Improving Medical OCR Information Extraction with Integrated Bert and LayoutXLM Models

Lianchi Zheng[1,2], Xiaoming Liu[1,2,3⋆],Zhihui Sun[1,3], Yuxiang He[1,3]

[1] Zhongyuan University of Technology, Zhengzhou, 450007,China
zhenglianchi@gmail.com,ming616@zut.edu.cn
xiaosun_wisdom@qq.com,1090765227@qq.com
[2] Henan Key Laboratory on Public Opinion Intelligent Analysis, Zhengzhou, 450007,China
[3] Zhengzhou Key Laboratory of Text Processing and Image Understanding, Zhengzhou, 450007,China

Abstract. Currently, medical records in most hospitals are paper-based and rely on manual input, but with the advancements in OCR and NLP technologies, it is now possible to convert such records into electronic and structured formats. In this paper, we explore the CHIP2022 evaluation task 4 and compare the performance of two pre-training models: Bert without additional coordinate information and LayoutXLM with additional coordinate information. We apply a selection and regularization process to refine the results and evaluate our framework's accuracy through a list in Ali Cloud Tianchi. Our results demonstrate that our framework achieved good performance .
Keywords: Bert · LayoutXLM · Regularization

1 Introduction

As the COVID-19 outbreak continues worldwide, hospitals are facing incr[1]easing medical pressure. Among the challenges that hospitals face daily, dealing with numerous types of medical invoices not only creates complexity but also adds pressure to hospital staff. In everyday life, individuals encounter invoices regularly, and hospitals issue a considerable number of invoices in the course of their operations. Tedious invoice management processes and tremendous manual labor[1] costs have become major concerns. Moreover, the vast majority of medical invoices used in hospitals is paper-based, which is prone to damage and leads to data and economic losses. With artificial intelligence advancing at an unprecedented pace, an increasing

⋆ Corresponding author

number of invoice recognition systems based on deep learning technology, such as Baidu's OCR identification system and Tencent's OCR identification system, are emerging in the market. This paper aims to digitize medical invoices and use OCR technology to extract invoice information accurately and efficiently. Such innovation can reduce invoice processing time, save labor costs in parallel, increase the accuracy of information records, and enable unified management in medication and other aspects, overhauling medical staff's responsibility, and preventing the occurrence of non-standard medical advice. Ultimately, it protects patients' rights and interests, avoids disputes and accidents, and safeguards the interests of medical staff[2].

Existing frameworks for automated critical field extraction tasks can be expensive and prone to errors, particularly when extracting data from medical and invoice forms following OCR recognition. Medical invoices comprise structured data or information fields such as vendor name, vendor address, invoice date, invoice number, invoice total, Goods and Services Tax (GST) number, and list of items. Developing an automated key field extraction framework that can extract this relevant data can significantly raise an organization's productivity by minimizing error-prone and manual work[3].

The MedOCR task in CHIP2022 presents complex scenarios, such as outpatient and inpatient invoices, which feature numerous extracted content, irregular layout formats, and low image quality. As a result, they are prone to text recognition errors. Additionally, labeling OCR data is also challenging. To address these difficulties, we propose a two-stage approach. Firstly, we use an advanced OCR model to extract the picture's text and then perform text sequence annotation as the later named entity recognition's annotation data to extract information. We also correct some text errors based on the label's text. Finally, we use the Bert[4] model for named entity recognition to obtain the desired information. Additionally, we directly select the advanced LayoutXML[5] multimodal model that includes coordinate information for direct end-to-end recognition. Ultimately, we combine the two recognition results to obtain the final recognition results effectively. To validate our model's effectiveness, we participated in Alibaba Cloud Tianchi and achieved an accuracy rate for each item, as shown in Table 1.

2 Related Work

Deep learning has achieved great success in text detection and recognition for the OCR domain. However, extracting key information from documents is a more

critical task, and there are currently three types of existing models: raster-based, graph-based, and end-to-end.

-- **Raster-based.** Raster-based methods convert images into a raster representation vector based on image pixels, which is then inputted into deep learning networks to learn and extract key information. However, the relationship between the texts in the document is not only affected by the text sequence but also by the layout distribution of each text. To address this issue, Katti et al. proposed the Chargrid method[6], which maps document images into a character-level 2D grid representation. One-hot encoding is adopted for each character grid, and the vector representation is used as input for Chargrid-Net. The method uses an encoder-decoder CNN network structure for text box detection and semantic segmentation of key information. Similarly, Zhao et al. developed the CUTIE method[7] to address the inability of NLP technology alone to process layout information among various texts in a document. CUTIE maps the document image to a raster vector representation that retains the spatial position relation of each text and uses two types of CNN models to extract key information.

-- **Graph-based.** The graph structure-based method regards the document picture as a graph structure composed of text slices, and uses a 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 a 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 entire model is optimized by text slice classification task and IOB sequence classification task.

Xu, Yiheng, et al.[9] pointed out that pre-training models have achieved great success in the field of NLP but lacked the utilization of layout and layout information, making them unsuited for the task of extracting key information from documents and images. Therefore, the LayoutLM model was proposed, which uses BERT (a powerful pre-training model in the NLP field) as the backbone network. To make use of the layout and layout information, a 2D position vector representation is introduced, obtaining vectors of the two-point annotations of each text slice (horizontal and vertical coordinates in the upper-left corner and the lower-right corner) through 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 fundamentally 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 direct extraction of the key information content of a document from the original input image. Guo, He et al. [11] highlighted that information extraction technologies based on detection and recognition processes may be affected by slight positional offsets. To overcome these problems, the EATEN method was proposed, which directly extracts the key information content of the document from its original input image. Zhang, Peng et al. [12] noted that existing key information extraction methods rely on multiple independent tasks, such as text detection, text recognition, and information extraction, which are not supervised and learned from each other. As a solution, the authors proposed an end-to-end network model called TRIE that simultaneously learns from the aforementioned three tasks.

Based on the existing work, this paper utilizes the LayoutXLM model with multi-modal task combination and directly applies the Bert model for named entity recognition. By comparing and exploring the combined results, it aims to achieve an improved outcome.

3 Proposed Framework

The framework proposed in this study is based on the BERT model, which employs a mask language model (MLM) and a layoutXML model to implement BioNER tasks. By integrating the two frameworks, we aim to improve the accuracy of recognition, particularly for inputs with poor eigenvalues, and to obtain more reasonable and accurate recognition results after matching. It is worth noting that we did not introduce additional annotated data in this mission prediction; instead, we directly obtained results from NER annotations.

3.1 Main Model

The BERT model is a highly potent yet simple concept model, constructed by stacking the simplest Transformer structures to form the core architecture of the

model as depicted in **Fig.1**. During pre-training, it is necessary to employ the Masked Language Model (MLM) because under standard conditions, the model can only be trained from either left-to-right or right-to-left directions. The MLM technology is introduced to predict target words in complex sentences. To address the limitation of the model, some tokens are replaced with a random percentage of [mask], which is then predicted. To mitigate any adverse effects of [mask], the following operation is performed during pre-processing:

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.

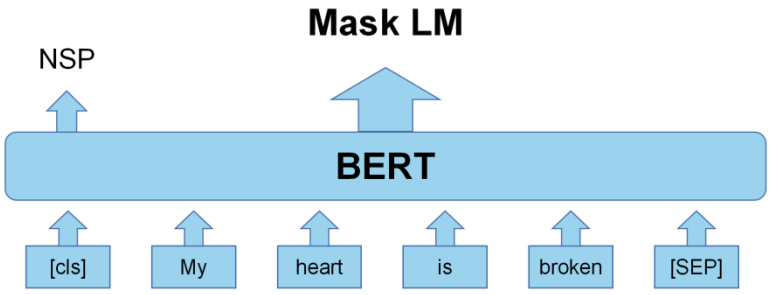

Fig.1.Main structure of BERT model

The LayoutXLM model is designed with a multimodal transformer architecture, similar to the LayoutLMv2[13] 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.The core framework of the model in **Fig.2**.

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:

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

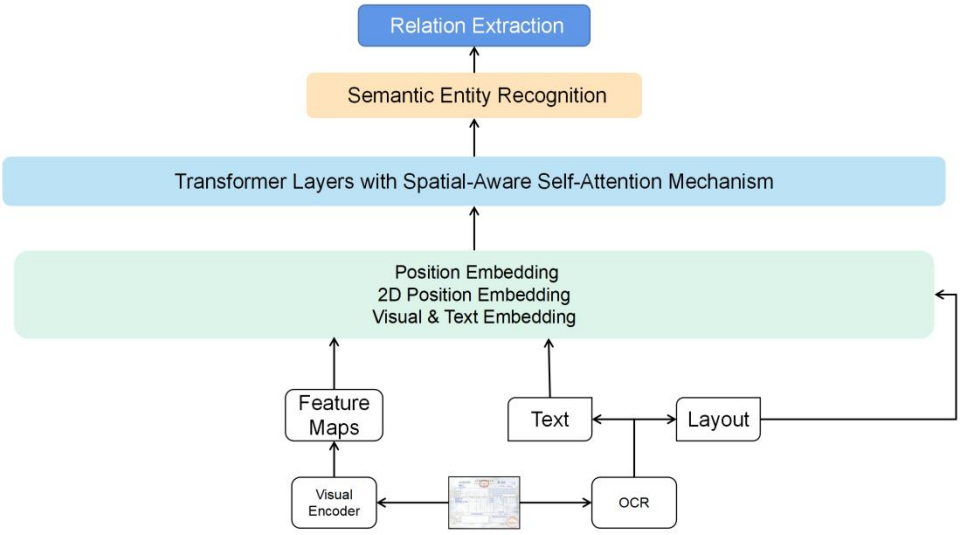

Fig.2.LayoutXML model structure diagram

4 Experiment

4.1 Datasets

The experiment was conducted using the OCR identification of electronic medical paper documents (ePaper) task dataset of the CHIP2022 Evaluation Task 4 medical invoice. The dataset employed in this study encompasses 1000 samples for recognition training. This dataset comprises 200 discharge summary invoices, 200 medicines invoices, 200 outpatient invoices, and 400 hospitalization invoices, all of which are real-world images, together with labeled results. The identification evaluation comprises two distinct lists – TestA and TestB. TestA lists 200 real-world images with corresponding marked results for identification assessment. TestB, on the other hand, lists 500 real-world images with matching marked results. It is worth noting that each class has different tags that need to be extracted.

To transform the given data into trainable structural data, we annotated the dataset while exploring the similarity threshold of tag and text matching. Subsequently, we performed experiments individually using the Bert and LayoutXLM models. Finally, we combined the results of both models to evaluate the overall model performance.

4.2 Image OCR recognition and error correction

In this study, we employed PaddleOCR[14] to recognize each image. The recognition result comprised sentence-based text content together with their corresponding coordinate information. We then corrected any identification errors using single word similarity in order to avoid the labeling effect being impacted by excessive identification errors. We found the corresponding relationship between each label and the sentence in each photo. Taking the length of the label as the window and a step size of one for matching, we computed the average value as the similarity between the contents of the window and the label. We subsequently set a threshold for similarity at 0.66. Any similarity scores exceeding this value were considered as errors. In such cases, the contents of the window were changed to match the label values.

4.3 Data labeling

Fig.3.(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.3(a)**, and the annotated data is shown in **Fig.3(b)**. Specific implementation of BIO annotation:

1. Read the annotations given in the dataset.
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.3(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 text and coordinates after image recognition.
2. Error correction of identification results.
3. Compare according to the labels to obtain the labels of each coordinate frame.
4. Structured storage.

4.4 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[15] 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.

4.5 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.

5 Conclusion

This paper explores the relationship between Bert and LayoutXLM for OCR information extraction tasks based on the CHIP2022 evaluation task 4. The results are refined through regularization and splicing to obtain better results. To label data in BIO format, we extrapolated word-to-word similarity to sentence-to-sentence similarity. Our findings indicate that Bert achieves higher accuracy in extracting information from discharge summary invoice, presumably because there are fewer entities to be identified, whereas LayoutXLM accuracy decreases due to coordinate information errors. For the other three types, the accuracy is higher in LayoutXLM since their coordinate information is relatively regular in the picture. Lastly, we regularized the integrated results and modified them to make them more reasonable. In the evaluation TestB, CYXJ-Acc, GYFP-Acc, MZFP-Acc, ZYFP-Acc, and Acc were 0.8094, 0.7384, 0.3495, 0.4237, and 0.4705, respectively, ranking our paper third in open source code.

6 Acknowledgment

We gratefully thank the anonymous reviewers for their helpful comments and suggestions. This study was supported partly by the National Natural Science Foundation of China (NSFC No.62076167), Ministry of Education industry-school cooperative education project(Grant No. 201902298016)，and Key Research Project of Henan Higher Education Institutions (Granted No. 23A520022).

References

1. S. Shi, C. Cui, and Y. Xiao, An invoice recognition system using deep learning [J].

in 2020 International Conference on Intelligent Computing, Automation and Systems (ICICAS). IEEE, 2020, pp. 416 – 423.
2. X. Yao, H. Sun, S. Li, and W. Lu, Invoice detection and recognition system based on deep learning [J]. Security and Communication Networks, vol. 2022, pp. 1 – 10, 2022.
3. D. Baviskar, S. Ahirrao, and K. Kotecha, Multi-layout unstructured invoice documents dataset: A dataset for template-free invoice processing and its evaluation using ai approaches [J]. IEEE Access, vol. 9, pp. 101494 – 101512, 2021.
4. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding [J]. arXiv preprint arXiv:1810.04805, 2018.
5. Y. Xu, T. Lv, L. Cui, G. Wang, Y. Lu, D. Florencio, C. Zhang, and F. Wei, Layoutxlm: Multimodal pre-training for multilingual visually-rich document understanding [J]. arXiv preprint arXiv:2104.08836, 2021.
6. A. R. Katti, C. Reisswig, C. Guder, S. Brarda, S. Bickel, J. Höhne, and J. B. Faddoul, Chargrid: Towards understanding 2d documents [J]. arXiv preprint arXiv:1809.08799, 2018.
7. X. Zhao, E. Niu, Z. Wu, and X. Wang, Cutie: Learning to understand documents with convolutional universal text information extractor [J]. arXiv preprint arXiv:1903.12363, 2019.
8. X. Liu, F. Gao, Q. Zhang, and H. Zhao, Graph convolution for multimodal information extraction from visually rich documents [P]. arXiv preprint arXiv:1903.11279, 2019.
9. Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, and M. Zhou, Layoutlm: Pre-training of text and layout for document image understanding [C]. in Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 1192 – 1200.
10. Garncarek, R. Powalski, T. Stanisawek, B. Topolski, P. Halama, M. Turski, and F. Gralinski, Lambert: Layout-aware language modeling for information extraction [P]. in Document Analysis and Recognition – ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5 – 10, 2021, Proceedings, Part I. Springer, 2021, pp. 532 – 547.
11. H. Guo, X. Qin, J. Liu, J. Han, J. Liu, and E. Ding, Eaten: Entity-aware attention for single shot visual text extraction [J]. in 2019 International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2019, pp. 254 – 259.
12. P. Zhang, Y. Xu, Z. Cheng, S. Pu, J. Lu, L. Qiao, Y. Niu, and F. Wu, Trie: end-to-end text reading and information extraction for document understanding [J]. in Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 1413 – 1422.
13. Y. Xu, Y. Xu, T. Lv, L. Cui, F. Wei, G. Wang, Y. Lu, D. Florencio, C. Zhang, W. Che et al., Layoutlmv2: Multi-modal pre-training for visually-rich document understanding [J]. arXiv preprint arXiv:2012.14740, 2020.
14. C. Li, W. Liu, R. Guo, X. Yin, K. Jiang, Y. Du, Y. Du, L. Zhu, B. Lai, X. Hu et al., Pp-ocrv3: More attempts for the improvement of ultra lightweight ocr system [J].

arXiv preprint arXiv:2206.03001, 2022.
15. T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz et al., Transformers: State-of-the-art natural language processing [C]. in Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations, 2020, pp. 38–45.
16. Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, Uniter: Universal image-text representation learning [J]. in Computer Vision–ECCV 2020, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 104–120.
17. B. Shi, X. Bai, and C. Yao, An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition [J]. in Arxiv 2015.
18. S. Mori, C. Suen, and K. Yamamoto, Historical review of ocr research and development [J]. Proceedings of the IEEE, vol. 80, no. 7, pp. 1029–1058, 1992.
19. D. Ming, J. Liu, and J. Tian, Research on chinese financial invoice recognition technology [J]. in Arxiv Pattern Recognition Letters, vol. 24, no. 1, pp. 489–497, 2003.
20. Zong H, Lei J, Li Z, et al. Overview of technology evaluation dataset for medical multimodal information extraction [J]. Journal of Medical Informatics, 2022, 43(12):2-5+22.
21. Liu L, Chang D, Zhao X, et al. MedOCR: the dataset for extraction of optical character recognition elements for medical materials [J]. Journal of Medical Informatics, 2022, 43(12):28-31.
22. Liu L, Chang D, Zhao X, et al. Information extraction of medical materials: an overview of the track of medical materials MedOCR [C]//Health Information Processing: 8th China Conference, CHIP 2022, Hangzhou, China, October 21-23, 2022, Revised Selected Papers. Singapore: Springer Nature Singapore.