Intelligent Data Capture in Digitized Business Documents

Rolandas Gricius and Igoris Belovas

E-mails: rolandas.gricius@mif.stud.vu.lt, igoris.belovas@mif.vu.lt

Vilnius University, Institute of Data Science and Digital Technologies

Abstract

Today most documents are produced in digital format directly, removing the need for OCR. Unfortunately, documents are primarily in free form. Thus data and information still need to be extracted for further processing in information systems. We aim to present possible approaches in dealing with this problem, including semantic frame-based approach, supervised and unsupervised learning and other techniques.

Introduction

The research area is in the realm of documents, more specifically - business documents used to conduct business. There is no single notion of the business document in the literature, but most authors agree on the main features of them. Important ones to note are:

- Business documents typically have some structure, i.e., they are semi-structured (as invoices) or structured (as CMR waybills)
- Business documents mandatorily contain some predefined sets of data (typically minimum requirements defined by law)
- Some data inside the document has semantic relations to other data (company code, name and address, invoice amount, tax rate and total)

The task of information extraction from business documents is actively researched as more and more documents are created. Typical steps of this process are defined in the literature:

1. *Extraction of characters* (using OCR to extract from scanned images, parsing complex document formats, such as PDF or HTML) and text in general.
2. *Identification of layout* combined with *document classification* to decide on how it should be further processed.
3. *Information extraction* for the purpose of storing relevant document contents in some data stores.

This last step is of main interest in this overview.

Main approaches

One of the first natural ideas would be to design the process the same way human accomplishes this task. It would be by understanding the document and extracting information as required. That is the basis of *the semantic-based approaches*.
Another type of approach would be to provide document examples, to show the information which should be extracted and to use some *supervised machine learning* technique to train the model.
As supervised learning requires a lot of data labeling, which is a substantial effort, *unsupervised learning* options are explored.
Finally, for special simpler cases, template and rule based systems are employed, and for specific areas of a document, such as tables, special techniques are developed.

Semantic-based approaches

The semantic-based approach tries to understand the meaning (semantics) of the text block as a whole and then retrieve needed parts (attributes) from it.
The text of a document is divided into constituent parts, and then the semantic model of each part and of the document as a whole is created. The classical approach is to map meaning into scenarios, described by semantic frames. For that purpose ones should be created beforehand. FrameNet is one of the popular semantic frame repositories used for this purpose.
Then Semantic role labeling (SRL) technique is typically used to extract information based on the meaning of sentences.
Finally, created frames are inspected, and information is extracted from the slots.
Alternatives to semantic frames for scenario representation, such as Schank schemas, are researched as well in semantic-based approaches.

Supervised learning

Supervised learning uses tagged documents as training examples to learn field extraction.
The most widely used techniques are Support Vector Machines (SVMs), K-nearest neighbour algorithm and Neural networks.
Support Vector Machines are useful for text classification tasks; field value is a result of assigning the corresponding class.
K-nearest neighbour algorithm helps in two ways: on the spatial perspective it helps to find specific text proximity and attribution to text block on the page, and on text analysis perspective it can find text similarities.
The neural network is yet another classification option, very well suited for more complicated cases, but requires more training and computational resources.
Newer supervised learning approaches use structured prediction learning methods. The structured prediction uses the fact that business documents have structure, data fields typically are near their labels, related data are often grouped. Main structured prediction techniques are conditional random fields, structured SVMs, Recurrent neural networks.
Conditional random fields take context into account, creating a graph of the predictions.
Structured SVMs can produce a parse tree of the document, showing the roles of text fields and their relations.
Recurrent neural networks are useful at text sequence labeling, again taking into account text connectivity.
Supervised learning techniques are very well established in the reviewed published research.

Unsupervised learning

The main shortcoming of supervised learning is the amount of work and human involvement to train the models. Thus, interest in unsupervised learning is growing.
Tasks approached by unsupervised learning include finding patterns within document data, similar document clustering, semantic annotation. The most common techniques are K-means clustering, Hierarchical clustering and Neural networks.
K-means clustering helps to find similarities between textual data (either inside a document or between documents).
Hierarchical clustering additionally creates hierarchies of clusters, so different aggregation levels can be explored.
Neural networks in unsupervised learning approach can also find clusters, extract named entities and do further semantic annotation.
There are observations in the literature that unsupervised learning approaches may have worse precision, but higher recall.

Other techniques

Some researchers employ other, often more straightforward, but effective techniques, especially useful when the document domain is narrow and/or the document is structured. In these cases, there is additional information from the domain knowledge (limited number of possible fields, static field positions).
The simplest one is to use predefined templates and start with document classification to select the correct template.
Another technique to choose a template is to use the document issuer logo to assign the document to the class. This leads to one class per document issuer, even if several of them use basically the same document structure (for example, by using the same software to produce documents).
Then more advanced technique addresses this by comparing strings and their positions in the document and assigning the class based on similarity ranking.
An important sub-field of research in information extraction is table extraction. Here techniques employ the regular structure of the tables and relations of column values, such as total = price * quantity.

Conclusions

We have surveyed the field of intelligent data capture in digitized business documents, and have found that this field is full of relevant ideas on how to apply the newest AI and ML approaches.
Supervised learning takes the most significant part in the field, but simpler techniques are useful in special cases, and emerging research tries the unsupervised learning approach.