Method of Fact Checking Contents in Non-Digital Documents Against Trade Finance Criteria

Technical Field

The present invention relates to a method of fact checking contents in a non-digital document against trade finance criteria which consists of one or more predefined rules for different document types.

Background

Business documents are critical to a company's efficiency and productivity. The business documents may be digital-born, occurring as electronic files, or they may be in scanned form that comes from reproducing written or printed documents. It is a very challenging and labour intensive task to check facts in non-digital business documents due to the diversity of layouts, poor quality of scanned documents as well as the complexity of template structures.

Currently, many companies extract data from business documents through manual efforts, which is very time-consuming and prone to high operation risks due to human errors. Besides, knowledge is not retained due to human attrition which leads to high training costs. In addition, trade finance criteria that consist of rules for each type of document often need to be hard-coded and updated with changes to the specific format. It is necessary to develop an automated approach in order to reduce the turnaround time to process each application and to allow companies and institutions (e.g. banks) to grow their trade finance business exponentially without facing scalability problems.

Automatic approaches typically focus on data digitization, which aims to convert scanned documents into digital formats. However, to better

understand and analyse business documents, it is indispensable to automatically classify, extract and structuralize information from business documents and to automatically perform cross-document check as well. Document AI, or Document Intelligence, is a new research topic that refers to the techniques for automated reading, understanding and analysing business documents. Contemporary approaches for Document AI need to be further built upon deep neural networks from a computer vision perspective or a natural language perspective, or a combination of them.

Although the previous Document AI models have made significant progress, most of them confronted the following limitations. First, they do not have enough financial related documents to train pre-trained language models which are better suited for financial extraction tasks. Second, they are not designed for the trade finance scenario which has standard but complex rules and various document layouts.

It is generally desirable to overcome or ameliorate one or more of the above described difficulties, or to at least provide a suitable alternative.

Summary

Disclosed herein is a computer-implemented method of generating a report including fact checking contents in a non-digital document against one or more predefined rules, comprising:

(a) providing, by an input module of a processing server, a document in an at least partially non-searchable format;

(b) digitizing, by a digitization module of the processing server, the document to produce a fully searchable, digitized document; and

(c) using an extraction module of the processing server to:

(i) determine one or more data objects for extraction;

(ii) extract the one or more data objects from the digitized document; and

(iii) predict the context of the one or more data objects extracted, based on content of the one or more data objects;

(d) generating code corresponding to the one or more predefined rules; and

(e) generating, by a verification module of the processing server, a report including whether the one or more data objects meet the one or more predefined rules.

Digitizing the document may comprise identifying a plurality of contents in the document, each content being from a set of desired contents, identifying location information of each content in the document, identifying layout information of the document, categorizing a plurality of said contents into searchable contents and non-searchable contents, and converting the non-searchable contents into searchable format.

Using the extraction module may comprise locating one or more data objects according to the layout information, and predicting the context of the one or more data objects.

Predicting the context may comprise training the extraction module to predict the context, by:

(c1) providing, to a neural network of the processing server, a training corpus in which a plurality of training contents, and context associated with each training content, have been identified from one or more training documents;

(c2) pre-training the neural network by:

(i) masking a portion of the context associated with a plurality of said training contents;

(ii) making a prediction of each masked portion based on an unmasked portion of the respective context; and

(iii) fine-tuning the prediction of each masked portion using an additional prediction layer that receives an embedding for the respective masked content and a classification ([CLS]) token embedding for at least one of the respective training document and the location information of the content associated with the masked portion;

(c3) training the pre-trained neural network on a domain-specific corpus from a domain comprising the document;

(c4) applying the trained neural network to the digitized document, to predict the context of the one or more data objects;

(c5) applying template based regular expression extraction method to predict the context of the one or more data objects; and

(c6) combining the results from (c4) and (c5) to give the final predictions.

Providing a training corpus may comprise providing non-searchable content having particular layout information, and training the extraction module to predict the context of the one or more data objects. Providing a training corpus may further comprise providing a training corpus of documents from the domain.

The [CLS] token embedding may be a [CLS] token embedding for the document.

The domain may be a financial domain.

Determining the context comprising the respective masked portion may be based on an embedding for a document in which the masked content was identified, and an embedding for the content determined prior to masking.

The context may be further determined based on an embedding for a location at which the masked portion appears in the document.

Disclosed herein is also a system of generating a report including fact checking contents in a non-digital document against one or more predefined rules, comprising:

(a) an input module of a processing server configured to provide a document in an at least partially non-searchable format;

(b) a digitization module of the processing server configured to digitize the document to produce a fully searchable, digitized document; and

(c) an extraction module of the processing server configured to:

(i) determine one or more data objects for extraction;

(ii) extract the one or more data objects from the digitized document; and

(iii) predict the context of the one or more data objects extracted, based on content of the one or more data objects; and

(d) a code generator configured to generate code corresponding to the one or more predefined rules; and

(e) a verification module of the processing server configured to generate a report including whether the one or more data objects meet the one or more predefined rules.

In the present context, a "corpus" refers to documents which can include for example, financial reports.

Brief description of the drawings

Preferred embodiments of the invention are hereafter described, by way of non-limiting example only, with reference to the accompanying drawings, in

which:

Figure 1 illustrates a computer-implemented method used for fact checking contents in a non-digital document against trade finance criteria which consist of one or more predefined rules;
Figure 2 illustrates a mechanism of digitizing a document to produce a fully searchable, digitized document;
Figure 3 illustrates a mechanism of predicting the context of the data objects extracted from the document;
Figure 4 illustrates an architecture of the pre-training model;
Figure 5 illustrates an example of fine-tuning model.

Detailed description

The computer-implemented method 100 set out in Figure 1 is used for generating a report including fact checking contents in a non-digital document against one or more predefined rules. The method, comprising:

(a) Step 101: providing, by an input module of a processing server, a document in an at least partially non-searchable format; and
(b) Step 102: digitizing, by a digitization module of the processing server, the document to produce a fully searchable, digitized document; and
(c) Step 103: using an extraction module of the processing server to:
    (i) extract one or more data objects from the digitized document; and
    (ii) predict the context of the one or more data objects extracted, based on content of the one or more data objects; and

(d) Step 104: generating, by a verification module of the processing server, a report including whether the one or more data objects meet the one or more predefined rules.

The present providing step 101 is intended to provide the non-searchable scanned financial documents such as bank invoices, bill of lading, bill of exchange to the next step for format conversion. In the present context, the term "non-searchable" has a meaning substantially the same as "non-digital", insofar as the non-digital document cannot be searched in its current form using a computer program. A digitized document refers to one that is searchable in all relevant parts (i.e. the parts that need to be analysed). In one embodiment, only the portions of a document in a partially non-searchable format that are unsearchable will be converted by the digitizing step 102 to a searchable format.

The present digitizing step 102 may consist of one or more of the following five sub-steps set out in Figure 2. These sub-steps are:

(a) The first sub-step 201 is to identify a plurality of contents in the document, each content being from a set of desired contents.

(b) The second sub-step 202 is to identify the location information of each content in the document. The location information can be obtained by applying Optical Character Recognition (OCR) to the document. This invention uses commercial OCR platforms (e.g. Google OCR, Baidu OCR) to generate text information with position information. In one embodiment, the digitization module may treat the scanned image provided by the input module as input, and identify the contents such as plain text, table, stamp and signature as well as the corresponding coordinate information of each content. The location of content in a document can assist in determining a context for that content – e.g. if content is towards the end of a document,

it is more likely to be conclusory (i.e. relating to the ultimate outcome of that which was set out earlier in the document).

(c) The third sub-step 203 is to identify layout information of the document. The layout information may contain columns forming a grid that has the appearance of, for example, a table or accounting ledger, as well as additional information such as formatting information, column grouping, column width, and so on. The relative positions of words in a document contribute a lot to the semantic representation. Taking form understanding as an example, given a key in a form, e.g. Passport ID, its corresponding value is much more likely on its right or below instead of on the left or above. In particular, the type of a document can be further identified based on the corresponding layout information. For example, financial numbers can be identified since a vertical column of financial numbers will often have the total at the bottom, or near the bottom and offset to the right. In one embodiment, the layout information may also be obtained by applying OCR to the document.

(d) The fourth sub-step 204 is to categorize a plurality of said contents into searchable contents and non-searchable contents.

(e) The fifth sub-step 205 is to convert the non-searchable contents into searchable format. In one embodiment, the digitization module may categorize the contents into text contents which are searchable using a computer program, and non-text contents such as images and tables which are non-searchable using a computer program (sub-step 204). In this embodiment, the digitization module may take the non-searchable contents as input and generate text contents output.

The present extracting step 103 is intended to extract one or more data objects from the document digitized by step 102, and predict the context of the one or more data objects extracted, based on content of the one or more data objects. The purpose of this step is to read and comprehend the

documents processed by step 102 so as to identify what the data objects (contents) represent. For example, Google LLC, January 2019, $12 in a document should be recognised as a company name, date and invoiced amount, respectively.

Before predicting the context of the one or more data objects, the present step 103 needs to locate one or more data objects for extraction first. In one embodiment, step 103 involves using the layout information identified by sub-step 203 to locate the potential zone of the data objects. In one embodiment, it may further use the searchable contents as well as the corresponding location information to locate the data objects for extraction.

Predicting the context of the data objects may comprise training the extraction module to predict the context. To be more specific, predicting the context of the data objects can be achieved through one or more of the following sub-steps as illustrated in Figure 3.

(a) The first sub-step 301 is to provide, to a neural network of the processing server, labelled content or labelled documents – e.g. a training corpus in which a plurality of training contents, and context associated with each training content, which have been identified from one or more training documents. In the present context, a corpus refers to documents which can include for example, financial reports. Providing a training corpus also comprises providing a training corpus of documents from the domain, wherein the domain may be a financial domain. Providing a training corpus may further comprise providing non-searchable content having particular layout information – e.g. to enable training for layout detection.

Data augmentation methods can be used to create more "fake" training data to overcome the lack of training data problem. In one embodiment, assuming there is a set of fields and their corresponding values for each labelled document, data augmentation methods may include:

(i) randomly selecting two documents which have same file type and share the same fields; and

(ii) deleting values for the first document; and

(iii) filling the corresponding values of second document into the first document.

For example, assuming there exist two invoices, each invoice has its own layout information. A "fake" data can be created, which has the first invoice's fields (e.g. layout information) while part of its content (values such as money amount, company name etc.) is copied from the second invoice.

(b) The second sub-step 302 is to pre-train the neural network, by:

(i) masking a portion of the context associated with a plurality of said training contents;

(ii) making a prediction of each masked portion based on an unmasked portion of the respective context; and

(iii) fine-tuning the prediction of each masked portion using an additional prediction layer that receives an embedding for the respective masked content and a [CLS] token embedding for at least one of the respective training document and the location information of the content associated with the masked portion.

This invention enhances the pre-train models such as BERT and LayoutLM. Figure 4 illustrates an architecture of the pre-training model. Compared to the original BERT model which is the base pre-training model and the LayoutLM model which introduces absolute location embedding, this invention enhances the embedding method by adding relative position embedding. In this way, the pre-training model can not only understand the searchable contents, but also utilize the position information. In particular, a document page can be considered as a coordinate system with the top-left origin. In this setting, the bounding box can be precisely defined by (x0,

y0, x1, y1), where (x0, y0) corresponds to the position of the upper left in the bounding box, and (x1, y1) represents the position of the lower right. The LayoutLM model introduces (x0, y0) information into the embedding model, while this invention introduces relative coordinate (x0 / max_x0, x1 / max_x1) information into the embedding model to better describe the location for each token.
For hidden layers which connect to the embedding layer, This invention follows the BERT model which consists of a 12-layer Transformer with 768 hidden sizes, and 12 attention heads, which contains about 113M parameters.

In one embodiment, the system is preferred to be deployed on GPU servers due to the computational complexity. Since the system employs complex algorithms to transform scanned documents to readable texts and extract a list of data objects, the system's maximum response time ranges from 10 to 20 minutes for these functions subject to the number of documents in a case.

The [CLS] token embedding may be a [CLS] token embedding for the document. Determining the context comprising the respective masked portion may be based on an embedding for a document in which the masked content was identified, and an embedding for the content determined prior to masking, wherein the context is further determined based on an embedding for a location at which the masked portion appears in the document.

(c) The third sub-step 303 is to train the pre-trained neural network on a domain-specific corpus from a domain comprising the document.

In particular, the main pre-train task uses masked language model (MLM), which means given a sequence of tokens as model input, this task will

randomly select some tokens and replace them. For example, assuming the original token sequence is "[CLS] ben ##ef ##icia ##ry name is X company [SEP]". After masking, the token sequence will become "[CLS] ben [MASK] ##icia ##ry [MASK] is X company [SEP]".

The pre-train model will then try to learn the correct tokens for the masked position. The learning objective is the probability distribution of the whole vocabulary for each masked tokens and the loss function is category cross-entropy loss.

In one embodiment, the above pre-train step can be enhanced to whole word masking method. In particular, masking will take effect on a whole word, not a single token. Regarding the example above, it will become "[CLS] [MASK] [MASK] [MASK] [MASK] name is X company [SEP]" because "beneficiary" is a word.

In one embodiment, the pre-train step can be enhanced by adjusting the masked word distribution. In particular, the word frequency may be counted first. The selected probability will then be reduced for low frequency words. This is because in the trade finance domain, there are lots of unique words (e.g. company name) which may cause overfitting problems.

Pre-train Step	
Model	MLM Accuracy
Original BERT	overfitting
Original LayoutLM	overfitting
Original BERT + Whole Word Mask	0.706
Original LayoutLM + Whole Word Mask	0.735
6Estates Enhanced Model + Whole Word Mask	0.751

The above table shows the performance of the pre-train task. According to the experiment results, the enhanced method can achieve better performance than the original models.

(d) The fourth sub-step 304 is to apply the trained neural network to the digitized document, to predict the context of the one or more data objects. Figure 5 illustrates an example fine-tune model. In particular, the input information for the fine-tune model is formatted as "[CLS] target_field_name [SEP] text_of_one_potential_paragraph [SEP] context_of_surrounding_paragraphs". The "target_field_name" is the textrepresentation of the specific field (e.g. the applicant name of the bill of lading, the shipment date of the bill of lading etc.). The "text_of_one_potential_paragraph" is the text of the detection area, and the "context_of_surrounding_paragraphs" is the text of the surrounding area. The objective of fine-tune model consists of two parts: the first part is a binary output (i.e., possible/impossible) represents whether the "text_of_one_potential_paragraph" contains the target field value or not; the second part are two category outputs represent the start position and end position of the "text_of_one_potential_paragraph" for target field value.

Fine-tune Step	
Model	Field Extraction Accuracy
Original BERT	0.862
Original LayoutLM	0.877
6Estates Enhanced Model	0.881

The table above shows the performance of the fine-tune task. According to the experiment results, the enhanced method can achieve better performance than the original models.

(e) The fifth sub-step 305 is to apply the template based regular expression extraction method to the digitized document, to predict the

context of the one or more data objects. In particular, some target fields in a document type (e.g. bill of lading) may have mainstream pattern or writing style, so the present teachings map such patterns into templates and use regular expression matching modules to extract target fields. For example, the issue date of the bill of lading is usually located in a table which has one cell with pattern 'date\s*of\s*issue|issue.*date', then its adjacent cell with pattern '\d{2}[-/](jan|feb|mar|apr|may|june|aug|sept?|oct|nov|dec)[-/]\d{2,4}|\d{4}[/-]\d{2}[/-]\d{2}' is the target field value.

(f) The sixth sub-step 306 is to apply the combination of the results from (d) and (e) together to give the final predictions. In particular, for a target field, if step (d) (neural model) predicts a result and the confidence score from the model is higher than a predefined threshold, the final predictions will be the model output; or else the final predictions will be from step (e).

The present verifying step 104 is intended to generate a report including whether the one or more data objects meet the one or more predefined rules. In particular, users can determine whether the context of the one or more data objects predicted by the extracting steps 303 to 305 meets the one or more predefined rules. Users can also implement a calculating formula for several extracted objects to generate a new object. The rules may be trade rules. The present verification module of the processing server contains no training model inside. In one embodiment, the verifying step 104 first transforms all predefined rules as stipulated in a rule appendix (e.g. list of known rules) into computer acceptable rules.

The present verifying step 104 may further comprises producing a database. In one embodiment, the database contains knowledge about the compliance check process for the letter of credit domain. The database mainly includes three parts including: knowledge about the targeted fields of the letter of credit related documents (e.g. letter of credit number in MT700, invoice

number in Invoice, shipper in Bill of Lading etc.), knowledge about the transformation from rules to executable codes according to the Uniform Customs & Practice for Documentary Credits (UCP 600) and the International Standard Banking Practices (ISBP), and knowledge about the prior information which is helpful for the check process (e.g. all ports information around the world, all banks information around the world, standard goods information around the world).

It will be appreciated that many further modifications and permutations of various aspects of the described embodiments are possible. Accordingly, the described aspects are intended to embrace all such alterations, modifications, and variations that fall within the spirit and scope of the appended claims.

In the present context, a "corpus" refers to documents which can include for example, financial reports.

Throughout this specification and the claims which follow, unless the context requires otherwise, the word “comprise”, and variations such as “comprises” and “comprising”, will be understood to imply the inclusion of a stated integer or step or group of integers or steps but not the exclusion of any other integer or step or group of integers or steps.

The reference in this specification to any prior publication (or information derived from it), or to any matter which is known, is not, and should not be taken as an acknowledgment or admission or any form of suggestion that that prior publication (or information derived from it) or known matter forms part of the common general knowledge in the field of endeavour to which this specification relates.