Document Information Extraction via Global Tagging

Shaojie He[1,2], **Tianshu Wang**[2], **Yaojie Lu**[2], **Hongyu Lin**[2]*, **Xianpei Han**[2]*, **Yingfei Sun**[1], **Le Sun**[2]
[1]University of Chinese Academy of Sciences, Beijing, China
[2]Chinese Information Processing Laboratory,
Institute of Software, Chinese Academy of Sciences, Beijing, China
{heshaojie2020,tianshu2020,luyaojie,hongyu,xianpei}@iscas.ac.cn
yfsun@ucas.ac.cn, sunle@iscas.ac.cn

Abstract

Document Information Extraction (DIE) is a crucial task for extracting key information from visually-rich documents. The typical pipeline approach for this task involves Optical Character Recognition (OCR), serializer, Semantic Entity Recognition (SER), and Relation Extraction (RE) modules. However, this pipeline presents significant challenges in real-world scenarios due to issues such as unnatural text order and error propagation between different modules. To address these challenges, we propose a novel tagging-based method – Global TaggeR (GTR), which converts the original sequence labeling task into a token relation classification task. This approach globally links discontinuous semantic entities in complex layouts, and jointly extracts entities and relations from documents. In addition, we design a joint training loss and a joint decoding strategy for SER and RE tasks based on GTR. Our experiments on multiple datasets demonstrate that GTR not only mitigates the issue of text in the wrong order but also improves RE performance.

1 Introduction

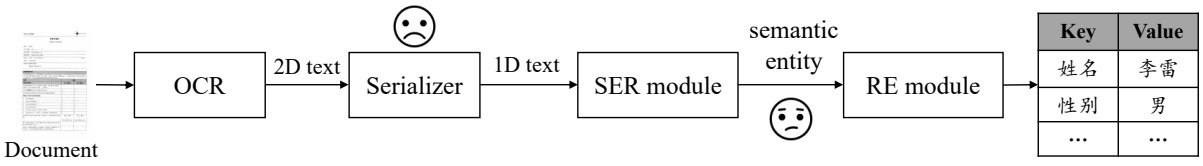
Document Information Extraction (DIE), which is to extract key information from document with complex layouts, has become increasingly important in recent years (Zhang et al., 2022; Hong et al., 2022). It not only enables us to efficiently compress document data, but also facilitates the retrieval of important information from documents. A typical pipeline approach for the DIE task is depicted in Figure 1(a) (Denk and Reisswig, 2019; Hwang et al., 2021a). First, the document with complex layout is converted into text blocks using Optical Character Recognition (OCR) tools. Next, the serializer module organizes these text blocks into a more appropriate order. Finally, the well-ordered text blocks are input sequentially into the Semantic Entity Recognition (SER) and Relation Extraction (RE) modules to extract key-value pairs.

However, the pipeline approach in Figure 1(a) presents significant challenges in real-world scenarios. (1) Mainstream models for the DIE task, such as LayoutLM (Xu et al., 2020), LayoutLMv2 (Xu et al., 2021) and LayoutXLM (Xu et al., 2022), usually use sequence labeling in the Beginning-Inside–Outside (BIO) tagging schema, which assume that tokens belonging to the same semantic entity are grouped together after serialization. If the serializer module fails to order the text blocks correctly, the final performance can be severely impacted. A potential solution is to train a strong and robust serializer module, but this is difficult due to the labor-intensive labeling process under rich and diverse styles of documents; (2) In addition to the issue of text order, this pipeline also suffers from error propagation when using a SER module and a RE module. In research settings, the results of the SER and RE tasks are generally tested separately, with the ground truth of the SER results being used as default auxiliary information for the RE task. However, in real-world scenarios, the SER module in the pipeline cannot provide 100% accurate results, which ineluctably leads to error propagation on RE performance.

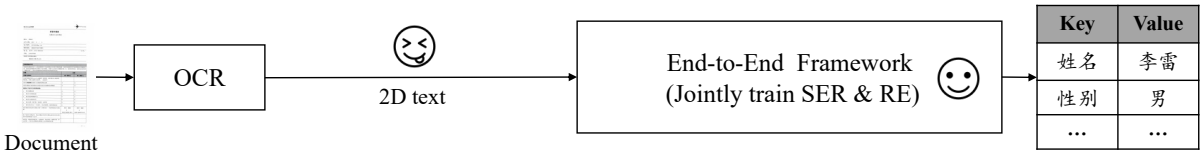
Researchers have explored alternative methods for modeling OCR results directly without serializer module to tackle the issue of text in the wrong order. Some have utilized graph convolution networks to

*Corresponding author.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 726-735, Harbin, China, August 3 - 5, 2023.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China



(a) The pipeline approach in the DIE task.



(b) The end-to-end approach in this paper.

Figure 1: A comparison between (a) current pipeline approach and (b) our end-to-end approach.

model the relationships between tokens (Yu et al., 2020; Lee et al., 2022; Wei et al., 2020) . Others have converted the DIE task into a parsing problem, modeling tree structure for the document (Hwang et al., 2021b; Mathur et al., 2023). Besides, generative encoder-decoder frameworks are applied to avoid the weakness of the BIO tagging schema essentially (Kim et al., 2022). While these methods can mitigate the problem of text in the wrong order, they still face challenges. For example, graph-based methods require a more delicate model design, and generative models are usually difficult to train and require a great amount of document data for pre-training.

To address abovementioned two problems, we propose a simple yet effective method named Global TaggeR (GTR). Our approach is inspired by Wu et al. (2020), which converts the original sequence labeling task into a token relation classification task. For the SER task, we tag all token pairs and design a decoding strategy based on disjoint sets to decode the semantic entities. And we find GTR naturally resistant to wrong text order to a certain extent. For example, there is a document fragment "登记表姓名李雷性别男" and we tag the token pair {姓, 名} so that we know "姓名" is a semantic entity. Even if we shuffle this fragment to "登记表姓李性男名雷别", we still can know "姓名" is a semantic entity using the same tag {姓, 名}. In other words, GTR enables us to recognize discontinuous semantic entities, regardless of text in the wrong order. Additionally, for the RE task, we combine RE and SER tags for joint training and extend the decoding strategy for joint decoding. The pipeline of this study is depicted in Figure 1(b). We remove the serializer module from the original pipeline to make it easier and propose an end-to-end extraction framework for jointly training the SER and RE tasks to prevent error propagation problem. The contributions of this work are summarized as follows:

- We propose an end-to-end extraction framework for the document information extraction, which simplifies the traditional pipeline approach and alleviates error propagation issues.
- In this end-to-end extraction framework, we propose the Global TaggeR (GTR) method, which contains a global tagging schema and a joint decoding strategy for the SER and RE tasks.
- Our experiments on multiple datasets demonstrate that the GTR proposed not only mitigates the issue of text in the wrong order but also facilitates the interaction of entity and relation information, resulting in improvement of RE performance.

2 Background

2.1 Task Definition

Given a document image I and its OCR results that containing a sequence of tokens $S = \{t_1, ..., t_n\}$ paired with corresponding bounding boxes $L = \{b_1, ..., b_n\}$, the goal of the DIE task is to extract a set of entities $E = \{e_1, ..., e_m\}$ in the document and their corresponding relations $R = \{(e_i, e_j)\}$. We usually divide the DIE task into two sub-tasks named SER and RE. For the SER task, we try to recognize

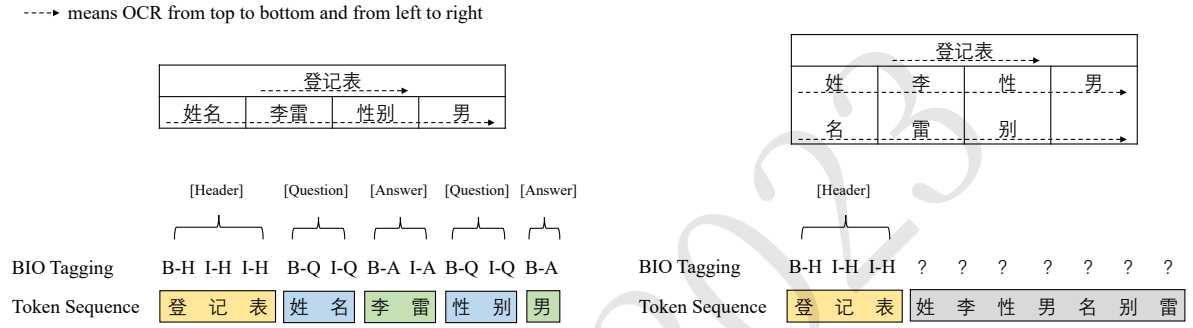
Proceedings of the 22nd China National Conference on Computational Linguistics, pages 726-735, Harbin, China, August 3 - 5, 2023.

all possible semantic entities in token sequence S and classify them with three entity types {[Header], [Question], [Answer]}. For the RE task, based on semantic entities that we have recognized, we match each two of them if they are question-answer pairs, or key-value pairs. The relations only have two types, paired or not.

2.2 LayoutXLM

We choose LayoutXLM (Xu et al., 2022) as our baseline model, which is a multilingual and multi-modal pre-trained language model designed with a single encoder architecture. The model first feeds token sequence S and bounding box sequence L, along with visual features extracted from document image I. Next, it adopts visual and text embedding, position embedding and layout embedding as the representation of tokens, and then employs multi-modal Transformer encoder layers to generate the representations of the given tokens $H = \{h_1, ..., h_n\}$. Finally, a simple classifier is connected to the encoder, enabling it to perform downstream SER and RE tasks.

2.3 BIO Tagging Schema



(a) The BIO tagging schema successfully tags when the input text is well-ordered.

(b) The BIO tagging schema fails to tag when the order of the input text is incorrect.

Figure 2: The illustration of the BIO tagging schema.

The BIO tagging schema, which is a popular sequence labeling technique, is widely used for the SER task. In this schema, each token in the document is labeled with a prefix that indicates whether it is the beginning (B), inside (I), or outside (O) of an entity span. Figure 2(a) provides a simple illustration. However, layout-rich documents often result in OCR text in the wrong order. Given text in the wrong order, the BIO tagging schema cannot express span boundaries correctly, illustrated in Figure 2(b). Therefore, it is necessary to find new approaches to tackle this issue.

3 Approach

In this section, we introduce our GTR approach in four parts. First, we propose the global tagging schema of the DIE task. Next, a token pair scoring layer added to baseline model is proposed. Then, we design a corresponding decoding strategy to decode entities and relations from the predicted tagging matrix. Finally, we introduce our training loss for jointly training SER and RE tasks.

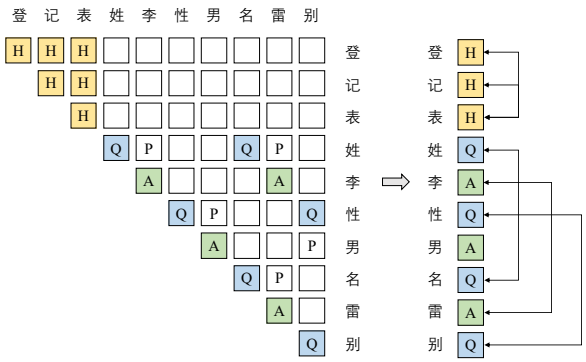
3.1 Global Tagging Schema

For the DIE task, we use five tags {O, H, Q, A, P} to represent relations between token t_i and t_j. Table 1 shows the meanings of these five tags.

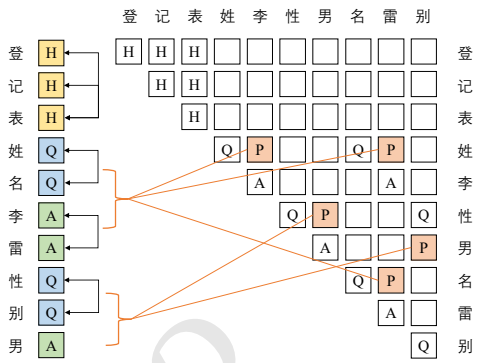
Figure 3(a) illustrates the global tagging schema tags entities that are difficult to tag using the BIO tagging schema in Figure 2(b). Tokens in the same semantic entity are tagged with the same label pairwise. The labels are {H, Q, A}, representing the entity types {[Header], [Question], [Answer]}, respectively. For example, in Figure 3(a), the token pair {姓, 名} = Q means that the tokens "姓" and "名" belong to the same entity span, and the entity type is [Question]. Similarly, {登, 记, 表} belongs

Tags	Meanings
H	Token t_i and t_j belong to the same entity span, and the entity type is [Header].
Q	Token t_i and t_j belong to the same entity span, and the entity type is [Question].
A	Token t_i and t_j belong to the same entity span, and the entity type is [Answer].
P	Token t_i and t_j belong to two paired entities, with the types of [Question] and [Answer].
O	No above four relations for token t_i and t_j.

Table 1: The meanings of tags for the DIE task.



(a) SER task. The left is the tagging matrix and the right is the corresponding semantic entities.



(b) RE task. The right is the tagging matrix and the left is the rearranged semantic entities.

Figure 3: The illustration of global tagging schema for jointly labeling (a) SER and (b) RE tasks. We only display the upper triangular of tagging matrix on account of its symmetry.

to the [Header] entity, {性, 别} belongs to the [Question] entity, and {李, 雷}, {男} belong to the [Answer] entity.

Figure 3(b) illustrates the global tagging schema tags relations after tagging entities. For each [Question]-[Answer] (QA) relation in the document, tokens from the two associated entities, are tagged with the same label P pairwise. For example, given the premise that {姓, 名} belongs to [Question] entity and {李, 雷} belongs to [Answer] entity, the token pairs {姓, 李}, {姓, 雷}, {名, 雷} = P, indicating that {姓, 名} and {李, 雷} are paired QA relation. Similarly, {性, 别} and {男} are paired QA relation.

The global tagging schema offers two primary advantages in the DIE task. (1) First, it allows for the tagging of discontinuous semantic entity spans. Due to the diversity of document layouts, the token sequence produced by OCR tools is usually in an incorrect order. Even if the tokens in the same semantic entity span are discontinuous in the token sequence, they can still be tagged using this global tagging schema. (2) Second, it supports joint training of the SER and RE tasks. Using the global tagging schema, the SER task can be expanded to token-to-token relationship classification task. This schema unifies task format and enables unified modeling and joint training for the SER and RE tasks.

3.2 Token Pair Scoring

For the representations $H = \{h_1, ..., h_n\}$ generated from given token sequence S, we employ simple linear transformation and multiplication operation to obtain the global score $s_{ij|c}$ of token pair t_i and t_j classified to class c:

$$q_{i,c} = W_{q,c}h_i + b_{q,c} \tag{1}$$

$$k_{j,c} = W_{k,c}h_j + b_{k,c} \tag{2}$$

$$s_{ij|c} = (\mathcal{R}_i q_{i,c})^T (\mathcal{R}_j k_{j,c}) \tag{3}$$

where $q_{i,c}$ and $k_{j,c}$ are intermediate representations created by linear transformation operation. $\mathcal{R}$ is a rotary position embedding (Su et al., 2021), which helps to embed relative position information and accelerate training process.

During the training stage, we directly use $s_{ij|c}$ to compute loss function. For the supervision signal of $s_{ij|c}$, we assign signal 1 to represent {H, Q, A, P} tags and signal -1 to represent the absence of any of the above four relations. Therefore, during the inference stage, we obtain the predicted tagging matrix by processing $s_{ij|c}$ with a threshold of 0, where values greater than 0 are regarded as tags.

3.3 Decoding Strategy

With the predicted tagging matrix, we design a decoding strategy to extract the semantic entities and relations, as shown in Algorithm 1. Following the proposed decoding strategy, we decode in two steps:

Algorithm 1 Decoding Strategy for DIE

Input: The predicted tagging matrix T. The predicted tag of token pair t_i and t_j is denoted as $T(t_i, t_j)$. The predicted tag of token pair t_i and t_i is abbreviated as $T(t_i)$. If all tokens in a set e share the same tag, abbreviated as $T(e)$.

Output: Entity set E and relation set R.

Initialize the entity set E and relation set R with $\varnothing$, and $n \leftarrow len(S)$.
while $i \leq n$ **do**
  if $T(t_i) \in \{\mathrm{H}, \mathrm{Q}, \mathrm{A}\}$ **then**
    $E \leftarrow E \cup \{t_i\}$
  end if
end while
while $i \leq n$ and $j \leq n$ **do**
  if $i \neq j$ **and** $T(t_i, t_j) \in \{\mathrm{H}, \mathrm{Q}, \mathrm{A}\}$ **and** $T(t_i, t_j) = T(t_i) = T(t_j)$ **then**
    $E \leftarrow$ Merge the set where t_i resides and the set where t_j resides in E.
  end if
end while
while $e_i \in E$ and $e_j \in E$ **do**
  if $T(e_i) = \mathrm{Q}$ and $T(e_j) = \mathrm{A}$ and any $T(t_k, t_l) = \mathrm{P}$ that $t_k \in e_i$ and $t_l \in e_j$ **then**
    $R \leftarrow R \cup \{(e_i, e_j)\}$
  end if
end while
return the set E and the set R

SER. Firstly, we recognize the diagonal tags, and use these tags to label the token sequence S. Then, we recognize the non-diagonal tags belonging to {H, Q, A}, and use these tags for merging tokens. Iterating through these tags, we use a disjoint set algorithm with additional judgement to merge semantic entity tokens. Therefore, we can extract the semantic entity set E.

RE. Using the semantic entity set E, we iterate through all possible `[Question]`-`[Answer]` entity pairs. If there exists any token pair t_k and t_l that t_k in an entity e_i with type `[Question]` and t_l in an entity e_j with type `[Answer]` and $T(t_k, t_l)$ is tagged with label P, we add (e_i, e_j) into relation set R. Finally, we can extract the semantic entity set E as well as the relation set R.

3.4 Training Loss

For token pair t_i and t_j, we denote y_{ij} as the ground truth tag and $P_{ij}(\hat{y} = k)$ as the predicted probability for class k. A cross entropy loss is applied:

$$\mathcal{L} = -\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k\in C}\mathbb{I}(y_{ij} = k)\log P_{ij}(\hat{y} = k), \quad P_{ij}(\hat{y} = k) = \frac{e^{s_{ij|k}}}{\sum\limits_{k'\in C} e^{s_{ij|k'}}} \tag{4}$$

where $\mathbb{I}$ is an indicator function and C is the label set {H, Q, A, P, O}. And $s_{ij|k}$ denotes the predicted score for token pair t_i and t_j classified to class k.

We attempt to train the baseline model using the above loss function but fail due to convergence issues. And the training results always output O tags. We suggest that our global tagging schema requires the

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 726-735, Harbin, China, August 3 - 5, 2023.

prediction of a probability matrix of $n*n$, which results in very sparse supervised signals, facing a severe class imbalance problem, and making it challenging to train the model effectively. Inspired by Su et al. (2022), we improve $\log P_{ij}(\hat{y}=k)$ with a class imbalance likelihood:

$$\log P_{ij}(\hat{y}=k) = \log(1+e^{-s_{ij|k}}) + \log(1+\sum_{k' \in C, k' \neq k} e^{s_{ij|k'}}) \quad (5)$$

which turns loss into a pairwise comparison of target category scores and non-target category scores.

4 Experiments

4.1 Experimental Setup

Dataset. We use FUNSD (Jaume et al., 2019) and XFUN (Xu et al., 2022) datasets to evaluate our proposed approach. (1) FUNSD is an English dataset for document understanding, comprising 199 annotated documents. The dataset is split into a training set of 149 documents and a testing set of 50 documents; (2) XFUN is a multilingual dataset for document understanding that comprises seven languages [Chinese (ZH), Japanese (JA), Spanish (ES), French (FR), Italian (IT), German (DE), Portuguese (PT)], totaling 1,393 annotated documents. Each language's data has separate training and testing sets, with 199 and 50 documents respectively.

Parameter Settings. For training, we follow the hyper-parameter settings of Xu et al. (2022), setting the learning rate to 5e-5 and the warmup ratio to 0.1. The max length of input token sequence is set to 512, which means a split of chunk size 512 if the input token sequence is too long. For a fair comparison, we set the batch size to 64 and run the training for 2000 steps to ensure that the models have well converged.

Input Settings. Golden input and OCR input are two types of input text order for experiment input settings. (1) Golden input means that we concatenate the ground truth text blocks into a token sequence and feed it into the model, which implies that all semantic entity spans are continuous. (2) OCR input means that we concatenate all tokens following the recognition pattern of a common OCR from top to bottom and left to right before feeding them into the model. This implies that under complex layouts, the same semantic entity span may be discontinuous.

Evaluation Metrics. For evaluation, we use F1-score on two sub-tasks: (1) Semantic Entity Recognition (SER), where semantic entities are identified by tagging as either {`[Header]`, `[Question]`, `[Answer]`}. When the entity type and all entity tokens are correct, the entity is regarded as a correct entity. (2) Relation Extraction (RE), where paired relation of question and answer entities are identified. We use a strict evaluation metrics that only the paired two entities are exactly correct at the token-level, the relation is regarded as a correct relation.

Baseline Model. We use LayoutXLM$_{\text{BASE}}$ model as the baseline model. Its original RE results are tested based on the given ground truth semantic entities. To test the RE results in the pipeline for baseline model, we first reproduce the results of Xu et al. (2022) and then re-test the RE results using the semantic entities generated by its SER module.

4.2 Result

We evaluate the baseline model with the BIO tagging and the global tagging on language-specific fine-tuning settings (training on X, and testing on X).

Table 2 presents the results under Golden input settings. We compare our global tagger approach with the reproduced baseline. The results show that our global tagger method outperforms the baseline model on average F1-score of the 8 languages for the SER task. Moreover, when combining the SER and RE tasks in an end-to-end extraction framework, the RE performance of average F1-score significantly surpassed that of the baseline model pipelined, and is even higher on two languages compared with the baseline model using ground truth semantic entity information .

Table 3 presents the result under OCR input settings. We directly use the baseline model trained under Golden input settings to predict the SER and RE results for evaluating the BIO tagging schema. We observe that the SER performance on average F1-score of the 8 languages for the baseline model is

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 726-735, Harbin, China, August 3 - 5, 2023.

	Model	FUNSD	ZH	JA	ES	FR	IT	DE	PT	Avg.
SER	BIO♣	0.7940	0.8924	0.7921	0.7550	0.7902	0.8082	0.8222	0.7903	0.8056
	BIO	0.8013	**0.8944**	0.7864	0.7426	0.7852	0.8073	0.7951	0.7848	0.7996
	GTR	**0.8079**	0.8818	**0.7972**	**0.7631**	**0.8067**	**0.8210**	**0.8032**	**0.8071**	**0.8110**
gtSER+RE	BIO♣	0.5483	0.7073	0.6963	0.6896	0.6353	0.6415	0.6551	0.5718	0.6432
	BIO	0.5560	0.7047	0.6519	0.7041	0.6664	0.6725	0.6485	0.5893	0.6492
SER+RE	BIO	0.4340	0.5965	0.5082	0.498	0.5064	0.4861	0.4258	0.3765	0.4789
	GTR	**0.5910**	**0.7739**	**0.6470**	**0.5363**	**0.6063**	**0.6594**	**0.5531**	**0.5247**	**0.6115**

Table 2: Main result under Golden input settings. ♣: results reported in Xu et al. (2022). Best results are in **bold** comparing reproduced BIO tagging (abbreviated as BIO) with our global tagger (abbreviated as GTR). gtSER+RE denotes evaluating the RE results using ground truth SER results. And SER+RE denotes evaluating the RE results using SER results of the model.

	Model	FUNSD	ZH	JA	ES	FR	IT	DE	PT	Avg.
SER	BIO	0.5735	0.3970	0.4017	0.6287	0.6916	0.7055	0.6823	0.6863	0.5958
	GTR	**0.7412**	**0.8444**	**0.7205**	**0.7165**	**0.7676**	**0.7772**	**0.7508**	**0.7811**	**0.7624**
SER+RE	BIO	0.2712	0.1441	0.1759	0.3665	0.4141	0.4206	0.3566	0.3086	0.3072
	GTR	**0.5828**	**0.6920**	**0.5427**	**0.5686**	**0.5712**	**0.5888**	**0.5933**	**0.5580**	**0.5872**

Table 3: Main result under OCR input settings. Best results are in **bold** comparing reproduced BIO tagging (abbreviated as BIO) with our global tagger (abbreviated as GTR).

significantly impacted, making it difficult to perform the RE process based on its SER results. However, with joint training and decoding using our global tagger approach, we are able to alleviate this issue.

4.3 Analysis

4.3.1 Golden Input vs. OCR Input

The BIO tagging schema requires well-ordered input, while the global tagging schema accepts unordered input. Comparing the average F1-score of SER performance in Table 2 and Table 3, we observe a significant drop from 0.7996 to 0.5958 when changing Golden input settings into OCR input, indicating a great impact by the order of input tokens using the BIO tagging schema. On the other hand, under the global tagging schema, we find the model's average SER performance only drops from 0.8110 to 0.7624 between Golden input settings and OCR input settings, demonstrating that the global tagging schema can effectively alleviate suboptimal input token order issue.

4.3.2 Pipeline Framework vs. End-to-End Framework

The pipeline framework with the BIO tagging schema suffers from error propagation, while the end-to-end GTR method can greatly mitigate it. In Table 2, we observe a drop of average F1-score on the RE results from 0.6492 to 0.4789 when combining the SER and RE modules in the pipeline, demonstrating that pipeline framework can greatly impact performance. Particularly, when both SER and RE modules have poor performance under OCR input settings, we observe a terrible performance, which is only 0.3072 average F1-score on the RE task. In such case, joint training and decoding in GTR method can significantly alleviate error propagation issue with the average F1-score of 0.5872 rather than 0.3072 of 8 language datasets on the RE task.

Besides, in Table 2, the SER+RE results using GTR approach are even higher than the baseline RE results using ground truth semantic entities on English(FUNSD) and Chinese(ZH) language datasets, indicating that the end-to-end extraction framework is potential for facilitating the interaction of entity and relation information, resulting in better RE performance.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 726-735, Harbin, China, August 3 - 5, 2023.

5 Related Work

In recent years, benefited from pre-training and fine-tuning paradigm, information extraction for documents has gained significant attention in both research and industry (Li et al., 2021b; Li et al., 2021a; Appalaraju et al., 2021; Wang et al., 2022; Li et al., 2022; Sun et al., 2023). However, there are still numerous challenges in the pipeline when applied in real-world scenarios. Addressing the issue of text order in the pipeline, related works are organized into three perspectives.

5.1 Sequence-based Perspective

Sequence-based models, such as LayoutLM (Xu et al., 2020) and LayoutLMv2 (Xu et al., 2021), aim to encode serialized token sequence from complex and diverse document, integrating layout, font, and other features. These models offer several advantages, such as simplicity, scalability, and suitability for Masked Language Modeling (MLM) pre-training. However, these models are constrained by the traditional BIO tagging mode and require a well-ordered token sequence as a basis.

5.2 Graph-based Perspective

Graph-based models usually treat tokens as nodes in a graph and allow interactions between tokens explicitly to enhance their representations (Yu et al., 2020; Lee et al., 2022; Wei et al., 2020). Even though these models leverage the graph structure to capture more complex relationships between entities, they still use the BIO paradigm for SER task. Alternatively, some works, like SPADE (Hwang et al., 2021b), take a different approach by converting DIE task into document parsing task. It models the document as a dependency tree to represent entities and relations.

Our work in this paper also lies in graph-based perspectives. Similar to the tack of SPADE that converting the DIE task to a different task, we view the DIE task as a token relation classification task. But unlike SPADE, we do not utilize a graph generator and graph decoder. Rather, we simply modify the tagging schema and do not change the encoding model.

5.3 End-to-End Perspective

End-to-End model typically combines the entire pipeline into one model. Dessurt (Davis et al., 2022), TRIE++ (Cheng et al., 2022), for example, unify OCR, reordering, and extraction into a single model. Meanwhile, models like Donut (Kim et al., 2022), GMN (Cao et al., 2022), use a generative encoder-decoder architecture to unify OCR and generation. In contrast to extraction-based works, they directly generate the structured output, making it more flexible for varying output formats.

6 Conclusion

In this paper, we propose an end-to-end approach named global tagger to solve the document information extraction task. Experiments on the FUNSD and XFUN datasets demonstrate its efficacy in effectively mitigating the gap between token order in OCR input and golden input. Furthermore, our experimental results indicate that joint training and decoding of semantic entity recognition and relation extraction tasks in this end-to-end extraction framework can alleviate the negative impact of error propagation and improve the performance of the relation extraction results.

Acknowledgements

We sincerely thank the reviewers for their insightful comments and valuable suggestions. This research work is supported by the National Natural Science Foundation of China under Grants no. U1936207, 62122077 and 62106251.

References

Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R. Manmatha. 2021. Docformer: End-to-end transformer for document understanding. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 973–983. IEEE.

Haoyu Cao, Jiefeng Ma, Antai Guo, Yiqing Hu, Hao Liu, Deqiang Jiang, Yinsong Liu, and Bo Ren. 2022. GMN: generative multi-modal network for practical document information extraction. In Marine Carpuat, Marie-Catherine de Marneffe, and Iván Vladimir Meza Ruíz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 3768–3778. Association for Computational Linguistics.

Zhanzhan Cheng, Peng Zhang, Can Li, Liang Qiao, Yunlu Xu, Pengfei Li, Shiliang Pu, Yi Niu, and Fei Wu. 2022. TRIE++: towards end-to-end information extraction from visually rich documents. *CoRR*, abs/2207.06744.

Brian L. Davis, Bryan S. Morse, Brian L. Price, Chris Tensmeyer, Curtis Wigington, and Vlad I. Morariu. 2022. End-to-end document recognition and understanding with dessurt. In Leonid Karlinsky, Tomer Michaeli, and Ko Nishino, editors, *Computer Vision - ECCV 2022 Workshops - Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part IV*, volume 13804 of *Lecture Notes in Computer Science*, pages 280–296. Springer.

Timo I. Denk and Christian Reisswig. 2019. BERTgrid: Contextualized embedding for 2d document representation and understanding. In *Workshop on Document Intelligence at NeurIPS 2019*.

Teakgyu Hong, DongHyun Kim, Mingi Ji, Wonseok Hwang, Daehyun Nam, and Sungrae Park. 2022. BROS: A pre-trained language model focusing on text and layout for better key information extraction from documents. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10767–10775.

Wonseok Hwang, Hyunji Lee, Jinyeong Yim, Geewook Kim, and Minjoon Seo. 2021a. Cost-effective end-to-end information extraction for semi-structured document images. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *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 3375–3383. Association for Computational Linguistics.

Wonseok Hwang, Jinyeong Yim, Seunghyun Park, Sohee Yang, and Minjoon Seo. 2021b. Spatial dependency parsing for semi-structured document information extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 330–343, Online. 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 *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 1–6.

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 Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *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.

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. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3735–3754. Association for Computational Linguistics.

Chenliang Li, Bin Bi, Ming Yan, Wei Wang, Songfang Huang, Fei Huang, and Luo Si. 2021a. Structurallm: Structural pre-training for form understanding. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *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 6309–6318. Association for Computational Linguistics.

Yulin Li, Yuxi Qian, Yuechen Yu, Xiameng Qin, Chengquan Zhang, Yan Liu, Kun Yao, Junyu Han, Jingtuo Liu, and Errui Ding. 2021b. Structext: Structured text understanding with multi-modal transformers. In Heng Tao Shen, Yueting Zhuang, John R. Smith, Yang Yang, Pablo César, Florian Metze, and Balakrishnan Prabhakaran, editors, *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pages 1912–1920. ACM.

Qian Li, Hao Peng, Jianxin Li, Jia Wu, Yuanxing Ning, Lihong Wang, Philip S. Yu, and Zheng Wang. 2022. Reinforcement learning-based dialogue guided event extraction to exploit argument relations. *IEEE ACM Trans. Audio Speech Lang. Process.*, 30:520–533.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 726-735, Harbin, China, August 3 - 5, 2023.

Puneet Mathur, Rajiv Jain, Ashutosh Mehra, Jiuxiang Gu, Franck Dernoncourt, Anandhavelu Natarajan, Quan Hung Tran, Verena Kaynig-Fittkau, Ani Nenkova, Dinesh Manocha, and Vlad I. Morariu. 2023. Layerdoc: Layer-wise extraction of spatial hierarchical structure in visually-rich documents. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2023, Waikoloa, HI, USA, January 2-7, 2023*, pages 3599–3609. IEEE.

Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. 2021. Roformer: Enhanced transformer with rotary position embedding. *CoRR*, abs/2104.09864.

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. *CoRR*, abs/2208.03054.

Kai Sun, Richong Zhang, Samuel Mensah, Yongyi Mao, and Xudong Liu. 2023. Learning implicit and explicit multi-task interactions for information extraction. *ACM Trans. Inf. Syst.*, 41(2):27:1–27:29.

Jiapeng Wang, Lianwen Jin, and Kai Ding. 2022. Lilt: A simple yet effective language-independent layout transformer for structured document understanding. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 7747–7757. Association for Computational Linguistics.

Mengxi Wei, Yifan He, and Qiong Zhang. 2020. Robust layout-aware IE for visually rich documents with pre-trained language models. In Jimmy X. Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu, editors, *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 2367–2376. ACM.

Zhen Wu, Chengcan Ying, Fei Zhao, Zhifang Fan, Xinyu Dai, and Rui Xia. 2020. Grid tagging scheme for aspect-oriented fine-grained opinion extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2576–2585, Online. 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 *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.

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. 2021. 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 (Volume 1: Long Papers)*, pages 2579–2591, Online. Association for Computational Linguistics.

Yiheng Xu, Tengchao Lv, Lei Cui, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, and Furu Wei. 2022. 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.

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.

Zhenyu Zhang, Bowen Yu, Haiyang Yu, Tingwen Liu, Cheng Fu, Jingyang Li, Chengguang Tang, Jian Sun, and Yongbin Li. 2022. Layout-aware information extraction for document-grounded dialogue: Dataset, method and demonstration. In João Magalhães, Alberto Del Bimbo, Shin'ichi Satoh, Nicu Sebe, Xavier Alameda-Pineda, Qin Jin, Vincent Oria, and Laura Toni, editors, *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pages 7252–7260. ACM.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 726-735, Harbin, China, August 3 - 5, 2023.