IntelliExtract: An End-to-End Framework for Chinese Resume Information Extraction from Document Images

Yijing Liu

School of Computing and Data Science, Xiamen University Malaysia, Selangor, 43900, Malaysia

CST2009133@xmu.edu.my

Abstract. Traditional document processing can be labor-intensive and time-consuming to manually extract and organize the information in a document. This manual process is often inefficient and error-prone. In order to improve processing efficiency and accuracy of document data, we develop IntelliExtract, an end-to-end framework designed for document information extraction. This is a comprehensive framework that includes image text detection and recognition, information extraction, and document intelligent question-answering. Some recent models and algorithms are employed, OCR models for converting scanned documents into machine readable text, layout analysis algorithms for understanding the spatial arrangement of document elements, and information extraction techniques for extracting structured data from unstructured documents. To evaluate the effectiveness of the framework, we conducted experiments by employing a Chinese Talent Resumes Dataset for visualizing the results. For named entity extraction, the confidence level of the extracted results from the text in the images is generally above 0.95. The proposed framework provides a powerful tool for enterprises, educational institutions, and other entities in processing document information, and holds promise for significant practical applications.

Keywords: End-to-End, document information extraction, document intelligent question-answering, named entity extraction.

1. Introduction

With the rapid growth of the Internet and digital technology, the volume of resumes and candidate information is growing exponentially. Recruiters are faced with the challenge of processing large amounts of document data and extracting key information, which can be a tedious and time-consuming task for manual operations. This information overload is putting significant pressure on recruiting departments, highlighting the need for more efficient ways to process and manage this massive amount of information. In addition, manual extraction and organization of document information is prone to errors due to the inevitable errors and fatigue of human operations. This can lead to the omission or misclassification of important information, which can affect the accuracy and efficiency of hiring decisions. Therefore, the recruiting industry needs a solution that can automate and improve the efficiency and accuracy of document data processing like information extraction.

Information Extraction (IE) technology extracts specific objects or events details from natural language text, and then to further automate the categorization or reconstruction of large amounts of content. It consists of three main subtasks: Named Entity Recognition (NER), Relationship Extraction (RE), and Event Extraction (EE). With the continuous development and popularity of Internet technology, information extraction technology has become an effective screening management method for the recruitment industry. Recruiters receive a lot of resume information through mailboxes, job boards and other channels every day. It is really time-consuming to deal with plenty of information of candidates and store key data. Therefore, information extraction of resumes from by using a powerful resume parsing system is become a more effective way in all screening management.

NER has seen significant advancements in information extraction, from traditional models like Hidden Markov Models and Conditional Random Fields (CRF) to recent models such as Bidirectional Long Short-Term Memory-CRF (BiLSTM-CRF) and Lattice LSTM. Additionally, TENER, a novel NER architecture, incorporates an adapted Transformer Encoder to effectively model both character-level features and word-level features [1]. The introduction of BERT, a pre-trained model Pre-training of Deep Bidirectional Transformers for Language Understanding [2], further enhanced NER performance. In the realm of RE, models like Convolutional Neural Networks (CNN) and Piecewise Convolutional Neural Networks (PCNN) gained popularity in the deep learning era, and a dynamic multi-pooling convolutional neural network, known as DMCNN, was proposed by researcher Yubo Chen [3] at the Chinese Academy of Sciences and became a notable example in EE.

Drawing upon the advancements mentioned above in NER, our work focuses on developing a comprehensive framework that seamlessly integrates multiple techniques to improve the accuracy and completeness of entity extraction from image-based resumes. We present an end-to-end framework called IntelliExtract, designed to address the task of identifying and extracting named entities from unstructured text data in resumes. A Chinese resume dataset is utilized to verify the efficient of our framework. Firstly, we implemented optical character recognition (OCR) using the differential binarization (DB) technique and the Scene Text Recognition with a Single Visual Model (SVTR) model [4] to successfully detect and recognize Chinese text. Secondly, we combined the Structure Location Alignment Network (SLANet) model [5] and the PicoDet model [6] to perform table recognition and layout analysis on complexly formatted resume data. The results of the final entity extraction exhibit significant completeness and accuracy, surpassing a threshold of 0.95. In addition, we incorporated the Visual-feature Independent LayoutXLM (VI-LayoutXLM) pre-training model [5], into our methodology to conduct Semantic Entity Recognition (SER) on the resume dataset. Then we reintegrated the recognized results and set the schema for entity recognition, and the efficiency and completeness of the extracted information were improved. Finally, we use the ERNIE-Layout [7] model for document intelligent answering to achieve information filtering. Compared with traditional methods, our framework can extract key information more accurately, reducing errors and omissions.

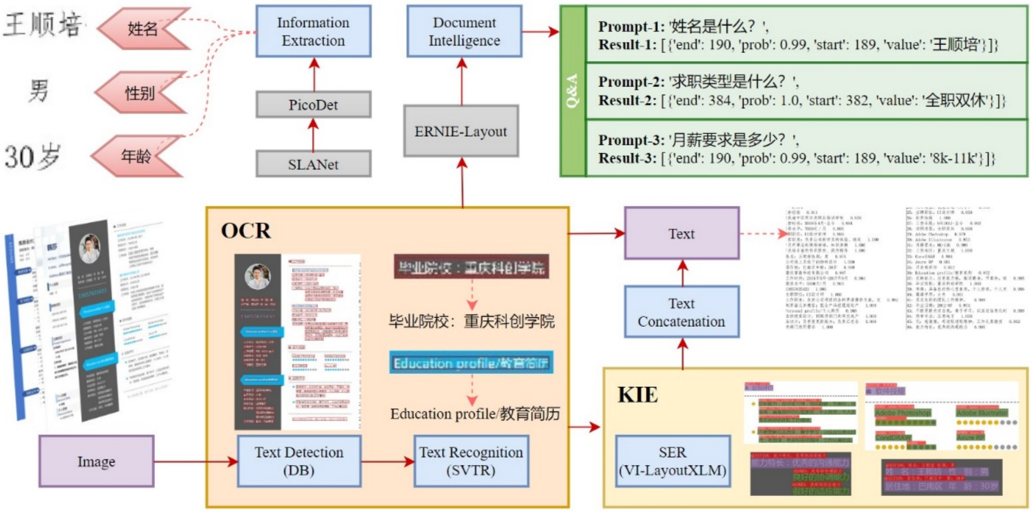

王顺培
姓名
男
性别
30岁
年龄
Information Extraction
PicoDet
SLANet
Document Intelligence
ERNIE-Layout
Q&A
Prompt-1: '姓名是什么？ ',
Result-1: [{'end': 190, 'prob': 0.99, 'start': 189, 'value': '王顺培'}]}
Prompt-2: '求职类型是什么？ ',
Result-2: [{'end': 384, 'prob': 1.0, 'start': 382, 'value': '全职双休'}]}
Prompt-3: '月薪要求是多少？ ',
Result-3: [{'end': 190, 'prob': 0.99, 'start': 189, 'value': '8k-11k'}]}
OCR
毕业院校：重庆科创学院
毕业院校：重庆科创学院
Education profile/教育简历
Education profile/教育简历
Image
Text Detection (DB)
Text Recognition (SVTR)
Text
Text Concatenation
KIE
SER (VI-LayoutXLM)

Figure 1: A Unified Framework for Information Extraction

Major contributions are summarized as follows:

We propose an end-to-end, application-oriented framework that integrates state-of-the-art techniques to

achieve the desired functionalities in each module.

We use deep learning models based on table recognition and layout analysis to accurately extract and parse information from tables, improving the accuracy and reliability of entity extraction.

We utilize a vision-independent LayoutXLM pre-training model to improve the effectiveness of traditional OCR in extracting key information from resume images.

The remaining sections of this paper are structured as follows: Section 2 provides a literature review about different approaches of information extraction; Section 3 outlines our utilized models and demonstrate our framework for extracting relevant information from resumes in image format; Section 4 discusses the implementation and resulting outputs, and finally, Section 5 concludes the work with suggestions for future research.

2. Related Work

With the integration and development of natural language processing (NLP) and deep learning techniques, the extraction of relevant information from unstructured documents such as resumes has become an important research topic in recent years. Various methods have been proposed to extract structured information from unstructured resumes, including rule-based methods, machine learning-based methods, deep learning-based methods, and semantic-based methods.

Rule-based resume information is extracted by parsing the resume file, converting it into a structured data format, and then using a series of rules to extract information, including personal information, educational background, work experience, etc. Several rule-based extraction methods treat resumes as webpages and extract detailed information using the structural hierarchy of the Document Object Model [8-10]. Another feasible method is to parse the resume file into XML format [11] and then use rule-based regular expressions and keyword matching techniques to extract metadata. Rule-based approach usually has higher accuracy due to the ability to design custom rules based on different resume formats. It is applicable to small data sets and domain-specific resume information extraction. However, the design and maintenance of rules require a lot of manual labor and domain knowledge, and cannot handle diverse resume formats.

Machine learning based approaches have been proposed to adapt resume information extraction to different formats and domains. The structure and features of resumes can be learned from data automatically without writing rules manually. For example, Ayishathahira et al. [12] proposed a hybrid system for resume parsing using deep learning models such as the CNN, Bi-LSTM and CRF. An end-to-end resume extraction system is created by Tobing et al. [13], who used some heuristic rules and machine learning algorithms to solve the problem of extracting

contents from an Indonesian resume. Nahm et al. [14] described an information extraction system based on transformation-based learning, which uses learned meta-rules on patterns for slots, to improve the performance of the underlying information extraction. Téllez-Valero et al. [15] proposed a combination of regular expressions and text classifiers to adapt an IE application to a new domain.

In recent years, deep learning has made remarkable progress in NLP tasks. Compared to shallow machine learning methods,

deep neural networks have better adaptability to unstructured data and can automatically discover features through their powerful representation learning capabilities. For example, Gaur et al. [16] proposed a semi-supervised approach, which involves training a neural network model on a small initial set of labeled data and using it to predict entities in unlabeled data.

And a high overall accuracy was achieved without the need of large annotated data. By utilizing the sequential relationships existing amongst different blocks of resumes, Xu et al. [17] found an effective hybrid model which combines a fully connected neural network model and a block-level recurrent neural network model, to deal with a special scenario in text classification which has a weak sequential relationship among different classification entities. Zu et al. [18] proposed an end-to-end pipeline for resume parsing based on neural networks-based classifiers and distributed embeddings. A resume was effectively segmented into predefined text blocks and BLSTM-CNNs-CRF's superiority in named entity recognition task was confirmed.

In large-scale text processing, semantic-based approaches are more applicable. By using domain knowledge and rules to model linguistic structures, it is possible to better deal with ambiguity and polysemy in text. Celik et al. [19] presented an information extraction system based on ontology, designed to handle millions of resumes in free-text format. The system aims to transform the resumes into a structured and semantically enriched format, which can be used for semantic data mining and extracting crucial information for human resources processes. Hao et al. [20] suggested a Semantic-based Resume Screening System for effectively screening resumes with the assumption that electronic resumes of all candidates will be saved in XML format.

3. Methods

Our framework combines OCR, pre-trained models, layout analysis, and state-of-art models like SLANet, PicoDet, VI-LayoutXLM to enhance information extraction and ERNIE-Layout to enable intelligent Q&A of resume documents.

3.1 Overall Framework

The framework improves the accuracy and completeness of resume entity extraction by integrating OCR techniques, various pre-trained models, and techniques such as layout analysis and question-answer matching. It uses OCR techniques to convert image resumes into machine-readable digital text format. For text extraction, we use NER techniques to extract specific information from documents, such as names, addresses, etc. We used models such as SLANet and PicoDet for layout analysis and target detection to further improve the accuracy and completeness of entity extraction. For text filtering, we utilize the document intelligence capability of the ERNIE-Layout model for resume screening. The model combines text, layout and image features to extract information related to the question by matching the question with the answer.

3.2 Text Detection and Recognition

OCR is to recognize the text in the image and return the content in the form of text. The overall framework of OCR utilizes a pipeline approach, where the detection module is optimized based on the DB algorithm, and the recognition module adopts the state-of-the-art text recognition algorithm SVTR.

Differentiable Binarization has been shown to be effective in handling various types of document images, including those with complex backgrounds, low contrast, and irregular-shaped text, due to the feature that it utilizes a segmentation-based text detection algorithm [21]. The goal of text detection is to accurately and efficiently locate text regions in an input image, regardless of their size, orientation, or font style. DB can improve text detection performance by allowing the segmentation network to learn the optimal threshold for binarization during training, instead of using a fixed threshold value that may not be optimal for all images. This approach enables the segmentation network to adapt to the variations in image brightness, contrast, and background complexity, resulting in more accurate and robust text detection. Furthermore, due to the differentiable characteristic, the gradient of the loss function can be backpropagated through the binarization step during training. Therefore, the end-to-end optimization of the segmentation network are achieved. This facilitates faster convergence and better results compared to traditional post-processing methods, which are often heuristic and non-differentiable.

The input image I is passed through the DB module to generate a binary image B (I; t) using the learned threshold parameter t:

$$B(I;\ t) = \begin{cases} 1, & if\ I \geq t \\ 0, & otherwise \end{cases} \tag{1}$$

The DB module applies a differentiable approximation of the binary step function to the input image, allowing the threshold t to be learned during training. The binary image B (I; t) is used to segment the text regions from the background using a segmentation network. Then the segmentation network is trained using a combination of binary cross-entropy loss and Dice loss:

$$Loss\ =\ -\frac{1}{N}\sum \begin{matrix}[y_i \log(p_i) +\ (1-y_i)\log(1-p_i)] \\ -\log\big(Dice(B(I;\ t), y)\big)\end{matrix} \tag{2}$$

where N is the number of pixels in the binary image, y_i is the ground-truth binary label of the i-th pixel, p_i is the predicted probability of the i-th pixel belonging to the text region, and Dice is the Dice loss, which measures the overlap between the predicted text regions and the ground-truth text regions. The binary cross-entropy loss term measures the pixel-wise classification error between the predicted probabilities and the ground-truth binary labels

style.

Scene Text Recognition with a Single Visual Model adopts a hierarchical approach for text recognition in images. The method begins by decomposing the image text into small patches referred to as character components. Afterwards, these character components undergo a series of hierarchical stages, which involve mixing, merging, and combining at the component level. To capture both inter-character and intra-character patterns, global and local mixing blocks are employed in the model. These blocks enable the perception of multi-grained character component information. By effectively integrating information from different levels of granularity, the recognition of individual characters is enhanced. The character recognition process in SVTR is achieved through a simple linear prediction, where the learned representations from the character components is leveraged to accurately predict the corresponding characters.

The internal architecture operates in three stages, progressively reducing the height of the input image. The input image is divided into character components, which are small patches associated with text characters. Mixing blocks and merging operations at different scales are employed to extract features from the character components. It captures stroke-like local patterns and inter-component dependencies to enhance the representation of character features. By combining the backbone network, component features, and dependencies, the SVTR is capable to encapsulate multi-grained character features. Finally, a parallel linear prediction with de-duplication is performed to obtain the character sequence, representing the recognized text.

3.3 Information Extraction

Key Information Extraction is a technique commonly employed to identify and extract particular pieces of information, such as name, address, and other relevant fields, from documents such as ID cards and forms [5].

Structure Location Alignment Network is a deep learning model for spatial layout analysis in document images. It implements spatial layout analysis and structural alignment through components such as PP-LCNet, CSP-PAN and SLAHead. SLANet can accurately locate and identify various parts of a resume, such as name, contact information, education experience, work experience, etc., and align their spatial location and structural information for subsequent information extraction.

PicoDet is a lightweight target detection model for embedded devices and mobile applications. It uses ESNet (Enhanced ShuffleNet) as the backbone network and CSP structure is used for feature connectivity and fusion between neighboring feature maps [6]. PicoDet can quickly and accurately detect the bounding box of key information in resume images and predict its class and confidence level.

The information module incorporates the SLANet and PicoDet network structures to implement resume information extraction. First, SLANet is responsible for spatial layout analysis of resume images to precisely locate and identify the position and content of each information block. Then, PicoDet performs target detection based on SLANet, detects the bounding boxes of key information, and outputs their categories and confidence levels.

Visual-feature Independent LayoutXLM is an enhanced version of LayoutXLM that focuses on improving inference speed while maintaining nearly lossless accuracy during downstream fine-tuning. This improvement is achieved by removing the visual backbone network module. By eliminating the visual backbone network module, VI-LayoutXLM streamlines the model architecture, resulting in faster inference times. Despite this modification, the model's accuracy remains largely unaffected, ensuring high-quality results in information extraction tasks.

The enhanced model strikes a balance between speed and accuracy, making it well-suited for scenarios where real-time or efficient processing is crucial without compromising the overall performance.

By passing the detected text to the VI-LayoutXLM model for SER, the output text information is extracted and combined and returned to information extraction module for named entity extraction. Through the combination of these models and algorithms, the accurate information extraction can be achieved.

3.4 Information Filtering

In general, it is better to filter resumes based on specific needs to meet the needs of the company. ERNIE-Layout is utilized to complete resume screening tasks based on specific Q&A through its powerful document intelligence capabilities.

ERNIE-Layout is an innovative approach to document pre-training that incorporates layout knowledge throughout the entire process. This enables the model to generate more effective representations by combining features from text, layout, and images [8]. ERNIE-Layout is an improved and extended model based on ERNIE, which is mainly used for document-level semantic understanding and information extraction. It contains two sub-models, ERNIE-Layout-X and ERNIE-Layout-Y. The former one is used to partition all the text in a document into multiple blocks, while the latter one is used to classify each block and determine whether the content in the block belongs to the information required by the user. ERNIE-Layout rearranges the token sequence with the layout knowledge and extracts visual features from the visual encoder. When performing Q&A task, users can provide a specific question as input, and the model will automatically match the semantic similarity between the question and the document to extract information related to the question in the document.

4. Module Evaluation

Our experiment employs a Chinese Talent Resumes Dataset for visualizing the results framework. In the following four sections, dataset, model performance, implementation details and results are demonstrated in details. The completeness of entity extraction from resumes are achieved and explained.

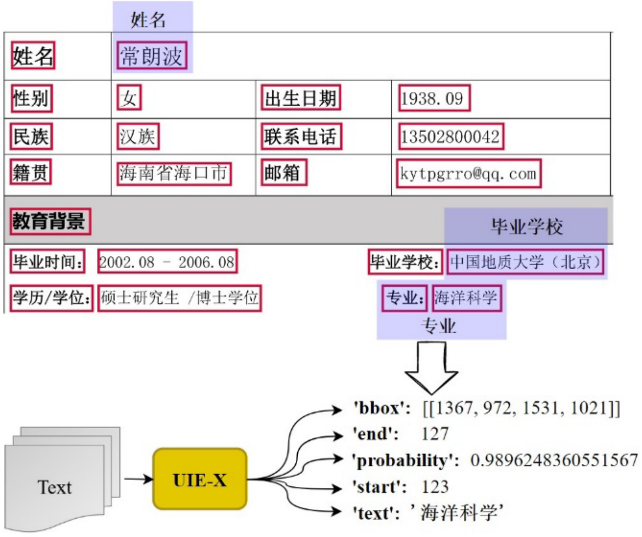
4.1 Dataset

In this study, our information extraction framework, as illustrated in Figure 1, is applied to the Chinese Talent Resume Dataset [22], which is a de-identification dataset including 2,000 manually generated resumes. Careful curation and annotation have ensured that the dataset encompasses various fields, such as personal information, educational background, work experience, and project involvement. The information is organized in the form of lists, where each "Graduation Institution, Academic Degree, Graduation Year" triplet represents an educational experience, each "Current Employer, Job Responsibilities, Job Title, Employment Duration" group represents work experience, and each "Project Name, Project Responsibilities, Project Duration" group represents project experience.

4.2 Model Performance

For identifying key entities from documents, the SER task is

conducted using VI-LayoutXLM model. The model is trained in XFUND [23], which is a multilingual form understanding benchmark dataset including form understanding samples in 7



姓名	常朗波		
性别	女	出生日期	1938. 09
民族	汉族	联系电话	13502800042
籍贯	海南省海口市	邮箱	kytpgrro@qq. com
教育背景			
毕业时间:	2002. 08 – 2006. 08	毕业学校:	中国地质大学（北京）
学历/学位:	硕士研究生 /博士学位	专业:	海洋科学

```
'bbox':  [[1367, 972, 1531, 1021]]
'end':   127
'probability':  0.9896248360551567
'start':  123
'text':  '海洋科学'
```

Figure 2: Illustrative Examples of Entity Extraction Results

languages (Chinese, Japanese, Spanish, French, Italian, German, and Portuguese). Each sample in XFUND has undergone manual labeling to annotate key-value pairs specific to each language.

For table recognition, we employed the pre-trained SLANet model and evaluated its performance on the PubTabNet dataset. PubTabNet is a large-scale benchmark comprising 568k+ images of tabular data annotated with HTML representations [24]. The algorithm achieved an accuracy of 76.31% in recognizing table structures, a table entity detection score (TEDS) of 95.89% in capturing table information, and exhibited a processing speed of 766ms per image inference on a CPU machine with MKL enabled.

For the task of document layout analysis, the pre-trained model PicoDet is utilized to perform the task on the CDLA [25] dataset (Chinese Document Layout Analysis dataset). CDLA is a comprehensive dataset that encompasses various document elements, such as text, title, figure, figure caption, table, table caption, header, footer, reference, and equation. 6,000 annotated images were included in this dataset, of which 5,000 images were used for training and 1,000 images for validation. The evaluation results demonstrated the effectiveness of the model, achieving a Mean Average Precision of 86.8% in accurately identifying and localizing document elements.

For the document question answering task, the ERNIE-Layout model is trained on the DocVQA dataset [26], which is specifically designed for visual question answering on document images. The dataset comprises over 12,000 document images accompanied by 50,000 associated questions. An impressive Average Normalized Levenshtein Similarity [27] score of 0.8321 in DocVQA is achieved by ERNIE-Layout, indicating its competence in accurately addressing questions based on the content within the document images.

4.3 Implementation Details

The resume files are converted from PDF format to JPEG image format firstly, and then the text contents are obtained by OCR. We set the attention dropout probability and hidden layer dropout probability as 0.1 to mitigate overfitting risks. And the image feature pooling shape as [7, 7, 256] for pooling and feature extraction. To accomplish the table recognition task, the hidden size and intermediate size of SLNet model are set as 768 and 3072 respectively. And 12 attention heads and12 hidden layers are utilized to enhance the model's representational power and inference capabilities. We also specified the layer normalization epsilon as$1e^{-12}$ to improve training and convergence. For performing the layout analysis, the PicoDet the coordinate size of PicoDet model is set to 128 to handle coordinate information in the images. And the size of hidden layers is set to 768 provide sufficient modeling capacity. Additionally, we enabled relative attention bias and spatial attention bias to facilitate the model's understanding of the document layout structure. Furthermore, the maximum relative 2D position embeddings are defined as 256 and the maximum relative position embeddings is defined as 128 to capture relative positional information.

To train the VI-LayoutXLM model, we employed a combination of loss functions, including distillation loss, SER-DML loss, and distance-based losses, which are carefully weighted to optimize the training process. The AdamW optimizer is employed with certain hyperparameter settings: a learning rate set to $5e^{-5}$, β_1 of 0.9 and β_2 of 0.999. L2 regularization was also applied to prevent overfitting. The training dataset, XFUND_zh is proceeded by using various data transformations such as decoding, encoding, padding, resizing, and normalization

These transformations are intended to enhance the understanding of the model in the layout and structure of the document while being independent of the visual features.

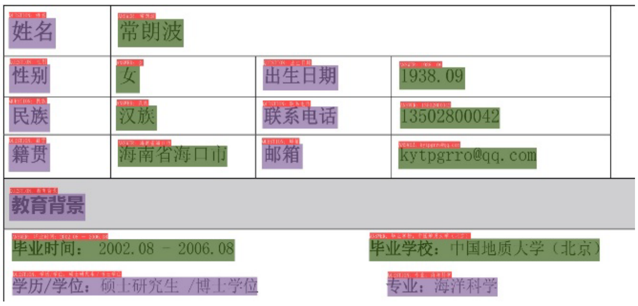
The ERNIE-Layout model underwent training for 6 epochs, employing a linear learning rate scheduler with a warm-up ratio of 0.05. In order to mitigate overfitting, attention dropout probability and hidden layer dropout probability of 0.1 were applied. The model was evaluated every 10,000 steps, and the best model was loaded at the end of training. The document's maximum sequence length was set to 128, and the target size was set to 1000. We used a batch size of 8 for training and performing evaluation on each device. The learning rate was set to $2e^{-5}$, and weight decay was disabled (set to 0). The experiment was conducted using 16 training shards, and a random seed of 1000 was used for reproducibility.

4.4 Schema-based Information Extraction

To visualize the performance of our model on a sample, a test image was selected, and the information was extracted and visualized. Figure 2 shows a partial intercept of one sample of original images and the named entity information extracted by the model. For example, the pattern is set to ['姓名', '专业', '毕业学校'], representing name, major and graduation institution, respectively. The defined entities are successfully recognized with a confidence level of over 98%. Meanwhile, the start and end positions are detected correctly with the rectangle markers attached.

4.5 Key Information Extraction

For semantic entity recognition of documents, VI-LayoutXLM algorithm is utilized. Figure 3 shows the visualization results of a partial interception of a sample document. Based on our results, it is observed that the model performs well in terms of key information extraction. Most of the keys and values can be accurately identified, showing the model's understanding of the document structure and its ability to extract key information. However, in some cases, there is room for improvement in the accuracy of the model in terms of question-and-answer determination.



姓名	常朗波		
性别	女	出生日期	1938.09
民族	汉族	联系电话	13502800042
籍贯	海南省海口市	邮箱	kytpgrro@qq.com

教育背景

毕业时间：2002.08 – 2006.08　毕业学校：中国地质大学（北京）

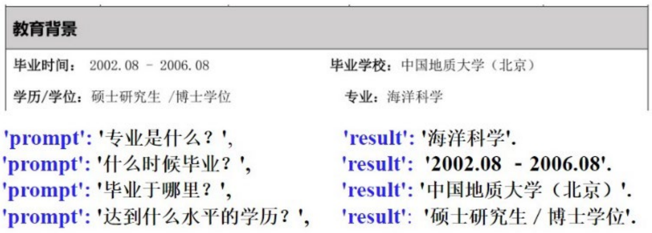
学历/学位：硕士研究生 /博士学位　专业：海洋科学

Figure 3: Illustration of SER Prediction Results

Then the SER prediction results are saved to the output file and then modified to json format, which already includes the text information we want. Then the file is read again for text stitching, and then named entities are extracted just as in the previous job. The accuracy shows an improvement of completeness of resume information extraction, due to the fact that VI-LayoutXLM SER is fine-tuned based on a form dataset that is particularly close to the form of the resume file.

4.6 Document Intelligent Question-Answering

To tackle the Q&A function, ERNIE-Layout is leveraged as a backbone of question-answering approach. Subsequently, a token-level classifier is built upon the output representation of ERNIE-Layout to predict the start and end positions of the answer. Compared with the previous SER task when we used VI-LayoutXLM, the questions and answers corresponding to the graduation time as well as the graduation school were identified as questions; the questions and answers for both degree and major were identified as answers, which indicates that the model is still lacking in locating the questions and answers. ERNIE-Layout, on the other hand, is able to accurately extract the answers related to the questions from the text under the guidance of the given questions, as shown in Figure 4. According to the prompt we set, it can locate and extract the correct result more accurately by understanding the semantics and context of the question and combining the layout information in the text.



教育背景

毕业时间： 2002. 08 – 2006. 08　　毕业学校：中国地质大学（北京）

学历/学位：硕士研究生 /博士学位　　专业：海洋科学

```
'prompt': '专业是什么？',          'result': '海洋科学'.
'prompt': '什么时候毕业？',        'result':  '2002.08  - 2006.08'.
'prompt': '毕业于哪里？',          'result': '中国地质大学（北京）'.
'prompt': '达到什么水平的学历？',  'result':  '硕士研究生 / 博士学位'.
```

Figure 4: Examples of Q&A results for ERNIE-Layout

5. Conclusion

In this paper, we develop IntelliExtract, an end-to-end framework specifically designed for Chinese resume information extraction from document images. The proposed system integrates various components including OCR, layout analysis, information extraction, and document understanding techniques to enable efficient and accurate processing of document data. To achieve this, some recent models and algorithms are employed, such as OCR models for converting scanned documents into machine readable text, layout analysis algorithms for understanding the spatial arrangement of document elements, and information extraction techniques for extracting structured data from unstructured documents. However, it is important to note that our system still faces some challenges. For example, it may encounter difficulties in accurately extracting information from documents with low image quality or complex formatting. Furthermore, the system's performance may vary depending on the specific characteristics of the document dataset being processed.

In conclusion, our research contributes to the field of document information extraction by proposing a comprehensive end-to-end system. The system's integration of OCR, layout analysis, information extraction, and document understanding techniques enables efficient and accurate processing of document data. Despite existing challenges, our work sets the foundation for further advancements in document intelligence and paves the way for future research in this area.

References

[1] Yan, H., Deng, B., Li, X., & Qiu, X. (2019). TENER: adapting transformer encoder for named entity recognition. arXiv preprint arXiv:1911.04474.

[2] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[3] Chen, Y., Xu, L., Liu, K., Zeng, D., & Zhao, J. (2015, July). Event extraction via dynamic multi-pooling convolutional neural networks. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pp. 167-176).

[4] Du, Y., Chen, Z., Jia, C., Yin, X., Zheng, T., Li, C., ... & Jiang, Y. G. (2022). Svtr: Scene text recognition with a single visual model. arXiv preprint arXiv:2205.00159.

[5] Li, C., Guo, R., Zhou, J., An, M., Du, Y., Zhu, L., ... & Yu, D. (2022). PP-StructureV2: A Stronger Document Analysis System. arXiv preprint arXiv:2210.05391.

[6] Yu, G., Chang, Q., Lv, W., Xu, C., Cui, C., Ji, W., ... & Ma, Y. (2021). PP-PicoDet: A better real-time object detector on mobile devices. arXiv preprint arXiv:2111.00902.

[7] Peng, Q., Pan, Y., Wang, W., Luo, B., Zhang, Z., Huang, Z., ... & Wang, H. (2022). ERNIE-Layout: Layout Knowledge Enhanced Pre-training for Visually-rich Document Understanding. arXiv preprint arXiv:2210.06155.

[8] S. Gupta, G. Kaiser, D. Neistadt, and P. Grimm, "DOM-based Content Extraction of HTML Documents," Defense Technical Information Center, 2003.

[9] P. M. Joshi and S. Liu, "Web document text and images extraction using DOM analysis and natural language processing," in Proceedings of the 9th ACM Symposium on Document Engineering, (DocEng '09), pp. 218–221, Germany, September 2009.

[10] F. Sun, D. Song, and L. Liao, "DOM based content extraction via text density," in Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, (SIGIR '11), pp. 245–254, China, July 2011.

[11] Al-Amoudi, A., Alomari, A., Alwarthan, S., & Rahman, A. (2021). A Rule-Based Information Extraction Approach for Extracting Metadata from PDF Books. ICIC Express Letters, 12, 121-132. https://doi.org/10.24507/icicelb.12.02.121

[12] Ayishathahira, C. H., Sreejith, C., & Raseek, C. (2018). Combination of Neural Networks and Conditional Random Fields for Efficient Resume Parsing. In 2018 International CET Conference on Control, Communication, and Computing (IC4) (pp. 388-393). IEEE. doi: 10.1109/CETIC4.2018.8530883.

[13] Tobing, B. C. L., Suhendra, I. R., & Halim, C. (2019). Catapa Resume Parser: End to End Indonesian Resume Extraction. In Proceedings of the 3rd International Conference on Natural Language Processing and Information Retrieval (NLPIR) (pp. 68-74). Association for Computing Machinery. doi: 10.1145/3342827.3342832

[14] Nahm, U.Y. (2005). Transformation-Based Information Extraction Using Learned Meta-rules. In: Gelbukh, A. (eds) Computational Linguistics and Intelligent Text Processing. CICLing 2005. Lecture Notes in Computer Science, vol 3406. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-30586-6_57

[15] Téllez-Valero, A., Montes-y-Gómez, M., Villaseñor-Pineda, L. (2005). A Machine Learning Approach to Information Extraction. In: Gelbukh, A. (eds) Computational Linguistics and Intelligent Text Processing. CICLing 2005. Lecture Notes in Computer Science, vol 3406. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-30586-6_58

[16] Gaur, B., Saluja, G.S., Sivakumar, H.B. et al. Semi-supervised deep learning based named entity recognition model to parse education section of resumes. Neural Comput & Applic 33, 5705–5718 (2021). https://doi.org/10.1007/s00521-020-05351-2

[17] Xu, Q. et al. (2021). An Effective Algorithm for Classification of Text with Weak Sequential Relationships. In: Strauss, C., Kotsis, G., Tjoa, A.M., Khalil, I. (eds) Database and Expert Systems Applications. DEXA 2021. Lecture Notes in Computer Science(), vol 12924. Springer, Cham. https://doi.org/10.1007/978-3-030-86475-0_28

[18] Zu, S., & Wang, X. (2019). Resume information extraction with a novel text block segmentation algorithm. Int J Nat Lang Comput, 8, 29-48.

[19] ÇELİK, D. (2016). Towards a semantic-based information extraction system for matching résumés to job openings. Turkish Journal of Electrical Engineering & Computer Sciences, 24(1), 143-157. doi: 10.3906/elk-1304-130.

[20] Hou, Y., Tao, L. (2019). Semantic-Based Resume Screening System. In: Arai, K., Bhatia, R., Kapoor, S. (eds) Proceedings of the Future Technologies Conference (FTC) 2018. FTC 2018. Advances in Intelligent Systems and Computing, vol 880. Springer, Cham. https://doi.org/10.1007/978-3-030-02686-8_49

[21] Liao, M., Wan, Z., Yao, C., Chen, K., & Bai, X. (2020, April). Real-time scene text detection with differentiable binarization. In Proceedings of the AAAI conference on artificial intelligence (Vol. 34, No. 07, pp. 11474-11481).

[22] Hainan Provincial Big Data Management Bureau. (2020). Desensitized Chinese Talent Resume Data and Annotation [Dataset description]. In The 2nd Hainan Big Data Innovation Application Competition - Intelligent Algorithm Competition. Retrieved from https://tianchi.aliyun.com/competition/entrance/231771/information

[23] Xu, Y., Lv, T., Cui, L., Wang, G., Lu, Y., Florencio, D., ... & Wei, F. (2021). Layoutxlm: Multimodal pre-training for multilingual visually-rich document understanding. arXiv preprint arXiv:2104.08836.

[24] Zhong, X., ShafieiBavani, E., & Jimeno Yepes, A. (2020). Image-based table recognition: data, model, and evaluation. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16 (pp. 564-580). Springer International Publishing.

[25] Hang, L. 2021. CDLA: A Chinese document layout analysis (CDLA) dataset. https://github.com/buptlihang/CDLA/.

[26] Mathew, M., Karatzas, D., & Jawahar, C. V. (2021). Docvqa: A dataset for vqa on document images. In Proceedings of the IEEE/CVF winter conference on applications of computer vision (pp. 2200-2209).

[27] Biten, A. F., Tito, R., Mafla, A., Gomez, L., Rusinol, M., Mathew, M., ... & Karatzas, D. (2019, September). Icdar 2019 competition on scene text visual question answering. In 2019 International Conference on Document Analysis and Recognition (ICDAR) (pp. 1563-1570). IEEE.