Based on Bert and LayoutXLM for Medical OCR Information Extraction

Lianchi Zheng[1*] , Zhihui Sun[1], Yuxiang He[1]

[1] ZHONGYUAN UNIVERSITY OF TECHNOLOGY, Zhengzhou, China
zhenglianchi@gmail.com,xiaosun_wisdom@qq.com,
1090765227@qq.com

Abstract. At present, the medical records used in hospitals are mainly paper, which mostly rely on manual input. With the development of OCR and NLP, the use of OCR and NLP technology to make its information electronic and structured has become a hot spot in the current industry. In this paper, CHIP2022 evaluation task 4 is explored. In this paper, the pre-training model, Bert model without adding coordinate i[1]nformation and LayoutXLM model with adding coordinate information are used to compare the task, and the results are selected and regularized to get a better result. In order to evaluate the accuracy of the framework, we made a list in Ali Cloud Tianchi, and the results show that the framework has achieved good performance.
Keywords: Bert · LayoutXLM · Regularization

1 Introduction

The medical pressure on hospitals is increasing as the COVID-19 outbreak hits. In the hospital, a variety of medical invoices make people dazzling, but also increase the pressure of the hospital to deal with the form data. At the same time, invoice is also widely used in daily life. Hospitals have to issue a lot of invoices every day. However, the invoice management is tedious and the workload is large. In particular, a lot of numbers on the invoice need to be recorded manually, which costs a lot of labor[1]. At present, all kinds of medical invoices used in hospitals are mainly made of paper, which may be damaged in the process of preservation, resulting in double losses of economy and data. This makes the digitized and structured medical invoice into a hot spot for smart medicine. With the rise of artificial intelligence, more and more invoice recognition systems are starting to appear on the market. For example, Baidu's OCR identification system and Tencent's OCR identification system

* Corresponding author

both use deep learning to detect and identify invoice information. We digitize invoices and other materials to extract invoice information quickly and accurately. In this process, invoice processing time can be greatly reduced while parallel savings of labor can be realized. At the same time, it can also improve the accuracy of invoice information record, and unified management of patients' medication and other aspects, which improves the responsibility of medical staff, effectively avoids the generation and implementation of non-standard medical advice, not only protects the rights and interests of patients, avoids disputes and accidents, but also protects the interests of medical staff[2].

Existing frameworks for automated critical field extraction tasks are costly and error-prone when medical forms or invoices need to be extracted after OCR recognition. Invoices contain common structured data or information fields such as vendor name, vendor address, invoice date, invoice number, invoice total, Goods and Services Tax number (GST number), and list of items. An automated key field extraction framework capable of extracting all of this valuable data can significantly increase an organization's productivity by reducing error-prone and laborious manual work[3].

We first use very mature OCR technology to identify all the form data and get its text data. In order to identify required data later, BERT model is selected, mainly because new masked language model (MLM) is used for pre-training BERT[4] model, which can allow two-way perception of context, which is a very excellent technique.

We also use the LayoutXML[5] model, which is an excellent multilingual and multimodal form recognition model. Finally, through data integration, we participated in CHIP2022 evaluation Mission 4 for evaluation.

2 Related Work

For the OCR domain, including text detection and text recognition, deep learning has been very successful. As a downstream task of OCR, extracting key information from documents is a more critical task, and introducing some existing models in three categories: raster-based, graph-based, and end-to-end.

-- **Raster-based.** This kind of method converts an image into a raster representation vector based on image pixels and inputs it into a deep learning network to learn and extract key information. The relationship between the texts in the document is not only affected by the text sequence, but also related to the

layout distribution of each text in the document.To solve the above problems, Katti, Anoop R. et al. [6] proposed the chargrid method, which maps the document image into a character-level 2D grid representation. For each character grid, one-hot encoding is adopted, and then the vector representation is used as the input of Chargrid-Net. Text box detection and semantic segmentation of key information based on encoder-decoder CNN network structure.Zhao, Xiaohui, et al. [7] pointed out that NLP technology alone is unable to process the layout information among various texts in a document. Therefore, the CUTIE method is designed to map the document image to the raster vector representation that retains the spatial position relation of each text, and then two types of CNN models are designed to extract the key information.

-- **Graph-based.** The graph structure-based method regards the document picture as a graph structure composed of text slices, and uses the neural network model to learn the relationship between text slices to extract the key information content of the document. Liu, Xiaojing, et al. [8] pointed out that the traditional NER method BiLSTM-CRF could not make use of the layout information between text slices in document images. This paper proposes to use graph convolutional neural network to learn the semantic information and layout information of text slices, and construct a fully connected graph structure by treating text slices as points and inter-text relations as edges. The graph vector representation of each text slice is learned by using the graph convolutional neural network, which is then spliced with the Word2Vec vector of each text token in the text slice and input into the BiLSTM-CRF network for key information extraction of documents and images.The whole model is optimized by text slice classification task and IOB sequence classification task. Xu, Yiheng, et al. [9] pointed out that the pre-training model has achieved great success in the field of NLP, but it lacks the utilization of layout and layout information, so it is not suitable for the task of extracting key information of documents and images. Therefore, the LayoutLM model was proposed.In this model, BERT (a very powerful pre-training model in the field of NLP) is used as the backbone network. In order to make use of the layout and layout information, 2D position vector representation is introduced, that is, the vectors of the two-point annotations of each text slice (the horizontal and vertical coordinates in the upper left corner and the horizontal and vertical coordinates in the lower right corner) are obtained through the index tables in the horizontal and vertical directions respectively.Optionally, visual vector representations of slices can be added to provide more information. Since BERT can essentially be viewed as a fully connected graph network, we also classify LayoutLM as a graph structure-based technique.Later, LayoutLm-like pre-training models such as Lambert[10] appeared, which obtained SOTA structure on key information extraction tasks of documents and images, proving the powerful ability of deep

learning models based on large corpora and large models.

-- **End-to-end.** End-to-end refers to the key information content of the document obtained directly from the original image as input. Guo, He, et al. [11] pointed out that the information extraction technology based on detection and recognition process would be affected by, for example, slight position offset. To solve the above problems, the EATEN method is proposed, which directly extracts the key information content of the document from the input of the original image. Zhang, Peng, et al. [12] pointed out that key information extraction in existing methods is carried out as multiple independent tasks, that is, text detection, text recognition and information extraction, which cannot be supervised and learned from each other. Therefore, the author proposed an end-to-end network model TRIE to conduct model learning on the above three tasks at the same time.

Based on the above existing work, this paper adopts the LayoutXLM model based on multi-modal combination of tasks, and uses the Bert model to directly do named entity recognition for the identified content for comparison and combined exploration of the results.

3 Data preprocessing

3.1 CHIP2022-Medical manifest invoice OCR factor extraction task dataset

The data set used this time includes 1000 recognition training sets, among which 200 discharge summary invoices, 200 medicines invoices, 200 outpatient invoices and 400 hospitalization invoices are all real data pictures and labeling results. Identification evaluation A list, namely TestA, contains 200 real data pictures and marked results; Identification evaluation B list, TestB, contains 500 real data pictures and marked results. The tags that need to be extracted for each class vary.

3.2 Image OCR recognition and error correction

We adopted PaddleOCR[13] to recognize each image, and the recognition result was the text content and coordinate information of each sentence. Then, we used the single word similarity to correct errors in the identified content to prevent excessive identification errors from affecting the labeling effect. For each label in each photo, the corresponding relationship was found in the sentence. The length of the label was taken as the window, and the step size was 1 for matching. The average value is the similarity between the contents of the window and the label. We set a threshold. If the similarity is bigger than 0.66, it is considered to be an

error, and the contents of the window are changed to the label value.

3.3 Data labeling

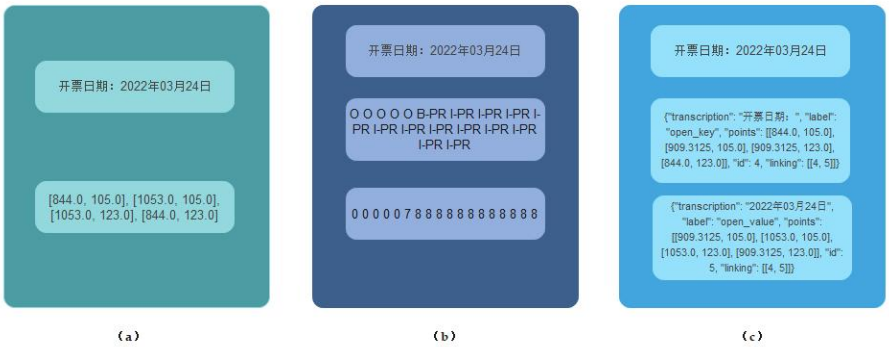

Fig. 1.(a)Sample one of the original data. (b)Sample of BERT's NER annotation. (c)Sample annotation of SER for LayoutXLM.

-- Bert uses named entity recognition (NER) to identify the key information, so it needs to use BIO annotation. The code is used for annotation. The original data is shown in **Fig. 1(a)**, and the annotated data is shown in **Fig. 1(b)**. Specific implementation of BIO annotation:

1. Read the annotations given in the data set.
2. Read the recognition result of each image and divide it into single words.
3. Match each sample word and get the result.
4. Structured storage.

-- Since the LayoutXLM training data is in the unit of sentences, the identified results are divided into sentences and then compared with the sentence label sentence by sentence. Sometimes coordinates need to be merged and split. The marked data is shown in **Fig. 1(c)**. It's a label for each entity but it's actually still a BIO tag on the inside but it's just a different representation on the outside. Specific implementation of annotation:

1. Read the annotations given in the data set.
2. Read the recognition result of each image and divide it into single

words.

3. Match each sample word and get the result.
4. Structured storage.

4 Proposed Framework

The framework proposed in this paper is based on BERT model and uses mask language model (MLM) and layoutXML model to implement BioNER tasks. By combining the two frameworks, we can produce different results when we encounter inputs with poor eigenvalues, and produce more reasonable and accurate recognition results after matching. It is important to note that we did not introduce additional annotated data in this mission prediction, but obtained results directly from NER annotations.

4.1 Main Model

BERT model is a simple concept, but the practical results are very powerful model. The main body of BERT model is formed by stacking the simplest Transformer. The model structure is shown in **Fig.2**. Moreover, in pre-training, we have to mention the Masked Language Model (MLM), because the models can only be trained from left to right or right to left under standard conditions. In order to predict target words in the paper, MLM technology is introduced. He replaces some tokens by inserting a random percentage of [mask], and then predicts it. To avoid the negative effects of [mask], he inserts [mask] as follows:

1. It has an 80% probability of being replaced with a normal [mask], for example: My heart is broken———>My heart is [mask].
2. It has a 10% chance of being replaced with a random word, for example: My heart is broken———>My heart is lonely.
3. There is a 10% chance that the original word will remain the same, for example: My heart is broken———>My heart is broken. We also add task-specific class tokens ([CLS]) to the input of the BERT model.

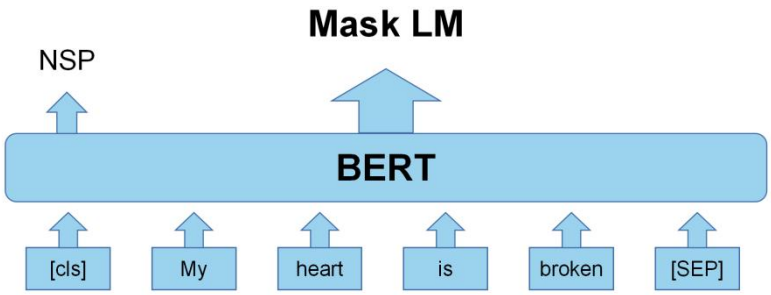

Fig.2.Main structure of BERT model

The LayoutXLM model is designed with a multimodal transformer architecture, similar to the LayoutLMv2[14] framework. The model accepts information from three different modes, including text, layout and image, which are coded as the text embedding layer, layout embedding layer and visual embedding layer respectively. After text and image embeds are concatenated, layout embeds are added to give model input. We use this model to realize the recognition of medical forms, and for key value extraction, one of the most critical tasks in form understanding, similar to FUNSD, this task is defined as two sub-tasks, namely semantic entity recognition and relationship extraction, task description. We're using semantic entity recognition.

In the Semantic Entity Recognition (SER) subtask, the descriptive methods are as follows:

Given A rich text document A, get the token sequence t={t_0, t_1... , t_n}, where each token can be expressed as t_i = (w, (x_0, y_0, x_1, y_1)), w is the token text, and (x_0, y_0, x_1, y_1) is the spatial coordinate position of the text in the document. Define the class of all semantic entities as C = {c_0, c_1,... c_m}. The semantic entity recognition task requires the model tag to extract all defined semantic entities and classify them into correct categories, that is, to find the function Fser:(A, C) ->, where E is the semantic entity set predicted by the model:

$$\varepsilon = \left\{ \left(\left\{ t_0^0, \cdots, t_0^{n0} \right\}, c_0 \right), \cdots, \left(\left\{ t_k^0, \cdots, t_k^{nk} \right\}, c_k \right) \right\}$$

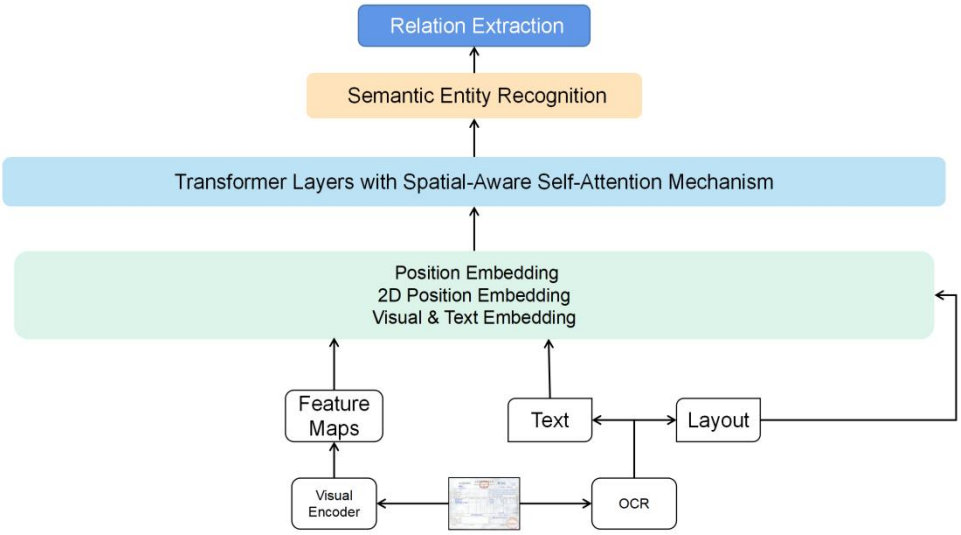

Fig.3.LayoutXML model structure diagram

5 Experiment

5.1 Datasets

The experiment was conducted on the OCR factor extraction task data set of the CHIP2022 Evaluation Task 4 medical invoice. Firstly, we annotate the given data to transform it into trainable structural data, in which we explore the similarity threshold of tag and text matching. Then, we use Bert and LayoutXLM models to conduct experiments respectively, and then combine the results of the two models to evaluate the model performance.

5.2 Setting

-- Bert adopts AdamW as the optimizer, lr is 2e-5, linear attenuation of learning rate is adopted, and the maximum number of training rounds is 96. Use transformers 4.7.0 and Pytorch 1.10.1 to structure the training model and push it to the huggingface community for invocation.

-- LayoutXLM employs AdamW as an optimizer, lr 2e-5, linear attenuation of

learning rate, maximum number of training rounds of 200, and training with paddle 2.3.2.

5.3 Experimental Results

In order to demonstrate the advantages of the two models in parallel, we have conducted experiments using only Bert and LayoutXLM models respectively. The score evaluation results of TestB in CHIP2022 Evaluation 4 are shown in **Table 1**:

Table 1.The score evaluation results of TestB in CHIP2022 Evaluation 4

Model	Evaluation Metrics	Score
Bert	CYXJ-Acc	0.8094
	GYFP-Acc	0.5705
	MZFP-Acc	0.1276
	ZYFP-Acc	0.1433
	Acc	0.2486
LayoutXLM	CYXJ-Acc	0.5162
	GYFP-Acc	0.6378
	MZFP-Acc	0.3437
	ZYFP-Acc	0.4337
	Acc	0.4435
Bert+LayoutXLM	CYXJ-Acc	0.8094
	GYFP-Acc	0.7384
	MZFP-Acc	0.3495
	ZYFP-Acc	0.4237
	Acc	0.4705

- As can be seen from Table 1, the effect of NER without coordinates is relatively good for the task of discharge summary, which has little recognition content but not obvious structure of the contents on the picture. However, the effect of NER without coordinates is very poor for the outpatient invoice and inpatient invoice with more recognition content but more structured content on the picture.
- LayoutXLM model is used only for the identification of medicines invoice, outpatient invoice and hospitalization invoice, whose data is formatted on the picture is better, but for the categories of discharge summary invoice with lower structured data but less identification content, the effect will be worse.

- Based on the above comparison, we combine the results of the two models through regularization query, and it can be concluded that each index of the two models has improved, which is 22.19% higher than that of Bert and 2.7% higher than that of LayoutXLM. Therefore, this method is effective.

6 Conclusion

Based on CHIP2022 evaluation task 4, this paper explores the relationship between Bert and LayoutXLM for OCR information extraction tasks, and the results are modified by regular and spliced to get a better result. In order to label data in BIO format, we extrapolated word-to-word similarity to sentence-to-sentence similarity and labeled it.The accuracy of discharge summary invoice is higher in Bert, because there are fewer entities to be identified, and decreases in LayoutXLM, because its coordinate information will lead to errors. The coordinate information of the other three types is relatively regular in the picture, so the accuracy is higher in LayoutXLM. Finally, the integrated results are regularized and some results are modified to make them reasonable. Finally, in the evaluation TestB, CYXJ-Acc, GYFP-Acc, MZFP-Acc, ZYFP-Acc and Acc are 0.8094 respectively. 0.7384, 0.3495, 0.4237, 0.4705, ranking third in open source code.

References

1. Shi S , Cui C , Xiao Y . An Invoice Recognition System Using Deep Learning[C]// 2020 International Conference on Intelligent Computing, Automation and Systems (ICICAS). 2020.
2. Yao X, Sun H, Li S, et al. Invoice Detection and Recognition System Based on Deep Learning[J]. Security and Communication Networks, 2022, 2022.
3. Baviskar D , Ahirrao S , Kotecha K . Multi-Layout Unstructured Invoice Documents Dataset: A Dataset for Template-Free Invoice Processing and Its Evaluation Using AI Approaches[J]. IEEE Access, 2021, PP(99):1-1.
4. Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.
5. Xu Y , Lv T , Cui L , et al. LayoutXLM: Multimodal Pre-training for Multilingual Visually-rich Document Understanding[J]. 2021.
6. Katti A R , Reisswig C , Guder C , et al. Chargrid: Towards Understanding 2D Documents[J]. 2018.
7. Zhao X , Niu E , Wu Z , et al. CUTIE: Learning to Understand Documents with Convolutional Universal Text Information Extractor[J]. 2019.
8. Liu X , Gao F , Zhang Q , et al. Graph Convolution for Multimodal Information Extraction from Visually Rich Documents:, 10.18653/v1/N19-2005[P]. 2019.

9. Xu Y, Li M, Cui L, et al. Layoutlm: Pre-training of text and layout for document image understanding[C]//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020: 1192-1200.
10. Garncarek U , Powalski R , Stanisawek T , et al. LAMBERT: Layout-Aware language Modeling using BERT for information extraction:, 10.48550/arXiv.2002.08087[P]. 2020.
11. Guo H , Qin X , Liu J , et al. EATEN: Entity-aware Attention for Single Shot Visual Text Extraction[J]. 2019.
12. Zhang P , Xu Y , Cheng Z , et al. TRIE: End-to-End Text Reading and Information Extraction for Document Understanding[J]. 2020.
13. Li C, Liu W, Guo R, et al. PP-OCRv3: More Attempts for the Improvement of Ultra Lightweight OCR System[J]. arXiv preprint arXiv:2206.03001, 2022.
14. Xu Y , Xu Y , Lv T , et al. LayoutLMv2: Multi-modal Pre-training for Visually-Rich Document Understanding[J]. 2020.
15. Wolf T , Debut L , Sanh V , et al. Transformers: State-of-the-Art Natural Language Processing[C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2020.