RE^2: Region-Aware Relation Extraction from Visually Rich Documents

Anonymous EACL submission

Abstract

Current research in form understanding predominantly relies on large pre-trained language models, necessitating extensive data for pre-training. However, the importance of layout structure (i.e., the spatial relationship between the entity blocks in the visually rich document) to relation extraction has been overlooked. In this paper, we propose **RE**gion-Aware **R**elation **E**xtraction (**RE**^{2}) that leverages region-level spatial structure among the entity blocks to improve their relation prediction. We design an edge-aware graph attention network to learn the interaction between entities while considering their spatial relationship defined by their region-level representations. We also introduce a constraint objective to regularize the model towards consistency with the inherent constraints of the relation extraction task. To support the research on relation extraction from visually rich documents and demonstrate the generalizability of **RE**^{2}, we build a new benchmark dataset, DIVERSEFORM, that covers a wide range of domains. Extensive experiments on DIVERSEFORM and several public benchmark datasets demonstrate significant superiority and transferability of **RE**^{2} across various domains and languages, with up to 18.88% absolute F-score gain over all high-performing baselines[1].

1 Introduction

Visually Rich Documents (VRDs) encompass various types such as *invoices*, *questionnaire forms*, *financial forms*, *legal documents*, and so on. These documents possess valuable layout information that aids in comprehending their content. Recent research (Liu et al., 2019; Jaume et al., 2019; Yu et al., 2020) has focused on extracting key information, such as entities and relations, from VRDs by leveraging their layout structures and Optical Character Recognition (OCR) results[2]. Figure 1 shows an example where entity recognition aims to identify blocks of text in certain categories, such as *Question*(Q), *Answer*(A), and *Header*(H). Relation extraction further predicts the links among the entities, especially Q-A links indicating that the A block is the corresponding answer to the Q block.

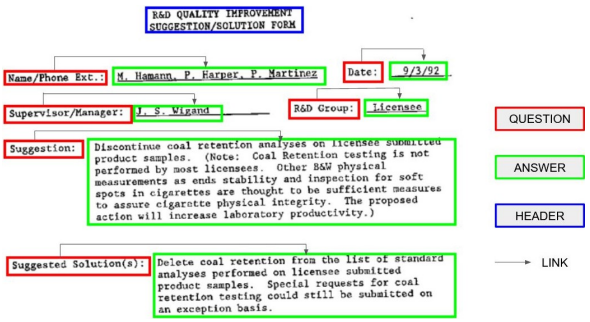


Figure 1: Example of entity and relation extraction from a visually rich document. The colored boxes represent three categories of semantic entities and the arrows represent relations between them.

Extracting key information, especially relations in VRDs is a challenging task. Though similar to traditional extraction tasks in text-only Natural Language Processing (NLP) (Grishman, 1997; Chen et al., 2022), inferring relations in VRDs poses additional challenges. They require not only understanding the semantic meaning of entities but also taking into account the layout information, e.g., the spatial structures among the entity blocks in original VRDs. Previous studies mainly focused on combining the text and layout with language model pre-training (Lu et al., 2019; Su et al., 2020; Chen et al., 2020; Powalski et al., 2021; Xu et al., 2022a; Wang et al., 2022a,b; Huang et al., 2022) or encoding the local layout information by constructing super-tokens (Qian et al., 2019; Liu et al., 2019; Yu et al., 2021; Lee et al., 2022, 2023). However, the

[1]We will make all the programs, model checkpoints and dataset publicly available once the paper is accepted.

[2]Optical Character Recognition will recognize a set of bounding boxes and their corresponding text from VRDs where each bounding box can represent a single word or a cohesive group of words, both semantically and spatially.

layout of the VRDs, especially the relative spatial relationship among the entity blocks, is still yet to be effectively explored for relation extraction.

To this end, we propose **RE**gion-Aware **R**elation **E**xtraction (**RE**^{2}) that leverages region-level spatial structures among the entities to reason about their relations[3]. Specifically, given the question and answer entities from each VRD, we define three categories of region-level representations for each entity block, through which we further characterize the relative spatial relationship between each pair of question and answer entities. We then employ a layout-aware pre-trained language model (i.e., LayoutXLM (Xu et al., 2022a)) to encode the entities and an Edge-aware Graph Attention Network (eGAT) to further learn the interaction between the question and answer entities in a bipartite graph while considering their spatial relationship. To ensure each answer is linked to at most one question, we design a constraint-based learning objective to guide the learning process, in combination with the relation classification objective.

To validate the effectiveness of **RE**^{2}, we conduct extensive experiments on various benchmark datasets for a wide range of languages and domains. We evaluate **RE**^{2} on two public datasets FUNSD (Jaume et al., 2019) and XFUND (Xu et al., 2022b), under supervised, multitask transfer, and zero-shot cross-lingual transfer settings. We also create a new benchmark dataset **DIVERSEFORM** that covers diverse domains, such as Veterans Affairs, visa applications, tax documents, air transport and so on, and evaluate **RE**^{2} for cross-domain transfer. Experimental results show that **RE**^{2} outperforms the previous state-of-the-art approaches with a large margin on (almost) all languages and domains across all settings. Our ablation studies also verify the significant benefit of the region-level spatial structures of entity blocks for relation extraction. The contributions of this work are summarized as follows:

- We are the first to propose the region-level entity representations and utilize them to characterize the spatial structure among the entity blocks, which have been proven to be significantly beneficial to relation extraction from visually rich documents.
- We develop a new framework **RE**^{2} that leverages the spatial structures among the question and answer entities with an effective eGAT network and regularizes model predictions with a novel constraint objective. **RE**^{2} demonstrates superior performance across (almost) all languages and domains under supervised, cross-lingual, and cross-domain transfer settings.
- We contribute **DIVERSEFORM**, a new benchmark dataset that covers a wide range of domains to support the research on information extraction from visually rich documents.

[3]This work mainly focuses on extracting *Q-A* relation given the gold *Question* and *Answer* entities.

2 Related Work

Recent research on visually rich document information extraction shows that incorporating 2D positional embedding and layout coordinates into the pre-trained language models improves VRD understanding (Xu et al., 2020, 2022a; Huang et al., 2022; Powalski et al., 2021; Wang et al., 2022b). To deal with the variation of relation definitions, DocRel (Li et al., 2022) proposes a contrastive learning framework that utilizes the coherence of existing relations in diverse enhanced positive views to generate relational representations. Zhang et al. (2021) further explores entity relation extraction as dependency parsing, incorporating minimum vertical and horizontal distances between the entities as layout heuristics. Compared with all these studies, our approach is the first to propose and incorporate multi-granular spatial structures among the entities, which have shown to significantly improve relation extraction from VRDs.

Graph Attention Networks (GAT) (Veličković et al., 2018) have proven to be efficient for learning on graph-structured data (Zhang et al., 2022a). This is exemplified by the work GraphDoc (Zhang et al., 2022b), a multimodal graph attention-based model that simultaneously utilizes text, layout, and image information for visually rich document understanding. Though several studies (Liu et al., 2019; Lee et al., 2022, 2023) have explored GNNs for entity extraction from VRDs, we are the first to design edge-aware GAT to improve relation extraction from VRDs, which presents additional challenges, encompassing spatial analysis to determine entity layout on the page and semantics between entities for identifying relationships. GNNs have also been applied to relation extraction from textual documents (Zhu et al., 2019; Guo et al., 2019; Zhang et al., 2018). However, these methods cannot be directly adapted to relation extraction from VRDs due to the fundamental differences in doc-

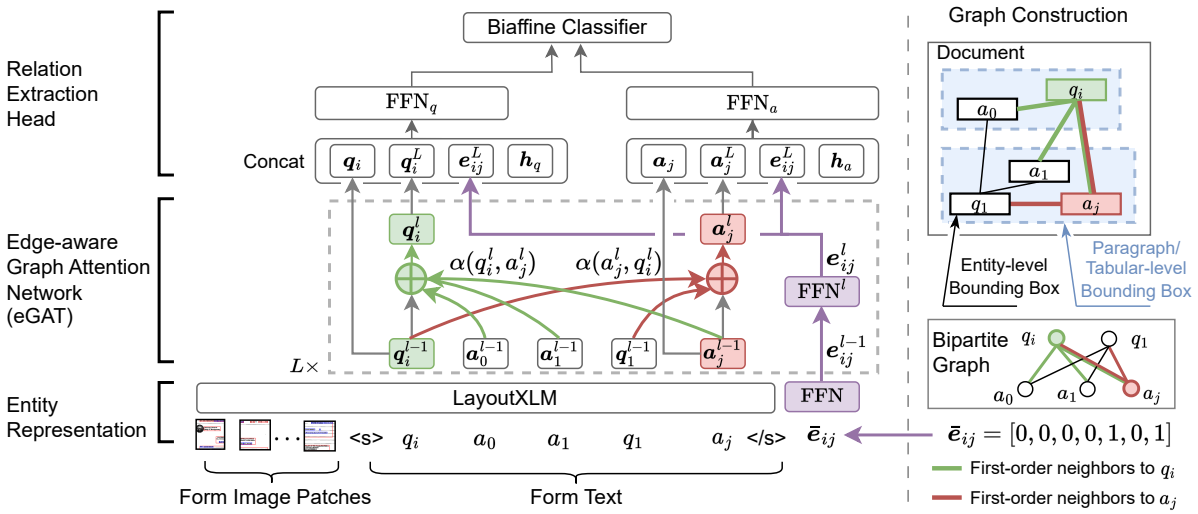


Figure 2: Overview of the **RE**gion-level **R**elation **E**xtraction (**RE**^{2}) framework. In the eGAT layer, the representation of each entity is updated based on the attention scores of its first-order neighbors.

ument formats, structures, and the key challenges encountered in relation extraction: text-only documents primarily rely on linguistic cues and phrases for relation extraction, whereas VRDs necessitate consideration of both semantics and spatial context. Given that, we innovatively incorporate a multi-granular layout heuristic into an edge-aware graph attention network, placing greater emphasis on capturing more fine-grained layout structures.

3 Approach

Given a visually rich document D, a set of question entities $Q = \{q_1, q_2, ..., q_m\}$ and answers $A = \{a_1, a_2, ..., a_n\}$, we aim to identify all the connected pairs (q, a) where $q \in Q$ and $a \in A$, indicating that a is the corresponding answer of q. Each q_i or a_j can be denoted as $\{[w_0, w_1, \cdots, w_t], (x_0, y_0, x_1, y_1)\}$, where $[w_0, w_1, \cdots, w_t]$ is the sequence of words denoting the entity span and (x_0, y_0, x_1, y_1) is the coordinates for the entity bounding box. Figure 2 illustrates our **RE**^{2} framework that aims to leverage region-level spatial structures among the question and answer blocks to detect their association.

3.1 Entity Representation

We first learn the encoding of question and answer entities based on LayoutXLM (Xu et al., 2022b), a layout-aware transformer-based model that has been extended to support multilingualism by pre-training on multilingual VRD datasets.

Given a set of question entities $Q = \{q_1, q_2, ..., q_m\}$ and answers $A = \{a_1, a_2, ..., a_n\}$ from document D, we obtain the entity embeddings $\boldsymbol{Q} = \{\boldsymbol{q_1}, \boldsymbol{q_2}, ..., \boldsymbol{q_m}\}$, $\boldsymbol{A} = \{\boldsymbol{a_1}, \boldsymbol{a_2}, ..., \boldsymbol{a_n}\}$, $\boldsymbol{q_i}, \boldsymbol{a_i} \in \mathbb{R}^{1\times \mathrm{F}}$, where F is the entity feature dimension[4]. For entities with multiple tokens, we use the embedding of their first tokens as their representations[5].

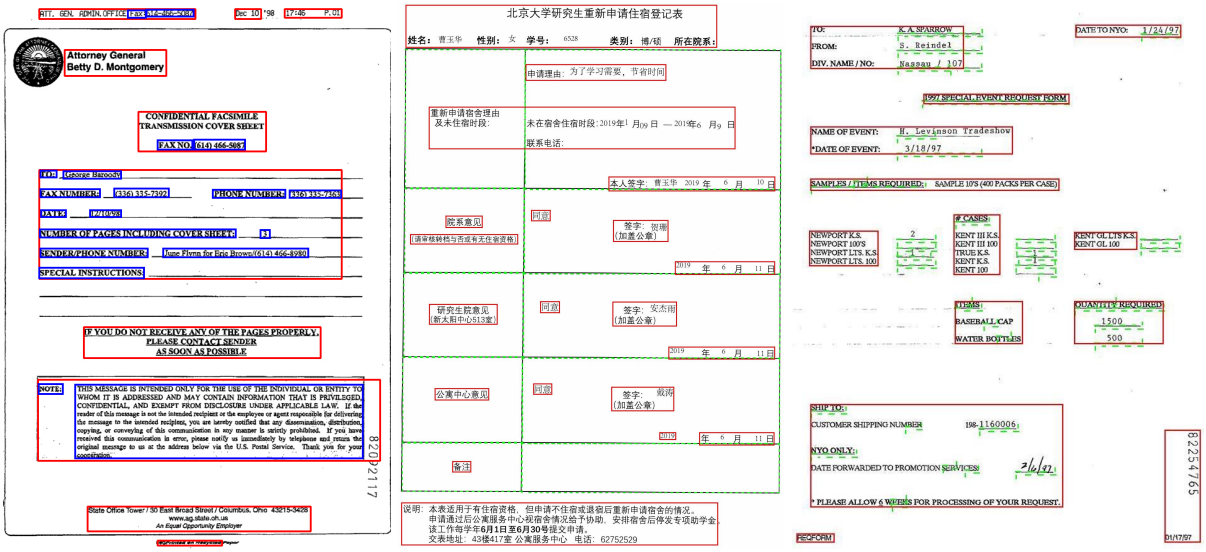
3.2 Region-Aware Graph Construction

Based on the spatial structures of the input VRD, we define three distinct categories of regions (i.e., bounding box) for each entity: (1) an **entity-level bounding box** that refers to the bounding box encompassing the entire entity span and is obtained by merging the bounding boxes of all the words in a span obtained by OCR (Liu et al., 2019; Yu et al., 2020); (2) a **paragraph-level bounding box** that is defined as a visually distinct section for the paragraph where the entity occurs within a document and corresponds to the clustering of words that are located within a dense region. The paragraph-level bounding boxes are extracted by an existing tool, EasyOCR[6], which takes the maximum horizontal and vertical distances between adjacent word-level bounding boxes as hpyerparameters to merge them into paragraph-level bounding boxes; and (3) a **tabular-based bounding box** if the entity occurs in a tabular structure. We define a tabular-based

[4]We use bold symbols to denote vectors.

[5]Preliminary experiments showed use of first subtoken performed better than average embedding of all subtokens.

[6]https://www.jaided.ai/easyocr/



(a) (b) (c)

Figure 3: Entity level bounding box (for question and answer entities) are shown in blue, paragraph-level bounding box in red and tabular-based bounding box in green.

bounding box as the coordinates of a table cell. Note that each entity can only appear in either a paragraph or a table, so other than its entity-level bounding box, we always assign either a paragraph-level or tabular-based bounding box for each entity, instead of both. Our preliminary results show that a tabular-based bounding box is vital because tabular structures are usually not well-captured by existing OCR tools. Illustrations of the three types of regions are shown in Figure 3. The pseudocode for extracting paragraph/tabular regions is present in Appendix H.

To characterize the links between the question and answer entities, we further propose to construct a complete bipartite graph, $G = (Q, A, E)$, for each visually rich document, where the question entities $Q = \{q_1, q_2, ..., q_m\}$ and answers $A = \{a_1, a_2, ..., a_n\}$ are the nodes, and for each pair of q_i and a_j, there is an edge $e_{ij} \in E$ connecting them. Each entity is represented by the encoding learned from LayoutXLM as detailed in Section 3.1, and each edge is represented by a one-hot encoding vector based on the spatial relationship between the three categories of bounding boxes of the question and answer:

$$\bar{\boldsymbol{e}}_{ij} = [\mathrm{I}, \mathrm{E}^1_{lr}, \mathrm{E}^1_{tb}, \mathrm{E}^0_{lr}, \mathrm{E}^0_{tb}, \mathrm{R}_{lr}, \mathrm{R}_{tb}],$$

where each term is an indicator variable: I indicates whether the two entities are within the same paragraph/tabular region. If so, $\mathrm{I} = 1$, otherwise, $\mathrm{I} = 0$. When the two entities are from the same paragraph/tabular region, E^1_{lr} and E^1_{tb} further indicate the left-right (lr) and top-bottom (tb) spatial relationship of their entity-level bounding boxes. For example, $\mathrm{E}^1_{lr} = 1$ indicates that the entity-level bounding boxes of the two entities have a left-right spatial relation, otherwise, $\mathrm{E}^1_{lr} = 0$. When the two entities are not from the same paragraph/tabular region, E^0_{lr} and E^0_{tb} indicate the left-right and top-bottom spatial relationship of their entity-level bounding boxes, while R_{lr} and R_{tb} indicate the left-right and top-bottom spatial relationship of their paragraph/tabular level bounding boxes. Note that when the two entities are from the same paragraph/tabular region, the indicators of E^0_{lr}, E^0_{tb}, R_{lr}, R_{tb} will be all zero.

To obtain a dense representation of each edge, we pass each one-hot encoding vector $\bar{\boldsymbol{e}}_{ij}$ to a feed-forward network, and the resulting vector $\boldsymbol{e}_{ij} = \mathrm{FFN}(\bar{\boldsymbol{e}}_{ij})$ is assigned as the edge weight between q_i and a_j, where $\boldsymbol{e}_{ij} \in \mathbb{R}^{1\times \mathrm{F}/2}$.

3.3 Edge-aware Graph Attention Network

We further propose an edge-aware graph attention network (eGAT) , extended from the graph attention network (GAT) (Veličković et al., 2018) by incorporating the edge weights inferred by spatial information to learn the interaction between the question and answer nodes. In our experiments, eGAT consists of 2 encoding layers, while each layer updates the node embeddings based on the first-order neighbors with masked self-attention.

Specifically, given the node embeddings at layer l, $\boldsymbol{Q}^l = \{\boldsymbol{q}^l_1, \boldsymbol{q}^l_2, ..., \boldsymbol{q}^l_m\}$, $\boldsymbol{A} = \{\boldsymbol{a}^l_1, \boldsymbol{a}^l_2, ..., \boldsymbol{a}^l_n\}$, we first compute the attention weight between q_i

and a_j as follows

$$\text{att}(\boldsymbol{W}^l \boldsymbol{q}_i^l, \boldsymbol{W}^l \boldsymbol{a}_j^l) = \boldsymbol{W}_{att}^\top (\boldsymbol{W}^l \boldsymbol{q}_i^l || \boldsymbol{W}^l \boldsymbol{a}_j^l)$$

$$\boldsymbol{c}(q_i^l, a_j^l) = \text{LeakyReLu}\Big(\text{att}\Big(\boldsymbol{W}^l \boldsymbol{q}_i^l, \boldsymbol{W}^l \boldsymbol{a}_j^l\Big)\Big)$$

$$\boldsymbol{\alpha}(q_i^l, a_j^l) = \underset{j}{\text{softmax}}\Big(\sum(\boldsymbol{e}_{ij}^l \cdot \boldsymbol{c}(\boldsymbol{q}_i^l, \boldsymbol{a}_j^l))\Big)$$

where $\cdot$ denotes scalar multiplication. $\boldsymbol{W}^l \in \mathbb{R}^{F' \times F}$ is a parameter matrix for shared linear transformation for $\boldsymbol{q}_i^l$ and $\boldsymbol{a}_j^l$. $\boldsymbol{W}_{att} \in \mathbb{R}^{2F'}$ is a weight vector for the attention mechanism. $||$ denotes the catenation operation. $\boldsymbol{e}_{ij}^l = \text{FFN}^l(\boldsymbol{e}_{ij}^{l-1})$ where $\boldsymbol{e}_{ij}^0$ is the initial dense representation $\boldsymbol{e}_{ij}$ of each edge.

The resulting edge-aware normalized attention scores are then used to update the hidden representations of the question and answer nodes, respectively, with residual connection:

$$\bar{\boldsymbol{q}}_i^{l+1} = \boldsymbol{q}_i + \sum_{j \in \mathcal{N}_i} \boldsymbol{\alpha}(q_i^l, a_j^l) \boldsymbol{W} \boldsymbol{a}_j^l$$

$$\bar{\boldsymbol{a}}_j^{l+1} = \boldsymbol{a}_j + \sum_{i \in \mathcal{M}_j} \boldsymbol{\alpha}(a_j^l, q_i^l) \boldsymbol{W} \boldsymbol{q}_i^l$$

where $\mathcal{N}_i$ and $\mathcal{M}_j$ denotes the first order neighbors of q_i and a_j respectively.

For each layer of eGAT, we apply multi-head attention (Vaswani et al., 2017), where each attention head performs operations independently, and the mean of all attention heads is taken for aggregation. The updated representation of question node q_i and answer a_j is computed as follows:

$$\boldsymbol{q}_i^{l+1} = \sigma\Big(\frac{1}{K}\sum_{k=1}^{K}\Big(\boldsymbol{q}_i + \sum_{j \in \mathcal{N}_i} \boldsymbol{\alpha}(q_i^l, a_j^l)^k \boldsymbol{W}^k \boldsymbol{a}_j^l\Big)\Big)$$

$$\boldsymbol{a}_j^{l+1} = \sigma\Big(\frac{1}{K}\sum_{k=1}^{K}\Big(\boldsymbol{a}_j + \sum_{i \in \mathcal{M}_j} \boldsymbol{\alpha}(a_j^l, q_i^l)^k \boldsymbol{W}^k \boldsymbol{q}_i^l\Big)\Big)$$

where K is the number of independent attention heads and $\boldsymbol{W}^k$ denotes the weight matrix for the k^{th} attention head. $\sigma(\cdot)$ denotes a non-linear function (ELU is used for experiments). $\boldsymbol{q}_i^{l+1}$ and $\boldsymbol{a}_j^{l+1}$ are then used as input node embeddings for layer $l+1$.

3.4 Relation Extraction

Binary Relation Prediction For each pair of question q_i and answer a_j, we aim to predict a binary label, indicating whether the answer corresponds to the particular question or not. The representation of q_i or a_j is formed by concatenating the LayoutXLM embedding, the final node embedding learned from eGAT, the edge representation of the pair of q_i and a_j, and an entity type representation (i.e., question or answer). The entity type representation is learned by an embedding layer during training. The inclusion of entity type embedding is important in determining the direction of the relation between the two entities. The resulting representations q_i and a_j are then passed through two feed-forward networks and a biaffine classifier (Dozat and Manning, 2017) to obtain a score $s_{i,j}$ for determining whether the pair is associated or not.

$$\boldsymbol{q}_i^{'} = \text{FFN}_\text{q}(\boldsymbol{q}_i \parallel \boldsymbol{q}_i^L \parallel \boldsymbol{e}_{ij}^L \parallel \boldsymbol{h}_q)$$

$$\boldsymbol{a}_j^{'} = \text{FFN}_\text{a}(\boldsymbol{a}_j \parallel \boldsymbol{a}_j^L \parallel \boldsymbol{e}_{ij}^L \parallel \boldsymbol{h}_a)$$

$$s_{ij} = \boldsymbol{q}_i^{'} \boldsymbol{U} \boldsymbol{a}_j^{'} + \boldsymbol{V}(\boldsymbol{q}_i^{'} \circ \boldsymbol{a}_j^{'}) + \boldsymbol{b}$$

where $\boldsymbol{h}_q$ and $\boldsymbol{h}_a$ are the type embeddings of question and answer entities. Note that $\boldsymbol{h}_q$ and $\boldsymbol{h}_a$ remain the same across all questions and answers, respectively. $\boldsymbol{U}, \boldsymbol{V}$ and $\boldsymbol{b}$ are trainable parameters. During training, the loss is computed following the cross-entropy loss

$$\mathcal{L}_b = -\sum y \cdot \log(p_{ij}).$$

where $y \in \{0, 1\}$ is the target binary label and $p_{ij} = \text{softmax}(s_{ij})$, indicating the probability of a relation between q_i and a_j.

Constraint Loss Our preliminary study shows that without any constraint, the model tends to predict multiple questions to be associated with one answer, which is against the definition of relation extraction for VRDs, where each answer is linked to at most one question. To address this issue, we incorporate the constraint into the learning process in the form of a constraint loss. Previous work (Li et al., 2019; Wang et al., 2020) demonstrated that declarative logical constraints can be converted into differentiable functions, and help regularize the model towards consistency with the logical constraints. We design a declarative logical constraint that holds true for relation extraction task from VRDs as follows, $\forall a_j \in A,\ \forall q_i \in Q$,

$$\text{rel}(q_i, a_j) \rightarrow \bigwedge_{q_k \in \mathbf{Q} \setminus \{q_i\}} \neg \text{ rel } (q_k, a_j).$$

This means, for any $a_j \in A$, if there exists one relation link between a_j and any particular q_i among all questions, there cannot be another relation link

Model	EN	ZH	JA	ES	FR	IT	DE	PT	Avg.
XLM-RoBERTa$_{\text{BASE}}$ (Conneau et al., 2020)	26.59	51.05	58.00	52.95	49.65	53.05	50.41	39.82	47.69
InfoXLM$_{\text{BASE}}$ (Chi et al., 2021)	29.20	52.14	60.00	55.16	49.13	52.81	52.62	41.70	49.10
LayoutXLM$_{\text{BASE}}$ (Xu et al., 2022b)	54.83	70.73	69.63	68.96	63.53	64.15	65.51	57.18	64.32
LiLT[InfoXLM]$_{\text{BASE}}$ (Wang et al., 2022a)	62.76	72.97	70.37	71.95	69.65	70.43	65.58	58.74	67.81
RE^2 (Our Approach)	**71.76**	**79.60**	**75.36**	**75.59**	**76.38**	**77.45**	**75.86**	**59.76**	**73.98**

Table 1: Language-specific fine-tuning results (F1%) on FUNSD(EN) and XFUND.

Model	EN	ZH	JA	ES	FR	IT	DE	PT	Avg.
XLM-RoBERTa$_{\text{BASE}}$	26.59	16.01	26.11	24.40	22.40	23.74	22.88	19.96	22.76
InfoXLM$_{\text{BASE}}$	29.20	24.05	28.51	24.81	24.54	21.93	20.27	20.49	24.23
LayoutXLM$_{\text{BASE}}$	54.83	44.94	44.08	47.08	44.16	40.90	38.20	36.85	43.88
LiLT[InfoXLM]$_{\text{BASE}}$	62.76	47.64	50.81	49.68	52.09	46.97	41.69	42.72	49.30
RE^2 (Our Approach)	**71.76**	**66.32**	**64.42**	**58.82**	**69.02**	**61.83**	**60.57**	**43.87**	**62.08**

Table 2: Zero-shot cross-lingual results (F1%) (trained on EN (FUNSD) and tested on other languages)

for this answer a_j. We further define the following constraint loss derived from the logical constraints:

$$\mathcal{L}_c = y \cdot \left| \log(p_{ij}) - \frac{1}{|Q|-1} \sum_{\substack{k=0 \\ k \neq i}}^{|Q|} \log\left(1 - p_{kj}\right) \right|$$

where Q denotes the whole set of questions in the document.

Overall Learning Objective The overall learning objective is a weighted combination of the binary cross entropy loss and the constraint loss:

$$\mathcal{L} = \beta \mathcal{L}_b + \delta \mathcal{L}_c$$

where β and δ are hyperparameters.

4 Experiment Settings

4.1 Datasets

The main challenge of relation extraction from visually rich documents lies in the varying layouts of form-like documents from different domains or languages. On the other hand, the region-level spatial structures defined in **RE**^{2} are domain and language-independent. To validate the effectiveness of **RE**^{2}, we perform experiments on several benchmark datasets covering a wide range of languages and domains.

FUNSD The FUNSD dataset (Jaume et al., 2019) is a subset of the RVL-CDIP dataset (Harley et al., 2015) containing realistic scanned document images with ground truth OCR. It provides bounding boxes for entity spans and annotations for four types: *Question*, *Answer*, *Header*, and *Other*. The dataset includes relational links among these entities, with a particular focus on Question-Answer links. We adopt the same data split and experimental settings as previous studies (Xu et al., 2022b; Wang et al., 2022a), using 149 documents for training, 50 documents for evaluation, and reporting the best performance on the evaluation set.

XFUND XFUND (Xu et al., 2022b) is a diverse multilingual dataset containing visually rich documents in seven languages: Portuguese, Chinese, Spanish, French, Japanese, Italian, and German. The dataset encompasses a wide range of document structures, reflecting variations in document conventions across languages and countries. With a total of 1,393 fully annotated forms, each language includes 149 forms in the train set and 50 forms in the test set. The dataset provides human-annotated ground truth OCR, entity annotations, and relation annotations. Notably, the XFUND dataset shares similarities with the FUNSD dataset in terms of its document format.

DIVERSEFORM To best demonstrate the performance of domain transfer of **RE**^{2}, we further create a new dataset, **DIVERSEFORM**, by curating government forms from Aggarwal et al. (2020) and Sarkar et al. (2020). These forms encompass a wide range of question types, including checkboxes, tables, multiple-choice questions (MCQs), and fill-in-the-blank fields. The domains of the forms cover various areas such as Veterans Affairs, visa applications, tax documents, air transport, legal forms, vehicle-related forms from the Department of Motor Vehicles (DMV), and miscellaneous forms from different government agencies. These forms are of single page and were originally empty and they are designed to collect confidential information such as health data and

Model	EN	ZH	JA	ES	FR	IT	DE	PT	Avg.
XLM-RoBERTa$_{\text{BASE}}$	36.38	67.97	68.29	68.28	67.27	69.37	68.87	60.82	63.41
InfoXLM$_{\text{BASE}}$	36.99	64.93	64.73	68.28	68.31	66.90	63.84	57.63	61.45
LayoutXLM$_{\text{BASE}}$	66.71	82.41	81.42	81.04	82.21	83.10	78.54	70.44	78.23
LiLT[InfoXLM]$_{\text{BASE}}$	74.07	84.71	**83.45**	**83.35**	84.66	**84.58**	78.78	**76.43**	81.25
RE^2 (Our Approach)	**74.11**	**88.25**	82.27	83.23	**86.83**	84.02	**81.89**	71.04	**81.46**

Table 3: Multitask fine-tuning performance (F1%) on FUNSD(EN) and XFUND.

Model	DIVERSEFORM	FUNSD → DIVERSEFORM	DIVERSEFORM → FUNSD
LayoutXLM$_{\text{BASE}}$	69.72	37.33	32.58
LiLT[InfoXLM]$_{\text{BASE}}$	64.15	41.56	30.26
RE^2	70.87	41.78	50.32

Table 4: Supervised results on **DIVERSEFORM** and cross-domain transfer results between **DIVERSEFORM** and FUNSD. (F1%)

tax details. To populate the forms, we employed two annotators who used synthetic data generated by The One Generator[7] for fields such as names, addresses, and other necessary information. This approach ensures the privacy and security of individuals' personal information while providing a realistic representation of the data typically found in these government forms. We then hire another annotator to label the *Question* and *Answer* entities as well as their relations for these documents using the annotation tool UBIAI[8], which also offers its customized OCR model for extracting text from uploaded images. However, due to the serialized top-left to bottom-right text extraction approach of the OCR, the spans of entities are sometimes fragmented in complex layout forms. During the annotation process, these fragmented spans are identified and merged to achieve the correct serialization of spans. After labeling the entities and relations for these documents, we further hire three annotators to validate the annotations. All the annotators are senior undergraduate students majoring in Computer Science and are paid a rate of $\$15/hour$. We name the final annotated dataset as **DIVERSEFORM**, which comprises a total of 150 training documents and 50 testing documents. Details of **DIVERSEFORM** annotation and statistics is in Appendix C.

4.2 Experiment Results

Language-specific fine-tuning results are presented in Table 1, where each model is fine-tuned on language X and tested on language X. The experimental findings show that the proposed model outperforms all the baselines across all evaluated languages. To evaluate the **cross-lingual zero-shot transfer** capability, the model is fine-tuned on the FUNSD dataset in English, followed by testing on multiple languages. The experimental results, as shown in Table 2, demonstrate the superiority of our model over the baseline approach in terms of zero-shot performance. This outcome provides compelling evidence that the incorporated region-level spatial structures and constraints for relation extraction exhibit effective transferability across different languages. We also conduct a significance test for both our approach and the best-performing baseline (i.e., LiLT[InfoXLM]$_{\text{BASE}}$ (Wang et al., 2022a)) under the settings of language-specific fine-tuning and cross-lingual zero-shot transfer. As shown in Table 7 in Appendix D, our approach significantly outperforms the baseline under both settings.

Furthermore, the **multitask fine-tuning** results are shown in Table 3, where the model is trained on the training sets of all languages and then tested on each individual language. The achieved results demonstrate superior performance, indicating that the model successfully learns the layout invariance present across different languages. By effectively capturing and leveraging the shared layout characteristics, the model exhibits improved generalization capabilities, leading to enhanced performance across diverse linguistic contexts. This highlights the significance of incorporating layout information in cross-lingual settings and underscores the model's ability to adapt and transfer its learned knowledge to effectively process and understand documents in various languages.

Note that **RE**^{2} shows less competitive performance on Portuguese (PT) due to more complex layout structures. Portuguese forms exhibit a com-

[7] https://theonegenerator.com/
[8] https://ubiai.tools/

Model	EN	ZH	JA	ES	FR	IT	DE	PT	Avg.
RE^2	**71.76**	**79.60**	**75.36**	**75.59**	**76.38**	**77.45**	**75.86**	**59.76**	**73.98**
- node embedding	70.19	78.93	75.00	74.60	76.00	76.82	73.20	57.29	72.75
- edge embedding	57.42	69.37	67.93	72.01	73.73	69.67	63.48	55.61	66.15
- constraint loss	68.52	77.77	74.49	74.78	75.20	75.66	73.61	57.48	72.19
- entity level regions	44.69	76.89	66.71	73.11	62.44	70.63	62.10	44.30	62.61
- paragraph/tabular regions	71.57	79.5	74.17	72.05	74.98	76.79	74.55	57.49	72.64

Table 5: Ablation study results (F1%) on eGAT (node and edge embeddings), constraint loss, paragraph/tabular regions and entity level regions.

bination of mixed tables and paragraph structures, making it challenging to determine the appropriate usage for paragraph-level regions or tabular regions. An example is shown in Appendix E.

We also assess the generalization of **RE**^{2} and two high-performing baselines based on **DIVERSEFORM** and FUNSD, which cover two sets of distinct domains. We conduct experiments under the settings of both domain-specific fine-tuning and cross-domain transfer where the models are trained on one dataset and tested on the other. As shown in Table 4, **RE**^{2} significantly outperforms the two strong baselines when fine-tuned on **DIVERSEFORM** and tested on **DIVERSEFORM** or FUNSD. The improvement of **RE**^{2} when it's trained on FUNSD and tested on **DIVERSEFORM** is marginal, probably due to the greater diversity and complexity in document layout of **DIVERSEFORM** compared to FUNSD.

4.3 Ablation Study

Effect of Node and Edge Embeddings from eGAT The node and edge embeddings from eGAT are concatenated with the entity representations before being passed to the biaffine classifier. A series of ablation studies are conducted to assess the individual contributions of the layout information. The results of these studies are presented in Table 5. Figure 6 in Appendix F provides visual evidence that solely relying on the updated node embeddings from eGAT fails to adequately capture the layout heuristics and results in the omission of numerous relations. Conversely, employing only the updated edge embeddings without considering the node embeddings leads to an over-prediction of relations with limited regard for the semantic relevance of the entities involved. Optimal performance is achieved through the joint utilization of both node and edge embeddings, indicating the importance of integrating both sources of information to effectively capture the region-level spatial structures and consider the semantic context of the relations.

Effect of Constraint Loss The constraint loss has been modeled to encourage each answer entity to be linked to at most one question. Table 5 shows that incorporating the constraint loss significantly improves the F1 score of **RE**^{2}, especially precision. The detailed experimental results are evidenced in Appendix G.

Effect of Region Information We also investigate the impact of each category of regions on characterizing the spatial relationship among the entities and further affecting the performance of **RE**^{2}. As shown in Table 5, the inclusion of each category of region information significantly improves the performance of **RE**^{2}. The absence of entity-level regions resulted in a substantial decrease in performance, underscoring the vital role of pairwise entity layout information, i.e., whether the question and answer entities are arranged vertically (top-bottom) or horizontally (left-right). Figure 7 in Appendix F shows an example to compare the relation predictions with and without paragraph/tabular regions, indicating that incorporating paragraph/tabular regions helps prevent the model from predicting relations across semantically different regions. The result of this ablation study proves the effectiveness of the multi-granular region information.

5 Conclusion

In this work, we propose a novel entity relation extraction model, **RE**^{2}, that incorporates layout heuristics and constraints that are generalizable across different languages. Experimental results on 8 different languages and our proposed dataset **DIVERSEFORM** show the effectiveness of our proposed method under four settings (language-specific, cross-lingual zero-shot, multi-lingual fine-tuning, and cross-domain transfer).

Limitations

In this work, we found the incorporation of layout heuristics to be compelling and we are excited by how leveraging region information improves performance drastically. One of the limitations of our model is its reliance on a relatively limited set of heuristics and features. For instance, we have not yet incorporated visual information and template-based knowledge, which could potentially improve the accuracy and robustness of the relation extraction task. Additionally, the current model employs an exhaustive inference approach, considering all possible relations during prediction. While this ensures comprehensive coverage, it also results in longer inference times for each relation type. These limitations indicate avenues for further improvement, such as exploring additional heuristics and incorporating more efficient inference strategies, to enhance the performance and efficiency of our model.

References

Milan Aggarwal, Mausoom Sarkar, Hiresh Gupta, and Balaji Krishnamurthy. 2020. Multi-modal association based grouping for form structure extraction. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 2075–2084.

Muhao Chen, Lifu Huang, Manling Li, Ben Zhou, Heng Ji, and Dan Roth. 2022. New frontiers of information extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorials*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *ECCV*.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing.

Ralph Grishman. 1997. Information extraction: Techniques and challenges. In *Information Extraction A Multidisciplinary Approach to an Emerging Information Technology: International Summer School, SCIE-97 Frascati, Italy, July 14–18, 1997*, pages 10–27. Springer.

Zhijiang Guo, Yan Zhang, and Wei Lu. 2019. Attention guided graph convolutional networks for relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 241–251, Florence, Italy. Association for Computational Linguistics.

Adam W. Harley, Alex Ufkes, and Konstantinos G. Derpanis. 2015. Evaluation of deep convolutional nets for document image classification and retrieval.

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 *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 4083–4091, New York, NY, USA. Association for Computing Machinery.

Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. Funsd: A dataset for form understanding in noisy scanned documents.

Chen-Yu Lee, Chun-Liang Li, Timothy Dozat, Vincent Perot, Guolong Su, Nan Hua, Joshua Ainslie, Renshen Wang, Yasuhisa Fujii, and Tomas Pfister. 2022. Formnet: Structural encoding beyond sequential modeling in form document information extraction.

Chen-Yu Lee, Chun-Liang Li, Hao Zhang, Timothy Dozat, Vincent Perot, Guolong Su, Xiang Zhang, Kihyuk Sohn, Nikolai Glushnev, Renshen Wang, Joshua Ainslie, Shangbang Long, Siyang Qin, Yasuhisa Fujii, Nan Hua, and Tomas Pfister. 2023. Formnetv2: Multimodal graph contrastive learning for form document information extraction.

Tao Li, Vivek Gupta, Maitrey Mehta, and Vivek Srikumar. 2019. A logic-driven framework for consistency of neural models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3924–3935, Hong Kong, China. Association for Computational Linguistics.

Xin Li, Yan Zheng, Yiqing Hu, Haoyu Cao, Yunfei Wu, Deqiang Jiang, Yinsong Liu, and Bo Ren. 2022. Relational representation learning in visually-rich documents.

Xiaojing Liu, Feiyu Gao, Qiong Zhang, and Huasha Zhao. 2019. Graph convolution for multimodal information extraction from visually rich documents. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, pages 32–39. Association for Computational Linguistics.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. *ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks*. Curran Associates Inc., Red Hook, NY, USA.

Rafał Powalski, Łukasz Borchmann, Dawid Jurkiewicz, Tomasz Dwojak, Michał Pietruszka, and Gabriela Pałka. 2021. Going full-tilt boogie on document understanding with text-image-layout transformer. In *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part II 16*, pages 732–747. Springer.

Yujie Qian, Enrico Santus, Zhijing Jin, Jiang Guo, and Regina Barzilay. 2019. GraphIE: A graph-based framework for information extraction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 751–761, Minneapolis, Minnesota. Association for Computational Linguistics.

Mausoom Sarkar, Milan Aggarwal, Arneh Jain, Hiresh Gupta, and Balaji Krishnamurthy. 2020. Document structure extraction using prior based high resolution hierarchical semantic segmentation. In *European Conference on Computer Vision*, pages 649–666. Springer.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks.

Haoyu Wang, Muhao Chen, Hongming Zhang, and Dan Roth. 2020. Joint constrained learning for event-event relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 696–706, Online. Association for Computational Linguistics.

Jiapeng Wang, Lianwen Jin, and Kai Ding. 2022a. Lilt: A simple yet effective language-independent layout transformer for structured document understanding.

Wenjin Wang, Zhengjie Huang, Bin Luo, Qianglong Chen, Qiming Peng, Yinxu Pan, Weichong Yin, Shikun Feng, Yu Sun, Dianhai Yu, and Yin Zhang. 2022b. Ernie-mmlayout: Multi-grained multimodal transformer for document understanding.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. 2022a. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding.

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 *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery; Data Mining*, KDD '20, page 1192–1200, New York, NY, USA. Association for Computing Machinery.

Yiheng Xu, Tengchao Lv, Lei Cui, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, and Furu Wei. 2022b. XFUND: A benchmark dataset for multilingual visually rich form understanding. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3214–3224, Dublin, Ireland. Association for Computational Linguistics.

W. Yu, N. Lu, X. Qi, P. Gong, and R. Xiao. 2021. Pick: Processing key information extraction from documents using improved graph learning-convolutional networks. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4363–4370, Los Alamitos, CA, USA. IEEE Computer Society.

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. *arXiv preprint arXiv:2004.07464*.

Shuaicheng Zhang, Qiang Ning, and Lifu Huang. 2022a. Extracting temporal event relation with syntax-guided graph transformer. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 379–390, Seattle, United States. 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.

Yuhao Zhang, Peng Qi, and Christopher D. Manning. 2018. Graph convolution over pruned dependency trees improves relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2205–2215, Brussels, Belgium. Association for Computational Linguistics.

Zhenrong Zhang, Jiefeng Ma, Jun Du, Licheng Wang, and Jianshu Zhang. 2022b. Multimodal pre-training based on graph attention network for document understanding.

Hao Zhu, Yankai Lin, Zhiyuan Liu, Jie Fu, Tat-Seng Chua, and Maosong Sun. 2019. Graph neural networks with generated parameters for relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1331–1339, Florence, Italy. Association for Computational Linguistics.

Appendix

A Data Preprocessing

To accurately determine the layout heuristics, it is important to get the bounding box of the entire entity span. If token-level bounding boxes are provided, the boxes can be merged to obtain a span-level box. All the paragraph/tabular regions are detected and their bounding boxes are obtained. We identify the region an entity belongs to by checking the entity's Intersection over Union (IoU) with the regions and assign the region with the maximum IoU.

B Experiment Setting and Hyperparameters

NVIDIA A40 GPU was used for all finetuning tasks. To create paragraph-level regions using EasyOCR, a horizontal merging of text boxes is carried out when their distance is within 2, and vertical merging is done when the distance is within 1, while setting the paragraph flag to True. The model is trained end-to-end, with fine-tuning of the LayoutXLM base model. The eGAT layers and relation extraction head are trained from scratch, using 2 eGAT layers for all experiments. The training process entails 5000 steps with a batch size of 4, a learning rate of 5e-5, and a warm-up ratio of 0.1. The cross entropy loss is weighted at 1 and constraint loss is weighted at 0.02.

C Annotation Details and Statistics of DIVERSEFORM

The guidelines of annotating entities are as follows:

- Question: A word, set of words, or sentence worded or expressed so as to elicit information from the person filling the form.
  - Questions are annotated even if they haven't been answered
  - Questions and sub-questions are labeled as the same type of entity.
- Header: A word, set of words, or sentences worded or expressed so as give context or encapsulate a set of questions.
  - Annotate headers even if their questions haven't been answered
  - Headers do not have answers directly attached to them
- Answer: A word, set of words, or sentence written in response to a question.
  - Responses in the form of checkbox options count as answers.
  - In multiple choice type questions, all the options are annotated as answers (following FUNSD and XFUND)

The guidelines of annotating relations are as follows:

- Question-Answer: A link exists between a question entity and an answering entity when the answer is a response to a particular question.
  - When multiple answers exist for a question, there are multiple Question-Answer links from the same question entity.
  - Answers to a sub-question should only be linked to the sub-question and not the parent question.
- Question-Question: A link exists between a question entity and another question entity if one question is a sub-question of another question or one question is conditioned on the answer of another question.
  - For example, "If yes, . . . " type of question has a Question-Question link with the parent question.
  - A question that is split into multiple fine-grained questions has a Question-Question link between them. For example, "Address" can have further questions such as "Apt. No", "Street Name", "City", "State", "Zip Code".
- Header-Question: A link exists between a header entity and a question entity if the questions are present under the section or subsection that is characterized by the header.
  - If multiple questions exist under a header, there are multiple Header-Question links from the same header entity.
  - Often confused with Question-Question links and can be differentiated based on

layout structure, font style, and other visual aspects of the questions from the form.

The guidelines of annotation of tables are as follows: We mainly deal with one dimensional tables. For the case that each cell in the table is related to both row and column questions, there will be a Question-Question link between the questions extracted from the row and column, indicating that one question is a sub-question of another question or one question is conditioned on the answer of another question. This is part of the annotation guidelines for FUNSD and our own dataset. Based on these annotation rules, the constraint of one answer having one question still holds.

Figure 4 shows the distribution of domains in **DIVERSEFORM**. Miscellaneous consists of forms for voter registration, agriculture, scholarship, immigration, property tax, etc. Veteran's Affairs encompasses varying forms ranging from child support payments to retirement funds. There is rich layout variation within each domain shown in the chart. The number of entities and relations of each type in **DIVERSEFORM** are tabulated in Table 6.

D Significance Test

Table 7 shows the significance test results for both our approach and the best performing baseline (i.e., LiLT[InfoXLM]$_{\text{BASE}}$ (Wang et al., 2022a)) under the settings of language-specific fine-tuning and cross-lingual zero-shot transfer. The results for all experiments reported were averaged across 3 runs.

E Case Study

Figure 5 visualizes paragraph-level regions, tabular regions, and predictions for a Portuguese form in FUNSD. It shows that paragraph-level regions are suitable for the top portion of the form, while tabular regions specifically pertain to the bottom table. In this particular form, the decision was made to adopt paragraph-level regions, resulting in the exclusion of the tabular layout despite its ability to convey more information. We acknowledge that there are instances where our proposed approach may struggle to accurately distinguish between paragraph-level and tabular regions, leading to a performance decrease.

F Visualizations of Ablation Results

Figure 6 shows the visualization of predictions of the ablation study of node and edge embeddings. Figure 7 shows the visualization of predictions of the ablation study of incorporating paragraph/tabular regions.

G Ablation Results of Constraint Loss

The constraint loss has been modeled to encourage each answer entity to be linked to at most one question. Table 8 shows that incorporating the constraint loss significantly improves the F1 score of $\mathbf{RE}^2$, especially precision.

H Pseudocode

The following pseudocode extracts the tabular and paragraph-level regions from a VRD.

Algorithm 1 IdentifyHorizontalAndVerticalLines(image)

```
Apply horizontal kernel to the image
Apply vertical kernel to the image
Find horizontal lines
Find vertical lines
return Combined horizontal and vertical lines
```

Algorithm 2 FindBoundingBoxes(lines)

```
TabularBoxList = []
Find contours in lines
for each contour do
    Compute the bounding box
    Append the box to TabularBoxList
end for
return TabularBoxList
```

Algorithm 3 SortBoxesByArea(boundingBoxes)

```
Sort the bounding boxes by area in increasing order
return boundingBoxes
```

Algorithm 4 AppendBoxToList(boundingBoxes, text)

```
FinalBoxList = []
for each box in boundingBoxes do
    if the box contains any text and has no intersection with existing boxes in FinalBoxList then
        Append the box to FinalBoxList
    end if
end for
return FinalBoxList
```

Algorithm 5 CheckAllTextPresent(FinalBoxList, text)

```
if all the text in the document is present in the boxes in FinalBoxList then
    return True
else
    return False
end if
```

Algorithm 6 GetMissingText(FinalBoxList, text)

```
missingText = []
if the text is not present inside the bounding boxes of any of the FinalBoxList then
    Append text to missingText
end if
return missingText
```

Algorithm 7 AppendMissingTextBoxes(FinalBoxList, missingText, ParagraphRegions)

```
for each missing text in missingText do
    if missing text is present in any paragraph region in ParagraphRegions then
        Append paragraph region to FinalBoxList
    end if
end for
return FinalBoxList
```

	Entities			Relations		
Split	Question	Answer	Header	Question-Answer	Question-Question	Header-Question
Training	3,087	3,585	230	1,172	594	546
Test	956	1,048	57	520	270	164

Table 6: Statistics of entities and relations in **DIVERSEFORM**

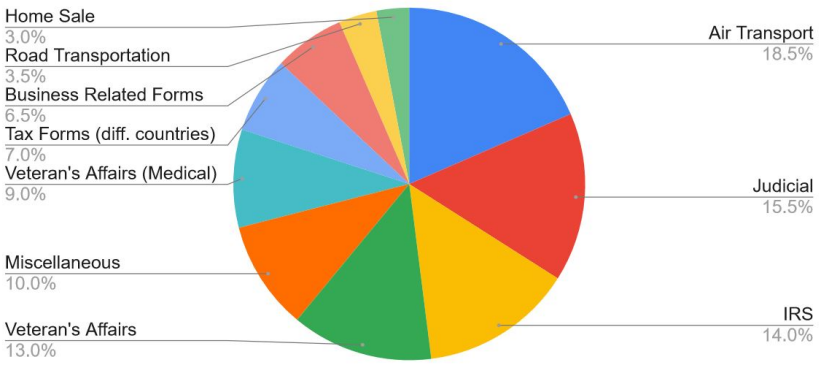
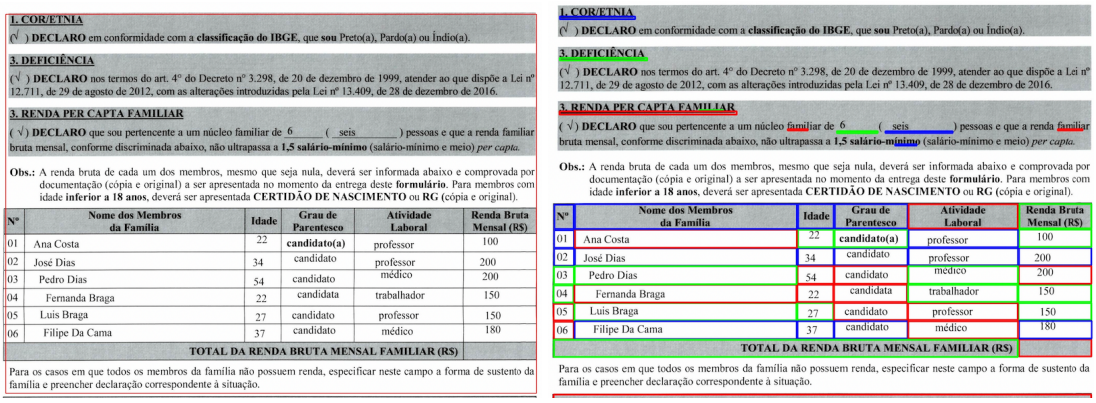


Figure 4: Domain distribution of **DIVERSEFORM**.



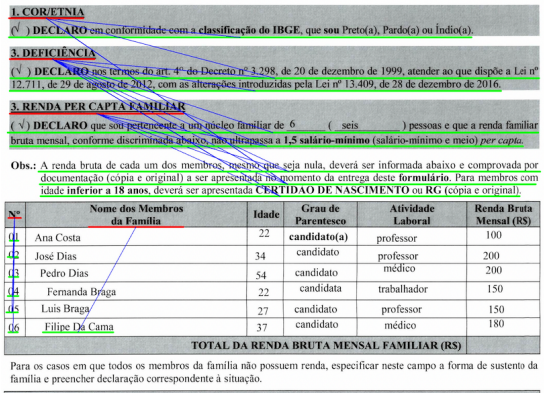
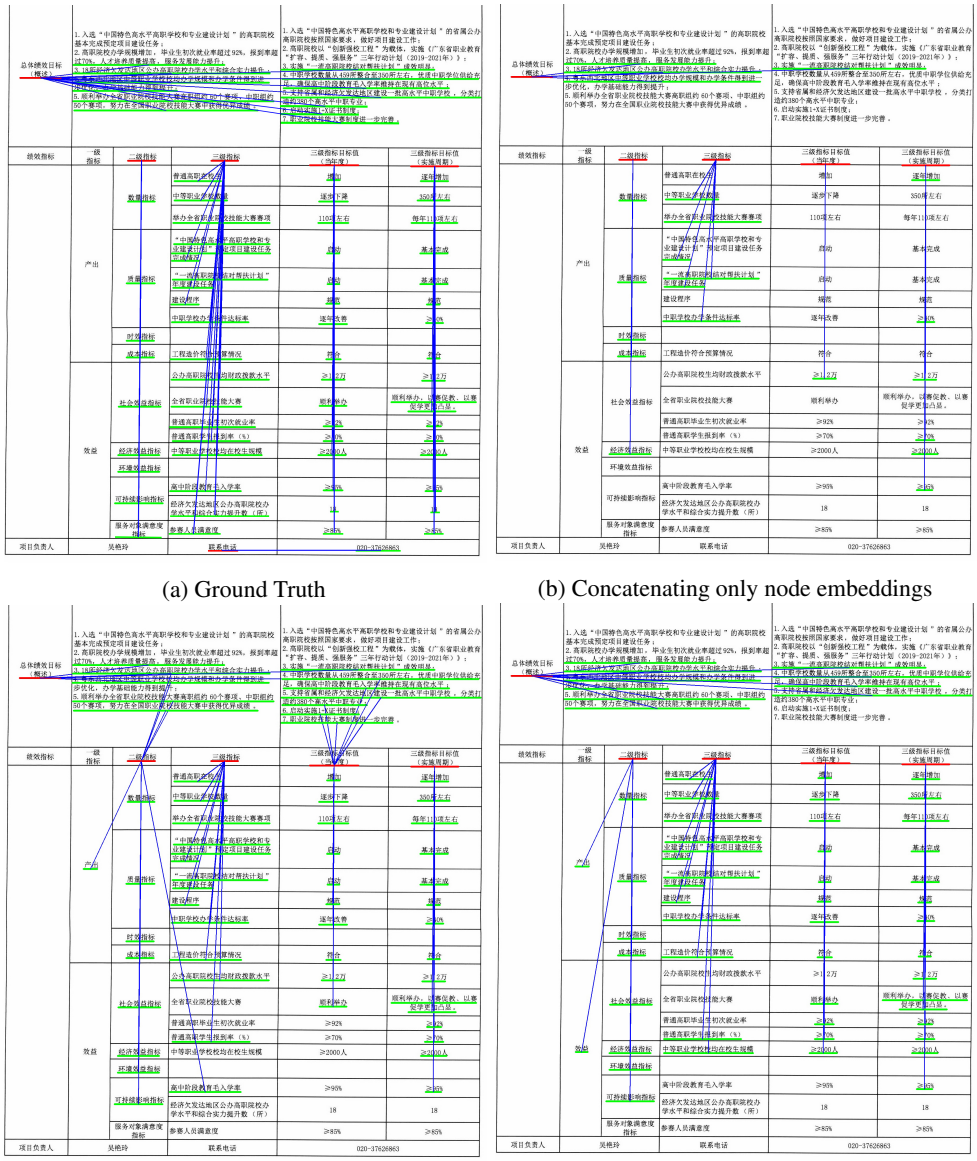


Figure 5: Visualization of paragraph-level regions (a), tabular regions (b) and predictions (c) for a Portuguese form in XFUND.



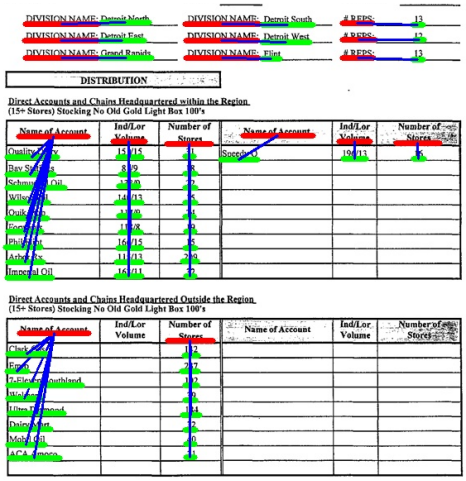
(a) Ground Truth (b) Concatenating only node embeddings

(c) Concatenating only edge embeddings (d) Concatenating node & edge embeddings

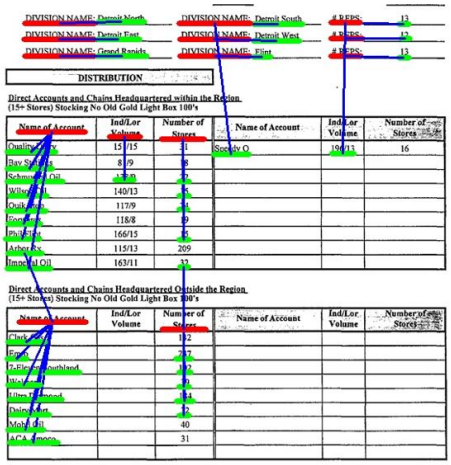
Figure 6: Visualization of predictions of the ablation study of node and edge embeddings, where red lines denote the question span, green lines denote the answer span, and blue lines denote the question answer relation predictions.

Algorithm 8 GetParagraphTabularRegions(imageFile)

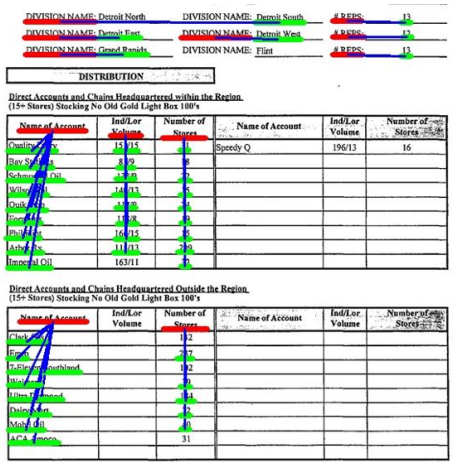
```
image = LoadImage(imageFile)
text = OCR(imageFile)
lines = IdentifyHorizontalAndVerticalLines(image)
boundingBoxes = FindBoundingBoxes(lines)
boundingBoxes = SortBoxesByArea(boundingBoxes)
FinalBoxList = AppendBoxToList(boundingBoxes, text)
if CheckAllTextPresent(FinalBoxList, text) then
    OutputResult(FinalBoxList)
else
    ParagraphRegions = GetParagraphRegionsFromEasyOCR(image)
    missingText = GetMissingText(FinalBoxList, text)
    FinalBoxList = AppendMissingTextBoxes(FinalBoxList, missingText, ParagraphRegions)
    OutputResult(FinalBoxList)
end if
```



(a) Ground Truth



(b) Predictions without paragraph/tabular region information



(c) Predictions with paragraph/tabular region information

Figure 7: Visualization of predictions of the ablation study of incorporating paragraph/tabular regions.

	RE^2		Baseline		
Setting	Mean	SD	Mean	SD	P-value
Language-Specific Fine-Tuning	73.98	6.15	67.81	4.98	0.0447
Zero-Shot Cross-Lingual	62.08	8.53	49.30	6.55	0.0047

Table 7: Significance Test Results

Model	EN			ZH			JA			ES		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
RE^2	69.71	73.74	71.67	76.80	82.62	79.60	70.16	81.39	75.36	70.26	81.78	75.59
RE^2- constraint loss	58.76	82.16	68.52	74.77	81.01	77.77	69.13	80.75	74.49	69.41	81.05	74.78
Model	FR			IT			DE			PT		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
RE^2	70.05	83.97	76.38	74.34	80.83	77.45	73.01	78.94	75.86	48.06	78.99	59.76
RE^2- constraint loss	71.54	80.57	75.79	72.32	79.32	75.66	71.74	75.59	73.61	46.98	74.02	57.48

Table 8: Precision, Recall and F1 score of ablation study of Constraint Loss on **RE**^{2}