R}^{S \times S \times P}$:

$$Y_p^{i,j} = FFN_a(H^i)^T FFN_b(H^j), \tag{3}$$

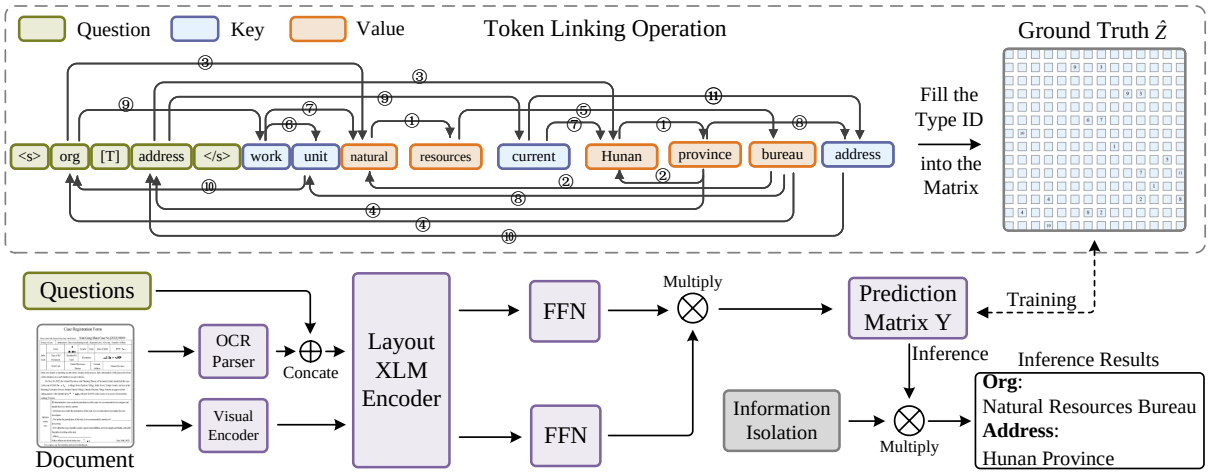

Figure 2: The architecture of P^{2}Net. It mainly has three steps: (1) The questions and context in the document are concatenated and encoded along with the visual features; (2) LayoutXLM output features are fed to two FFNs, forming the matrix Y; (3) Use ground truth $\hat{Z}$ to train the prediction matrix Y, and finally utilize Y for inference with the information isolation mechanism.

where S is the token numbder, P is the pre-defined token linking type, $Y_p^{i,j}$ is the score and the mask value from the i-th token to the j-th token for the p-th token linking type. Besides, to enhance the models' ability to identify entity span, we also add the sinusoidal position embedding (Vaswani et al., 2017) before calculating the score matrix.

4.3 Token Linking Operations

The core of P^{2}Net is how to link the head or tail tokens between questions and those entities in the context. Take the OCR-parsed sentence in Fig. 2 as an example, following the mechanism in Li et al. (2022), if P^{2}Net predicts (org→natural), (natural→resources), (resources→bureau), and (bureau→org), we could get "natural resources bureau" as the "org" type. This operation could be considered as building a directional word graph, which represents the word-word relations. The decoding object is to find certain paths from one word to another word in the graph using the predicted token linking relations.

To construct the directional word graph, we define 5 directed token linkings. Specifically, as depicted in the upper side of Fig.2, we design a relation matrix $\hat{Z} \in \mathbb{R}^{S\times S}$, where different linking types will be assigned with different type ids. $\hat{Z}^{i,j} \neq 0$ means there is a link between the i-th token and the j-th token; otherwise, no link exists. Next, we show how to build the 5 linking types:

Value_Head to Value_Tail Linking. This kind of linking connects continuous parts of the value entities. Assuming one value entity spans from the i-th token to the j-th token, we set $\hat{Z}^{k,k+1} = 1$, where k varies from i to $j-1$. For example, the "Hunan province" in Fig. 2 is the ground-truth value entity, so we let the linking points from "Hunan" to "province" and give this linking a type id as 1.

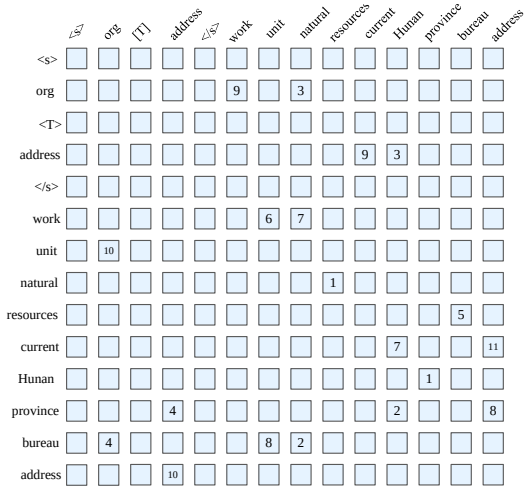

Figure 3: Illustration of the token linking operations to build the ground truth matrix $\hat{Z}$ for training.

Value_Tail to Value_Head Linking. The tail token of each value entity is pointed to the head token of this value entity. For instance, given "Hunan province" as value entity, a connection exists from "province" to "Hunan". This kind of linking is given a type id 2.

Question_Head to Value_Head Linking. The head token of each question is pointed to the head token of the corresponding value entity. For in-

stance, the head token[1] "address" of the question will point to the corresponding value entity head token "Hunan". The linking type is is 3.

Value_Tail to Question_Tail Linking. The tail token of the value entity is pointed to the corresponding tail token of each question. Likewise, the tail token "province" will point to the tail of the question "address". This linking type id is 4.

Discontinuous_Value_Head to Value_Tail Linking. Since there are problems with value entities such as discontinuities or folded rows, we connect different parts of the same entity first and last. As illustrated in Fig. 3, the folded-row problem makes the "natural resources" and "bureau" discontinuous, thus we point from "resources" to "bureau" and give this kind of linking the type id as 5. If the entity is continuous, we directly point the tail of the entity to the tail itself.

Additionally, to further utilize the clues from key entities and key-value pairs, we design additional 6 linking operations (type id ranges from 6 to 11) about the key entities. Due to the page limitation, please see Appendix A.1 for more details about the token linking operations about key entities.

4.4 Information Isolation

In order to filter out those impossible results and further boost the performance of extraction during inference stage, we design two mask methods for information isolation:

Question Context Isolation. Since we concatenate the questions to the OCR-parsed context and the expected entities must come from the context instead of from the questions, we masked out the question part and hope the model to find the answers from the OCR-parsed context. Specifically, if the p-th token type in prediction score matrix $Y \in \mathbb{R}^{S\times S\times P}$ related to the key/value entities (i.e., p=1, 2, 5, 6, 7, 8, 11), we use the question context isolation to mask out improbable predictions of keys/values at question positions.

Question Head/Tail Isolation. Since the order of the input concatenated questions is self-determined, the start/end positions of each question could be known before inputting to model. Therefore, we design the question head/tail isolation to tell the model where the head and tail positions of questions are. For example, when predicting about the question head pointing to key/value head (i.e., type id is 3 or 9), we only keep the predictions start from question head/tails and mask others.

Then we elementally multiply the mask matrix constructed by the information isolation mechanism to the prediction matrix Y, obtaining the score matrix $W \in \mathbb{R}^{S\times S\times P}$. Please refer to Appendix A.2 for more details and examples of the information isolation mechanism.

4.5 Training and Inference

During the training process, we expand the linking type ids in $\hat{Z} \in \mathbb{R}^{S\times S}$ to a new matrix $\hat{Y} \in \mathbb{R}^{S\times S\times P}$. Specifically, if the id in $\hat{Z}$ is p at the i-th row and j-th column, the corresponding position on the $\hat{Y}_p^{i,j}$ is set as 1.

We leverage the constructed matrix $\hat{Y}$ as the the ground-truth and let the predicted score matrix $Y \in \mathbb{R}^{S\times S\times P}$ fit $\hat{Y}$. Next, we flatten the score matrix Y and the ground truth matrix $\hat{Y}$ into a 1D vector, and leverage the circle loss (Sun et al., 2020; Su et al., 2022) to balance the sparse matrices:

$$\mathcal{L} = \log\left(1 + \sum_{\hat{Y}^i=0} e^{Y^i}\right) + \log\left(1 + \sum_{\hat{Y}^j=1} e^{-Y^j}\right), \tag{4}$$

where we flat the ground truth $\hat{Y}$ and score matrix Y, and then calculate the loss function.

During the inference stage, after information isolation, we compare the score matrix $W \in \mathbb{R}^{S\times S\times P}$ with a predefined threshold δ at the last dimension, and the $\tilde{W} \in \mathbb{R}^{S\times S\times P}$ could be obtained, where $\tilde{W}$ contains only 0 or 1:

$$\tilde{W}_p^{i,j} = \begin{cases} 1 & \textit{if } W_p^{i,j} \geq \delta \\ 0 & \textit{otherwise} \end{cases}. \tag{5}$$

After obtaining the directional word graph $\tilde{W}$, we can utilize it to decode the value entities corresponding to different questions based on the token linking operations.

5 Experiment

Experimental Settings. To verify the capability of the models, we conduct experiments on the SROIE (Huang et al., 2019) and CORD (Park et al., 2019), and the proposed CLEX datasets. Since CLEX is a multi-lingual dataset, we adopt *LayoutXLM-base* (Xu et al., 2021b) as backbone. Following previous works (Jaume et al., 2019; Zhang et al., 2021), we take the entity-level precision (P), recall (R), and F1 as the measure standards for value entity extration. Please see Appendix A.3

[1] If a question or entity has only one token, the head and the tail are itself.

zero-shot	P	R	F1
SL	1.62	0.28	0.48
Donut	22.84	21.28	22.03
GPT4V (w/o text)	24.11	21.68	22.83
ChatGPT	56.71	54.33	55.50
GPT4V (w text)	67.28	65.20	66.22
SimpleDLM	76.23	65.47	70.44
QA	77.86	66.72	71.86
WPN	78.94	**68.22**	73.19
P^{2}Net	**81.62**	67.75	**74.04**

Table 2: The zero-shot results on CLEX dataset.

	1-shot	5-shot	10-shot
SL	0.23	5.13	13.94
Donut	35.47	37.59	40.12
SimpleDLM	74.05	77.86	80.79
QA	75.33	78.79	81.39
WPN	76.76	80.73	82.58
PPN	**80.84**	**82.92**	**83.96**

Table 3: The F1 results of few-shot experiments.

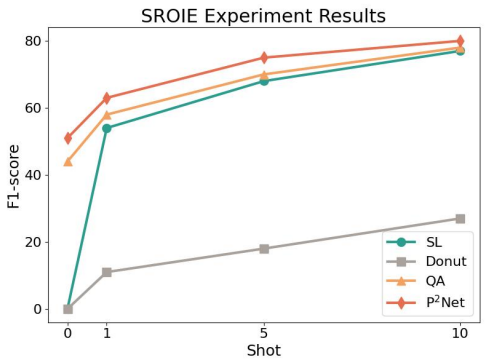

Figure 4: The experiment results on SROIE dataset.

	P	R	F1
P^{2}Net	81.62	67.75	74.04
- *sin*	78.82	68.55	73.33
- *key*	78.94	68.22	73.19
- *QCI*	79.85	68.92	73.98
- *QHI*	79.41	68.60	73.61
- *QTI*	79.20	68.55	73.49

Table 4: The results of ablation study.

for baseline and hyper-parameter details.

Baselines. Since CLEX is a novel dataset without relevant reports, we select the recent state-of-the-art methods conducted on other datasets as the baselines, including (1) **Sequence labelling (SL)** (Xu et al., 2021b); (2) **SimpleDLM** (Gao et al., 2022); (3) **QA** (Hu et al., 2023); (4) **Donut** (Kim et al., 2022); (5) **P^{2}Net**; (6) **WPN**, which is a variant of P^{2}Net that does not consider the 6 linking operations about the key entities; (7) **ChatGPT**[2], version *gpt-3.5-turbo-1106*; (8) GPT4V (OpenAI, 2023), version *gpt-4-1106-vision-preview*. Please refer to Appendix A.5 for the details of baselines. We also compare: **GPT4V (w/o text)**, which direct input with the images and the instructions; **GPT4V (w text)**, which further incorporate the OCR parsed text into the input instructions. Please see details of MLLM experiments in Appendix A.6.

The zero-shot experiment results on CLEX in Table 2 shows: (1) SL underperforms, highlighting challenges in classifying across 1,162 categories and transferring knowledge to unseen data. (2) Generation-based models like ChatGPT and Donut outperform SL but face limitations in KIE. In particular, generation-based methods are not controllable and hallucinations problem (e.g., the generated words do not exist in the original text) sometimes occurs. (3) GPT-4V (w/ text) significantly outperforms GPT-4V (w/o text) by leveraging textual context. A key factor is that GPT-4V (w/o text) relies solely on image inputs, making it prone to OCR errors. (4) WPN and P^{2}Net surpass the QA method by 1.33% and 2.18% in F1 scores, respectively. It shows that the word-word classification mechanism can effectively model the KIE task, and it could transfer knowledge learned from other entity categories to the unseen category.

The results of the 1/5/10-shot experiments on the CLEX dataset are shown in Table 3. It could be observed that: (1) As the number of training data in the same category increases, the experimental results of most models improve. This indicates that the models are able to learn the knowledge from a small number of samples. (2) SL struggles in 1-shot scenarios, indicating difficulty in knowledge transfer from extremely limited data. Until given more samples, the performance of SL has a certain improvement. (3) But P^{2}Net consistently outperforms baseline models in all scenarios, demonstrating its migration ability in few-shot learning.

To evaluate the generalization of the models, we also conduct zero-shot and 1/5/10-shot experiments on the English SROIE (Huang et al., 2019) dataset. In the experiment, we trained the models on the training set of CLEX training sets. Since the CLEX dataset is mostly in Chinese, we further introduce DocVQA (Mathew et al., 2021) during training to strengthen the models' ability of English understanding. As for the zero-shot experiment, we directly test on the SROIE test set. And for the 1/5/10-shot experiments, we randomly

[2]https://openai.com/blog/chatgpt

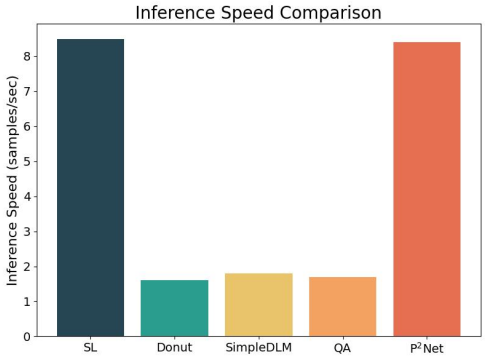

Figure 5: The speed comparison of different models.

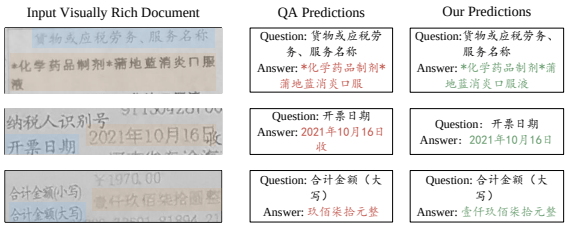

Figure 6: The case study experiment.

select 1/5/10 samples from the SROIE training set for training, and then directly test on the test set of SROIE. Considering the instability during training, we run the training and testing process for 10 times and report the average results. As illustrated in Fig. 4, we find P^{2}Net still outperforms other strong baselines under few-shot settings. Moreover, for P^{2}Net, there is a great performance increase under the few-shot situations compared to those under zero-shot settings. Those findings illustrate the effectiveness of the pointer-based method. Please refer to Appendix A.4 for more experiment results on CLEX, SROIE, and CORD.

Ablation Study. We construct ablation experiments in the zero-shot scenario to demonstrate the validity of the components in P^{2}Net, where we eliminate the following modules: (1) sinusoidal position embedding (*-sin*); (2) the 6 key-related entity linking operations (*-key*); (3) the Question-Context Isolation (*-QCI*); (4) the Question-Head Isolation (*-QHI*); (5) the Question-Tail Isolation (*-QTI*). Please note that if any token linking is missing from the *-key* module, the decoding process is incomplete. As illustrated in Table 4, we could observe that removing the 6 linking operations related to the key entity brings a 0.85% performance drop. It satisfies our intuition that the key-value relations could bring implicit clues to assist extraction. Besides, information isolation methods have positive impacts on the final results since they bring priori knowledge and block out those false results. Meanwhile, because sinusoidal position embedding can effectively record the relative position relationship between tokens, it brings some performance gains.

Speed Comparison. We conduct experiment to compare the inference speed between different models. We test the models directly on 196 samples of the *VAT invoices (rolled)* category and record the inference time, including the time for predicting and decoding. Please refer to Appendix A.8 for the details of the experimental settings. The experiment results are shown in Fig. 5. We find SimpleDLM and QA exhibit slow inference speeds due to multiple queries per sample based on the total categories for the target document. Meanwhile, because those auto-regressive methods such as Dount need to generate the answers word by word, they have a slow inference speed. Despite SL obtains a fast inference speed, the performance under zero/few-shot is far from satisfactory. In contrast, P^{2}Net achieves the balance between performace and inference speed.

Case Study. Several typical cases in the zero-shot experiment on CLEX are visualized in Fig. 6, where we can observe that QA model suffers from errors such as extraction across rows and inaccurate classification of prediction boundaries, but P^{2}Net performs better, even in long and discontinuous visually rich document situations. An important reason is that P^{2}Net considers the key-value relationship, which contributes to extracting the correct results.

6 Conclusion

In this work, to simulate the real-world situation that has various types of complex layout styles and unseen entities are emerging, we propose the CLEX, a human-annotated complex layout KIE dataset with 5,860 images and 1,162 entity categories. And we mainly consider the few/zero-shot KIE task. We also propose P^{2}Net. By leveraging several token linking operations, P^{2}Net incorporates the implicit clues from keys and key-value pairs and extracts value entities in parallel. The zero/few-shot KIE experiments on relevant datasets illustrate that P^{2}Net outperforms state-of-the-art methods with a fast inference speed.

Limitation

Although PPN outperforms traditional methods, there are still challenges, such as inaccurate classification of prediction boundaries, that need to be addressed. We plan to explore these improvements in future work. Furthermore, despite the impressive capabilities of large language models, they still lag behind our approach in the KIE task, both in performance and efficiency. Large models generally require significantly more parameters, whereas our method achieves a balance between inference speed and performance. As a result, we believe our approach holds practical value.

Ethics Statement

We recognize that the information security is crucial. Although all the data from CLEX are crawled from the Internet, our approach involves rigorous data desensitization practices. Specifically, to address the critical concerns associated with information security within the domain of natural language processing, the annotation methodology in this paper incorporates a series of meticulously designed data desensitization practices. These practices are delineated as follows:

Initially, a process of content obfuscation is employed, wherein all identifiable information contained within the original datasets, harvested through web crawling, is obscured. This step ensures the preservation of only the structural templates of the forms, effectively eliminating any direct references to the original content. Such an approach is pivotal in mitigating the risks of inadvertent data leakage.

Subsequently, leveraging the retained document templates, we engage individuals possessing profound expertise and familiarity with real-world data pertinent to the scenarios under investigation. These professionals are tasked with the reconstruction of content, drawing upon publicly available documents, such as adjudication cases, supplemented by their personal expertise. This reconstructed content, while reflective of plausible scenarios, is devoid of any real-world personal data, thereby safeguarding privacy. The synthesis of this content is meticulously designed to maintain a logical coherence, thereby ensuring its utility in subsequent analyses.

The final phase of the data sanitization process encompasses a thorough manual review. This comprehensive inspection serves to validate the effectiveness of the preceding steps in obfuscating sensitive information. It is a critical measure to ensure that the reconstructed datasets bear no resemblance to the original data, thus significantly diminishing the potential for privacy breaches.

Upon completion of this rigorous desensitization protocol, we believe that the refactored samples, now stripped of any original content and reduced to mere templates, present minimal risk of private information leakage.

Acknowledgements

The work is supported by the National Natural Science Foundation of China (62206267 and 62176029), China Postdoctoral Science Foundation Funded Project (2024M763867), Chongqing Key Project of Technological Innovation and Application Development (CSTB2023TIAD-KPX0064).

References

Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Jianfeng Gao, Songhao Piao, Ming Zhou, and Hsiao-Wuen Hon. 2020. Unilmv2: Pseudo-masked language models for unified language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 642–652. PMLR.

Haoyu Cao, Xin Li, Jiefeng Ma, Deqiang Jiang, Antai Guo, Yiqing Hu, Hao Liu, Yinsong Liu, and Bo Ren. 2022. Query-driven generative network for document information extraction in the wild. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pages 4261–4271. ACM.

Panfeng Cao, Ye Wang, Qiang Zhang, and Zaiqiao Meng. 2023. Genkie: Robust generative multimodal document key information extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 14702–14713. Association for Computational Linguistics.

Mengli Cheng, Minghui Qiu, Xing Shi, Jun Huang, and Wei Lin. 2020. One-shot text field labeling using attention and belief propagation for structure information extraction. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 340–348. ACM.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of*

the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pages 4171–4186. Association for Computational Linguistics.

Joseph L Fleiss, Bruce Levin, Myunghee Cho Paik, et al. 1981. The measurement of interrater agreement. *Statistical methods for rates and proportions*, 2(212-236):22–23.

Mingfei Gao, Le Xue, Chetan Ramaiah, Chen Xing, Ran Xu, and Caiming Xiong. 2022. Docquerynet: Value retrieval with arbitrary queries for form-like documents. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 2141–2146. International Committee on Computational Linguistics.

He Guo, Xiameng Qin, Jiaming Liu, Junyu Han, Jingtuo Liu, and Errui Ding. 2019. EATEN: entity-aware attention for single shot visual text extraction. In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*, pages 254–259. IEEE.

Ahmed Hamdi, Elodie Carel, Aurélie Joseph, Mickaël Coustaty, and Antoine Doucet. 2021. Information extraction from invoices. In *16th International Conference on Document Analysis and Recognition, ICDAR 2021, Lausanne, Switzerland, September 5-10, 2021, Proceedings, Part II*, volume 12822 of *Lecture Notes in Computer Science*, pages 699–714. Springer.

Teakgyu Hong, Donghyun Kim, Mingi Ji, Wonseok Hwang, Daehyun Nam, and Sungrae Park. 2021. BROS: A layout-aware pre-trained language model for understanding documents. *CoRR*, abs/2108.04539.

Anwen Hu, Haiyang Xu, Liang Zhang, Jiabo Ye, Ming Yan, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. 2024. mplug-docowl2: High-resolution compressing for ocr-free multi-page document understanding. *CoRR*, abs/2409.03420.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Kai Hu, Zhuoyuan Wu, Zhuoyao Zhong, Weihong Lin, Lei Sun, and Qiang Huo. 2023. A question-answering approach to key value pair extraction from form-like document images. *CoRR*, abs/2304.07957.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. Layoutlmv3: Pre-training for document AI with unified text and image masking. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pages 4083–4091. ACM.

Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and C. V. Jawahar. 2019. ICDAR2019 competition on scanned receipt OCR and information extraction. In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*, pages 1516–1520. IEEE.

Wonseok Hwang, Jinyeong Yim, Seunghyun Park, Sohee Yang, and Minjoon Seo. 2021. Spatial dependency parsing for semi-structured document information extraction. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 330–343. Association for Computational Linguistics.

Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. FUNSD: A dataset for form understanding in noisy scanned documents. In *2nd International Workshop on Open Services and Tools for Document Analysis, OST@ICDAR 2019, Sydney, Australia, September 22-25, 2019*, pages 1–6. IEEE.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. Ocr-free document understanding transformer. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXVIII*, volume 13688 of *Lecture Notes in Computer Science*, pages 498–517. Springer.

Jianfeng Kuang, Wei Hua, Dingkang Liang, Mingkun Yang, Deqiang Jiang, Bo Ren, and Xiang Bai. 2023. Visual information extraction in the wild: Practical dataset and end-to-end solution. In *Document Analysis and Recognition - ICDAR 2023 - 17th International Conference, San José, CA, USA, August 21-26, 2023, Proceedings, Part VI*, volume 14192 of *Lecture Notes in Computer Science*, pages 36–53. Springer.

Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. 2022. Unified named entity recognition as word-word relation classification. pages 10965–10973. AAAI Press.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023a. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Qiwei Li, Zuchao Li, Xiantao Cai, Bo Du, and Hai Zhao. 2023b. Enhancing visually-rich document understanding via layout structure modeling. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, pages 4513–4523. ACM.

Yulin Li, Yuxi Qian, Yuechen Yu, Xiameng Qin, Chengquan Zhang, Yan Liu, Kun Yao, Junyu Han, Jingtuo Liu, and Errui Ding. 2021. Structext: Structured text understanding with multi-modal transformers. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pages 1912–1920. ACM.

Jintao Liu, Kaiwen Wei, and Chenglong Liu. 2024a. Multimodal event causality reasoning with scene graph enhanced interaction network. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 8778–8786. AAAI Press.

Nayu Liu, Kaiwen Wei, Yong Yang, Jianhua Tao, Xian Sun, Fanglong Yao, Hongfeng Yu, Li Jin, Zhao Lv, and Cunhang Fan. 2024b. Multimodal cross-lingual summarization for videos: A revisit in knowledge distillation induced triple-stage training method. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(12):10697–10714.

Bodhisattwa Prasad Majumder, Navneet Potti, Sandeep Tata, James Bradley Wendt, Qi Zhao, and Marc Najork. 2020. Representation learning for information extraction from form-like documents. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 6495–6504. Association for Computational Linguistics.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021. Docvqa: A dataset for VQA on document images. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pages 2199–2208. IEEE.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. 2019. Cord: a consolidated receipt dataset for post-ocr parsing. In *Workshop on Document Intelligence at NeurIPS 2019*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8024–8035.

Qiming Peng, Yinxu Pan, Wenjin Wang, Bin Luo, Zhenyu Zhang, Zhengjie Huang, Yuhui Cao, Weichong Yin, Yongfeng Chen, Yin Zhang, Shikun Feng, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2022. Ernie-layout: Layout knowledge enhanced pre-training for visually-rich document understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 3744–3756. Association for Computational Linguistics.

Jian Song, Di Liang, Rumei Li, Yuntao Li, Sirui Wang, Minlong Peng, Wei Wu, and Yongxin Yu. 2022. Improving semantic matching through dependency-enhanced pre-trained model with adaptive fusion. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 45–57. Association for Computational Linguistics.

Tomasz Stanislawek, Filip Gralinski, Anna Wróblewska, Dawid Lipinski, Agnieszka Kaliska, Paulina Rosalska, Bartosz Topolski, and Przemyslaw Biecek. 2021. Kleister: Key information extraction datasets involving long documents with complex layouts. In *16th International Conference on Document Analysis and Recognition, ICDAR 2021, Lausanne, Switzerland, September 5-10, 2021, Proceedings, Part I*, volume 12821 of *Lecture Notes in Computer Science*, pages 564–579. Springer.

Jianlin Su, Ahmed Murtadha, Shengfeng Pan, Jing Hou, Jun Sun, Wanwei Huang, Bo Wen, and Yunfeng Liu. 2022. Global pointer: Novel efficient span-based approach for named entity recognition. *arXiv preprint arXiv:2208.03054*.

Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. 2020. Circle loss: A unified perspective of pair similarity optimization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 6397–6406. Computer Vision Foundation / IEEE.

Guozhi Tang, Lele Xie, Lianwen Jin, Jiapeng Wang, Jingdong Chen, Zhen Xu, Qianying Wang, Yaqiang Wu, and Hui Li. 2021. Matchvie: Exploiting match relevancy between entities for visual information extraction. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 1039–1045. ijcai.org.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Jiapeng Wang, Chongyu Liu, Lianwen Jin, Guozhi Tang, Jiaxin Zhang, Shuaitao Zhang, Qianying Wang, Yaqiang Wu, and Mingxiang Cai. 2021a. Towards

robust visual information extraction in real world: New dataset and novel solution. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 2738–2745. AAAI Press.

Jiapeng Wang, Tianwei Wang, Guozhi Tang, Lianwen Jin, Weihong Ma, Kai Ding, and Yichao Huang. 2021b. Tag, copy or predict: A unified weakly-supervised learning framework for visual information extraction using sequences. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 1082–1090. ijcai.org.

Sirui Wang, Di Liang, Jian Song, Yuntao Li, and Wei Wu. 2022. DABERT: dual attention enhanced BERT for semantic matching. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 1645–1654. International Committee on Computational Linguistics.

Zilong Wang and Jingbo Shang. 2022. Towards few-shot entity recognition in document images: A label-aware sequence-to-sequence framework. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 4174–4186. Association for Computational Linguistics.

Zilong Wang, Yichao Zhou, Wei Wei, Chen-Yu Lee, and Sandeep Tata. 2023. VRDU: A benchmark for visually-rich document understanding. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2023, Long Beach, CA, USA, August 6-10, 2023*, pages 5184–5193. ACM.

Kaiwen Wei, Xian Sun, Zequn Zhang, Li Jin, Jingyuan Zhang, Jianwei Lv, and Zhi Guo. 2023. Implicit event argument extraction with argument-argument relational knowledge. *IEEE Trans. Knowl. Data Eng.*, 35(9):8865–8879.

Kaiwen Wei, Xian Sun, Zequn Zhang, Jingyuan Zhang, Zhi Guo, and Li Jin. 2021. Trigger is not sufficient: Exploiting frame-aware knowledge for implicit event argument extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4672–4682. Association for Computational Linguistics.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei A. F. Florêncio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. 2021a. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 2579–2591. Association for Computational Linguistics.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. Layoutlm: Pre-training of text and layout for document image understanding. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 1192–1200. ACM.

Yiheng Xu, Tengchao Lv, Lei Cui, Guoxin Wang, Yijuan Lu, Dinei Florêncio, Cha Zhang, and Furu Wei. 2021b. Layoutxlm: Multimodal pre-training for multilingual visually-rich document understanding. *CoRR*, abs/2104.08836.

Wenwen Yu, Ning Lu, Xianbiao Qi, Ping Gong, and Rong Xiao. 2020. PICK: processing key information extraction from documents using improved graph learning-convolutional networks. In *25th International Conference on Pattern Recognition, ICPR 2020, Virtual Event / Milan, Italy, January 10-15, 2021*, pages 4363–4370. IEEE.

Chong Zhang, Ya Guo, Yi Tu, Huan Chen, Jinyang Tang, Huijia Zhu, Qi Zhang, and Tao Gui. 2023. Reading order matters: Information extraction from visually-rich documents by token path prediction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 13716–13730. Association for Computational Linguistics.

Yue Zhang, Bo Zhang, Rui Wang, Junjie Cao, Chen Li, and Zuyi Bao. 2021. Entity relation extraction as dependency parsing in visually rich documents. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 2759–2768. Association for Computational Linguistics.

A Appendix

A.1 More Details about Token Linking Operations

In addition to the token linking operations about the value entities, to further utilize the clues from key entities and key-value pairs, we design additional 6 linking operations about the key entities:

Key_Head to Key_Tail Linking. Similar to the key_head to key_tail linking, the head token of each key entity is pointed to its tail token. For example, the head token "work" will be pointed to the tail token "unit" from the key entity "work unit". The linking type id is 6.

Key_Head to Value_Head Linking. The head token of each key entity is pointed to the head token

of the corresponding value entity. The linking type id is 7.

Value_Tail to Key_Tail Linking. The tail token of each value entity is pointed to its corresponding key tail token. The linking type id is 8.

Question_Head to Key_Head Linking. The head token of each question is pointed to the head token of the corresponding key entity. The linking type id is 9.

Key_Tail to Question_Tail Linking. The tail token of each key entity is pointed to the tail token of the corresponding question. The linking type id is 10.

Discontinuous_Key_Head to Key_Tail Linking. Likewise, the key values could have discontinuous problems, thus we conduct the same operation as in discontinuous value tail-value head linking. For example, given the discontinuous entity "current address", we will connect the "current" and "address". The linking type id is 11.

As for the situations that there are no key entities, we directly skip these 6 key-related linking operations.

A.2 More Details about Token Linking Operations

Question Context Isolation. Since we concatenate the questions to the OCR-parsed context and the expected entities must come from the context instead of from the questions, we masked out the question part and hope the model to find the answers from the OCR-parsed context. Specifically, if the p-th token type in prediction score matrix $Y \in \mathbb{R}^{S\times S\times P}$ related to the key/value entities (i.e., p=1, 2, 5, 6, 7, 8, 11), we use the question context isolation to mask out improbable predictions of keys/values at question positions.

For example, as shown in Fig. 7, as for the OCR-parsed sentence “work unit natural resources current Human province bureau address”, we want to extract the entity type “organization” and “address”, the input is: "$\langle s\rangle$ organization [T] address $\langle/s\rangle$ work unit natural resources current Human province bureau address." The ground truth label “Natural Resources Bureau” and “Human province” must come from the context part, we masked out the question part “$\langle s\rangle$ organization [T] address $\langle/s\rangle$” during inference stage.

Question Head/Tail Isolation. Since the order of the input concatenated questions is self-determined, the start and end positions of each question could be known before inputting to the model. Therefore, to make it easier to judge the start or end position associated with the questions, we design the question head/tail isolation. With this mask mechanism, more prior knowledge will be given to identify the linking operation related to the question head/tail.

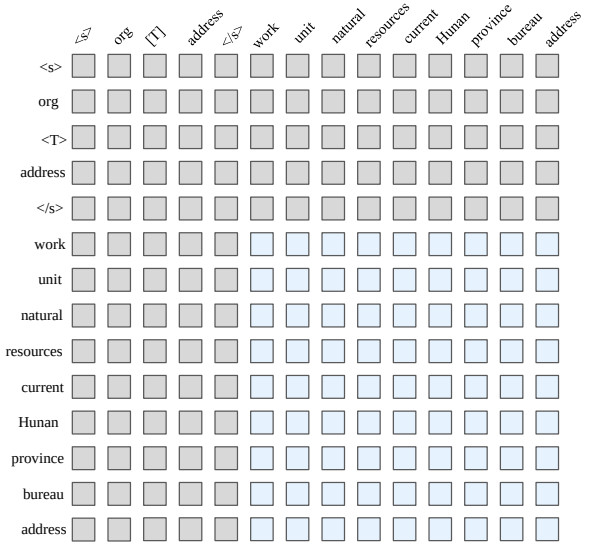

Figure 7: Illustration of the information isolation mechanism for token linking ids equals to 1, 2, 5, 6, 7, 8, 11. In those situations, the head and tail must come from the context, so we mask those in question positions. The boxes in gray means they are masked out.

For example, as shown in Fig. 8, when predicting about the question head pointing to value/key head (i.e., type id is 3 or 9), it would be impossible to predict question head pointing to question tail. Likewise, take the OCR-parsed sentence “work unit natural resources current Human province bureau address” as an example, the head of question "org" and "address" could be known during template construction. As a result, we only need to keep the prediction about the question head/tails and mask other positions.

Moreover, as illustrated in Fig. 9, when predicting about the value/key tail point to question tail (i.e., type id is 4 or 10), we need to keep those predictions end with the question tail and mask other parts.

Then we elementally multiply the mask matrix $M_p^{i,j}$ at the p-th linking type from information isolation mechanism to the prediction matrix Y, obtaining the score matrix W:

$$W_p^{i,j} = Y_p^{i,j} \otimes M_p^{i,j} \tag{6}$$

A.3 More Details about Experiment Settings

To verify the capability of the models, we conduct experiments on the widely-used SROIE (Huang

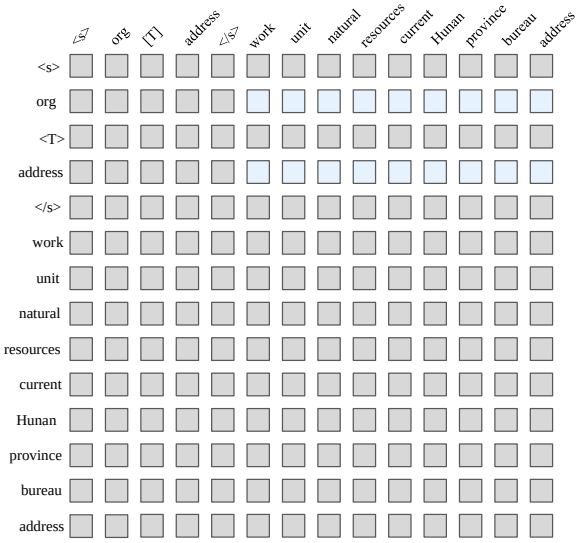

Figure 8: Illustration of the information isolation mechanism for token linking ids equals to 3 and 9. In those situations, the head must come from the question and the tail must come from the key/value (context), so we mask out other parts. The boxes in gray means they are masked out.

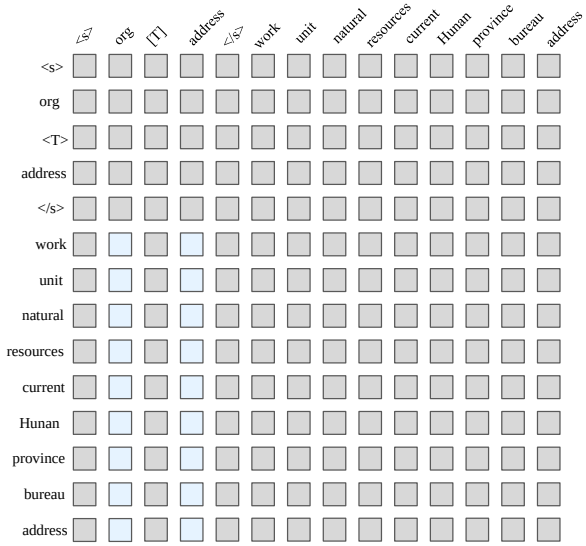

Figure 9: Illustration of the information isolation mechanism for token linking ids equals to 4 and 10. In those situations, the head must come from the key/value (context) and the tail must come from the question, so we mask out other parts. The boxes in gray means they are masked out.

et al., 2019) and CORD (Park et al., 2019) datasets, and the proposed CLEX dataset with complex layouts. In the experiment, Since CLEX is a multilingual dataset, we adopt *LayoutXLM-base* (Xu et al., 2021b) as the backbone. We leverage the OCR tools[3] to get the text content. All the experiments are conducted with Pytorch (Paszke et al., 2019) on one A6000 GPU. The models are trained 30 epochs with a 5e-5 learning rate. The models are evaluated once every 500 training steps. We set the batch size and warmup ratio as 8 and 0.1, respectively. The maximum question slice window size is set as 128. The threshold δ during decoding is set as 0.5. We leverage LoRA (Hu et al., 2022) mechanism for training MLLMs. The optimal hyper-parameters are obtained by grid search.

Following previous works (Jaume et al., 2019; Zhang et al., 2021), we take the entity-level precision (P), recall (R), and F1 score as the measure standards. To evaluate the generation-based methods, we first extract the contents generated for each category, then calculate the start and end positions in the original text, and compare them with the ground-truth positions.

A.4 Full Volume of Data Experiment Results

To explore the maximum capability of the models, we also conduct the experiments under the full volume of data on CLEX dataset, where we divide the test set according to the ratio of 7:3, and let the 70% of the data participate in training. As shown in Table 5, we can find that compared to the experiment results in zero-shot and few-shot situations, the performance of all models has further improved. Meanwhile, the gap between different models is becoming smaller. It shows that the amount of training data has a great influence on the final experimental results.

We also conduct the full volume data experiment on the SROIE and CORD datasets. The experiment results are illustrated in Table 6. We could find that: (1) At the point of the dataset, considering the relatively high experimental results for all the methods (e.g., the state-of-the-art methods has reached to 97% F1 score), it is necessary to propose a new dataset. As a result, in this paper, we release a more complex and larger dataset CLEX. (2) Moreover, at the aspect of the method, the proposed P^{2}Net obtains experimental results comparable to the state-of-the-art methods, reaching to more than 96% F1 score on both datasets. Despite the QGN method still outperforms the other methods on SROIE, but it utilizes data augmentation mechanism for training, so it is not quite fair for comparison. In addition, this paper mainly seeks to discuss the zero-shot and few-shot experiments to fit the complex real-world situations. From the ex-

[3]https://duguang.aliyun.com/

	P	R	F1
SL	27.14	34.57	30.41
Donut	50.54	50.00	50.27
SimpleDLM	93.24	78.98	85.52
QA	93.55	79.30	85.84
WPN	93.89	**79.58**	86.14
P^{2}Net	94.57	79.11	**86.15**

Table 5: The results of different methods in the full volume data situation on CLEX.

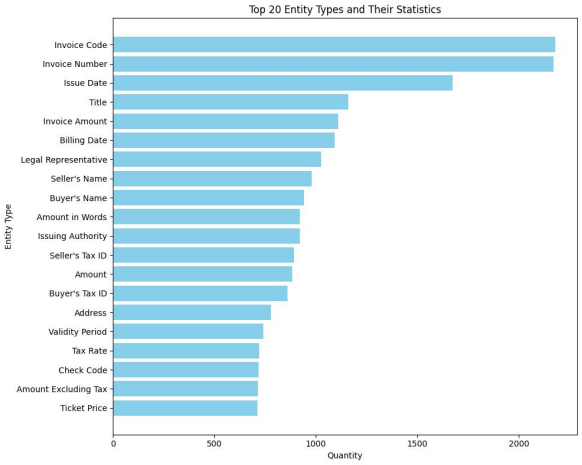

Figure 10: The top 20 entity types and their statistics in CLEX.

periments under the zero/few-shot settings, P^{2}Net outperforms other baselines on both CLEX and SROIE datasets, illustrating its effectiveness and migration ability.

	SROIE	CORD
BERT$_B$ (Devlin et al., 2019)	90.99	89.68
UniLMv2$_B$ (Bao et al., 2020)	94.59	91.98
LayoutXLM$_B$ (Xu et al., 2021b)	94.80	94.84
BERT$_L$ (Devlin et al., 2019)	92.00	90.25
UniLMv2$_L$ (Bao et al., 2020)	94.88	92.05
LayoutLM$_L$ (Xu et al., 2020)	95.24	94.93
LayoutLMv2$_L$ (Xu et al., 2021a)	97.81	96.01
BROS (Hong et al., 2021)	95.48	97.28
PICK (Yu et al., 2020)	96.12	-
VIES (Wang et al., 2021a)	96.12	-
TCPN (Wang et al., 2021b)	96.54	-
P^{2}Net (ours)	96.46	96.14
MatchVIE (Tang et al., 2021)	96.57	-
StrucTexT (Li et al., 2021)	96.88	-
GraphLayoutLM$_B$ (Li et al., 2023b)	-	97.28
GraphLayoutLM$_L$ (Li et al., 2023b)	-	**97.75**
QGN (Cao et al., 2022)	**97.90**	96.84

Table 6: Experiment results (F1 score %) on SROIE and CORD datasets under the full volume of data setting.

A.5 Baseline Details

(1) **Sequence labelling (SL)** (Xu et al., 2021b), which selects from all the categories and predicts which category each token belongs to; (2) **SimpleDLM** (Gao et al., 2022), which formulates the problem as value retrieval problem; (3) **QA** (Hu et al., 2023), which models the KIE as a question answering task, where the questions are the likely entity categories; (4) **Donut** (Kim et al., 2022), which is a method based on an OCR-free transformer trained in an end-to-end manner. (5) **P^{2}Net**, which is the proposed pointer-based network that extracts value entities in parallel; (6) **WPN**, which is a variant of P^{2}Net that does not consider the 6 linking operations about the key entities; (7) **ChatGPT**[4], version *gpt-3.5-turbo-1106*, which has a powerful ability in natural language processing tasks; (8) GPT4V (OpenAI, 2023), version *gpt-4-1106-vision-preview*, which supports multi-modal tasks.

A.6 MLLMs Experiment Details

We leverage the API from OpenAI to conduct the ChatGPT experiment. We select the *gpt-3.5-turbo-0301* version of ChatGPT and *gpt-4-1106-vision-preview* version of GPT4V for zero-shot inference. The input template for those MLLMs is as follows:

Please extract all entities of category *'[Entity_type1]', '[Entity_type2]', ... , '[Entity_typeN]'* from the text and return the corresponding entity content. The text content is: *[TEXT]*. The return format is: *{Entity_type1}: [Entity1_content, Entity2_content, ..., EntityN_content] or []; {Entity_type2}: [Entity1_content, Entity2_content, ..., EntityN_content] or []; ... ;{Entity_typeN}: [Entity1_content, Entity2_content, ..., EntityN_content] or []*. The return format is json, please answer in Chinese, do not return other content.

where *[Entity_type1], [Entity_type2], ..., [Entity_typeN]* indicate all possible entity categories of this kind of form. *[TEXT]* is the text after OCR recognition. The goal is to find the content corresponding to each value semantic entity category.

A.7 Top 20 Entity Types of CLEX

We list the top 20 entity types of the proposed dataset CLEX in Fig. 10.

A.8 Speed Comparison Experimental Settings

All the relevant experiments are conducted on one NVIDIA Tesla-A6000 GPU. The batch size of

[4]https://openai.com/blog/chatgpt

the inference speed experiment is set as 1. The codes are all runs under the *torch=1.8.0, transformers=4.25.1* environment. Finally, we calculate how many samples are finished for inference per second.

A.9 The Differences between Li et al. (2022)

Please note there are certain differences between P^{2}Net and the work in Li et al. (2022). First, Li et al. (2022) is conducted only for text modality, while P^{2}Net supports multi-modal situation. Secondly, the input of Li et al. (2022) is the raw context, P^{2}Net further leverages the key-value relation, where the token linking operations are more complex.