International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor (2018): 7.426

Improved Information Extraction with NLP & CRF for Invoices

Monalin Pal

M.Tech. Software Systems – Data Analytics, BITS Pilani, Rajasthan, India

Abstract: ***In a general Invoice processing system, the process starts from manually uploading the different format of invoices like paper, pdf, excel, document, image etc. for handling a supplier invoice, from its receipt to upload in the ERP system and is ready for payment. This invoice process is manually handled for any organization. Each process of manual upload and payments was time consuming. This is an attempt to use AI in invoice processing system, where we can process invoices automatically by extracting the relevant information from the invoices and feeding it to the ERP systems through RPA (Robotic Process Automation). This process will help to reduce costs while improving both the accuracy and the speed of data extraction. Invoice Processing Advisor that can read invoice and extract the accurate information in a representable format. Annotations and Machine Learning for data extraction from the preprocessed data. Intelligent Invoice Processing will help to cater the increasing business needs as we see that majority of organizations spend lot of man FTE's and time for processing invoices. With the help of AI, this solution will be helpful to extract the information from the invoices more accurately and reduce the manual effort as well as the cost. We can easily scale it to multiple type of invoices as well introduce various business rules check that can communicate with the RPA's to perform any type of complex actions.***

Keywords: Artificial Intelligence, CRF Model, Natural Language Processing, Python.

1. Introduction

Invoice processing mainly refers to the complete process for handling a supplier invoice, from its receipt to upload in the ERP system and is ready for payment. An invoice can be obtained in different formats – paper, pdf, excel, document, image etc. and the invoice data must be scanned or manually updated in the recipient's ERP system by an agent. Invoice Processing when handled manually, is a very expensive and time-consuming process for any organization. By introducing AI in invoice processing system, we can process invoices automatically by extracting the relevant information from the invoices and feeding it to the ERP systems through RPA (Robotic Process Automation). This will help to reduce costs while improving both the accuracy and the speed of data extraction.

2. Process

2.1 Data Ingestion/Extraction

In this process, we receive the invoice data in various formats viz excel, word, pdf, image, handwritten etc. We can receive the attachments through various channels like email, web service API, SFTP Folder etc. In our current scenario, we will be receiving the attachments through web service API. API will send us the attachments along with other information like from, subject, body, to, date from the email.

2.2 Data Annotation

Once we receive the attachment, we must convert the attachment to text file format that can be further used for annotation. In our case we will be considering the pdf attachments and will convert them to text. We have various tools available for annotation which are html based and standalone tools. We will be using eHost tool for annotation.

2.3 Model Train

We will use the annotated files and converted text files for feature extraction and generation. We will use the features for building the model. We will use CRF algorithm and pycrfsuite python module to train the model.

2.4 Post Processing

Once the model is trained, we will test on the test invoices and perform the prediction. Predicted output needs to be cleaned and postprocessed. We will be creating various functions like remove special characters, split address etc.

2.5 Response Generation

This is final step where we will generate the response json from the post processed output that can be utilized further for any RPA to consume.

3. Architecture

Architecture will be based on a web service framework where it can be deployed as an application in any ERP tool or as an integration with various customer service platform like Salesforce or Mainframe. We will be using below tools for architecture

- Python 2.7, Django Web Service Framework
- PDFtoTxt standalone Module
- Ehost Annotator, PyCRFSuite, Pandas, Json

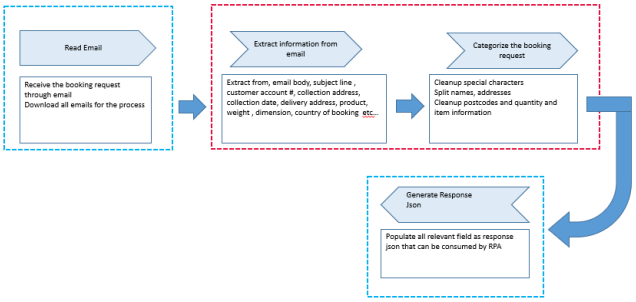

Figure 1: Architecture Design

4. Implementation

4.1 Read Email

We will be receiving an email with an attachment through a web service API. The fields that we receive through the API would be from, to, subject, body, date and attachment. Below is sample for request.json that will be fed to Intelligent Invoice Processing service.

```
{
    "attachments": "invoice.pdf",
    "body": "Can I please get this booked on for tomorrow. Thanks",
    "from_address": "Monalin.pal@gmail.com",
    "has_attachment": "True",
    "request_id": "5143",
    "subject": "IVOR",
    "to_address": "iip@gmail.com"
}
```

Figure 2: POST request details

Typical Invoice looks as below

Fax: 0987654321 Email: test@test.com

Account Number	13456789
Account Name	John
Order Contact Name	Mike Taylor
Order Contact Telephone Number	012345 56789
Reference For Your Invoice	1j654rty
Reference For Driver To Quote	Collection of service

Collection Information	
Company Name	Mac Transports
Company Address	Graphite India Hoodi Bangalore Postcode : 560048
Collection Contact Name: mike	
Collection Contact Telephone Number: 0912345678	
Delivery Information	
Company Name	John Medicals
Company Address	Chandni Chowk New Delhi India Postcode : 110006
Receiver Contact Name: Smith	
Receiver Contact Telephone Number: 012345678	

4	Number of items	08/01/2018	Ready Date
80	Total Weight	10:30	Ready Time
65 x 65 x 65	Dimensions of your parcel(s)	16:30	Close Time
			Lunch

Next day	Service Required
boxes	How are the goods packaged?
UPS	Description of goods
UN No: Packing Group:	Dangerous Goods Details

Is confirmation of booking required?	YES	
Fax / Email confirmation to:	confirm@iip.com	
Booking Reference:	Agent:	Origin:

Figure 3: invoice.pdf

4.2 Extract Information from Email

Once we received the attachment through web API, we will download the attachment, convert it to text and predict the various fields using trained CRF Model.

4.3 Categorize the Booking request

We will cleanup the predicted output by post processing functions like remove special characters, split the addresses, postcode and item description.

4.4 Generate Response Json

After post processing, fields will be generated in a representable format as Json that can be fed in to a RPA. Below sample shows response.json from Intelligent Invoice Processing Service. Below is sample response.json with extracted values


```
{
        "collection_address_city_town": {
                "value": "Bangalore",
                "conf_score": 0.961971920166791
        },
        "collection_contact_phone_number": {
                "value": "0912345678",
                "conf_score": 0.977901377317948
        },
        "collection_org_name": {
                "value": "Mac Transports",
                "conf_score": 0.996110395291647
        },
        "collection_address_1": {
                "value": "Graphite India",
                "conf_score": 0.663555245757666
        },
        "collection_address_2": {
                "value": "Hoodi",
                "conf_score": 0.921595653434297
        },
        "delivery_contact_first_name": {
                "value": "John Medicals",
                "conf_score": 0.948230336218706
        },
        "delivery_contact_phone_number": {
                "value": "012345678",
                "conf_score": 0.948230336218706
        },
        "delivery_address_1": {
                "value": "Chandni Chowk",
                "conf_score": 0.992502920208482
        },
        "delivery_address_city_town": {
                "value": "Delhi",
                "conf_score": 0
        },
        "delivery_address_postal_code": {
                "value": "110006",
                "conf_score": 0.976313278857954
        },
        "goods_description_of_goods": {
                "value": "UPS",
                "conf_score": 0
        },
```


Figure 4: Web Service Response Json

5. Other recommendations

We can use deep learning to improve the data extraction process as we see all the fields are not extracted accurately.
We can use distance metrics to extract values around a particular key.

References

[1] Hamza H, Belaïd Y, Belaïd A. "A case-based reasoning approach for invoice structure extraction". Document

Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on. 2007 Oct: p. 327-331.
[2] Aamodt A, Plaza E. "Case-based reasoning: Foundational issues, methodological variations, and system approaches". AI communications. 1994 Mar: p. 39-59.
[3] Sorio E, Bartoli A, Davanzo G, Medvet E. "A Domain Knowledge-based Approach for Automatic Correction of Printed Invoices". Information Society (iSociety), 2012 International Conference on. 2012 Jun: p. 151-155.
[4] Kulkarni P. "Knowledge Augmentation: A Machine Learning Perspective". In Reinforcement and Systemic Machine Learning for Decision Making.: WileyIEEE Press; 2012. p. 209 - 236.
[5] AR Kinjo, F Rossello, G Valiente. Profile Conditional Random Fields for Modeling Protein Families with Structural Information. BIOPHYSICS. 2009; 5: 37-44.

Author Profile

Monalin Pal received the Mtech Degree in Software Systems with Data Analytics from BITS, Pilani. She has done her Btech in Computer Science from Utkal University. Currently she is working as a Senior Analyst at Accenture and delivering values to clients in Artificial Intelligence, NLP, Big data.