Method of Analysing Unstructured Financial Documents

Technical Field

The present invention relates to a computer-implemented method of generating a report including correlation/reason between the changes of one or more numerical values in an unstructured document and one or more text values in the document.

Background

Financial documents such as balance sheets, income statements, cash flow statements and statements of shareholders' equity are critical in evaluating a company's financial activities and performance. The financial documents may be born digital or in scanned form. It is becoming increasingly important to check facts in financial documents, which is very challenging due to the diversity of layouts and the complexity of content structures.

Currently, many companies perform financial analysis through manual efforts. A typical manual analysis procedure is time-consuming, expensive and prone to error. Also, knowledge is not retained due to human attrition, resulting in high training costs.

It is desirable to develop an automated approach to reduce the processing time from reading original financial documents all the way through to generating an analysis report. Besides, the financials reported under different reporting standards need to be reconciled to a standard format. Advantageously, such efficiencies will allow banks to grow their credit assessment and underwriting business exponentially without facing scalability problems.

Several existing automated approaches focus on data digitization, which aims

to convert scanned documents into digital formats. However, to better understand and analyse financial documents, it is indispensable to automatically classify, extract and structuralize information from them. Document artificial intelligence (AI), or Document Intelligence, is a new research topic that refers to the techniques to automatically read, understand and analyse financial 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 are confronted with a lot oflimitations.

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 correlation/reason between the change of one or more numerical values in an unstructured document and one or more text values in the document, comprising:

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

(b) structuring, by a structuration module of the processing server, the document to produce a fully structured document;

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

(i) extract one or more numerical and text values from the structured document;

(ii) predict a context of the structured document based on content of the one or more numerical and text values extracted;

(iii) identify relationship between the one or more numerical values in the structured document and the context of the structured document; and

(iv) calculate the difference among at least two numerical values having the same relationship to the context of the structured document;

(d) detecting, by a detector of the processing server, correlation/reason between the one or more text values and the difference between the two numerical values; and

(e) generating, by a generator of the processing server, a report including correlation/reason between the one or more text values and the difference between the two numerical values.

Structuring the document may comprise identifying a plurality of contents in the document, each content being from a set of desired contents, identifying layout information of the document, categorizing a plurality of said contents into structured contents and unstructured contents, and converting the unstructured contents into structured format.

Extracting the one or more numerical and text values may comprise locating one or more numerical and text values according to the layout information and context structures.

Predicting the context of the structured document comprises 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;

(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;

(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 layout 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 structured document, to predict the context of the structured document;

(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) together to give the final predictions.

Providing a training corpus may further comprise providing unstructured content having particular layout information, and training the extraction module to predict the context of the one or more numerical and text values.

Detecting the correlation/reason between the one or more text values and the difference between the two numerical values comprises:

selecting one or more target text values for detection; and

predicting a numeric score on how well each text value is relevant to the difference between the two numerical values.

Selecting the one or more target text values for detection may comprise at least one of:

detecting whether there is relationship between the one or more text values and the two numerical values;

detecting whether the one or more text values containing at least one text value in a predefined text value list; and

training a neural network, the neural network identifying the most similar target text value based on the similarity score calculated by mapping the one or more text values to an embedding representation.

Predicting the numeric score may comprise training the detector to predict the numeric score associated with a text value in a structured document, by:

providing, to a neural network of the processing server, a training corpus in which a plurality of training text values, and numeric score associated with each training text value, have been identified;

pre-training the neural network by:

(i) masking a portion of the numeric scores associated with a plurality of said training text values;

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

(iii) fine-tuning the prediction with an additional prediction layer that receives an embedding for the respective masked numeric score and a [CLS] token embedding for a text value in which the masked portion was associated with;

training the pre-trained neural network on a domain-specific corpus from a domain comprising the structured document, the structured document containing at least one text value; and

applying the trained neural network to the structured document, to predict the numeric score associated with the at least one text value.

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

The domain may be a financial domain.

As disclosed herein is a system of generating a report including correlation/reason between the change of one or more numerical values in an unstructured document and one or more text values in the document, comprising:

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

(b) a structuration module of the processing server configured to structure the document to produce a fully structured document;

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

(i) extract one or more numerical and text values from the structured document;

(ii) predict a context of the structured document based on content of the one or more numerical and text values extracted;

(iii) identify a relationship between the one or more numerical values in the structured document and the context of the structured document; and

(iv) calculate the difference among at least two numerical values having the same relationship to the context of the structured document;

(d) a detector of the processing server configured to detect correlation between the one or more text values and the difference between the two numerical values; and

(e) a generator of the processing server configured to generate a report including correlation/reason between the one or more text values and the difference between the two numerical values.

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 is a flow diagram showing a computer-implemented method used for generating a report including correlation/reason between the change of one or more numerical values in an unstructured document and one or more text values in the document;
Figure 2 is a flow diagram showing a mechanism of structuring a document to produce a fully structured document;
Figure 3 is a flow diagram showing a mechanism of predicting the context of the data objects extracted from the document;
Figure 4 illustrates an example embedding layer for the plain text;
Figure 5 illustrates an example embedding layer for the financial tables;
Figure 6 illustrates an example fine-tune model;
Figure 7 illustrates an example prediction on how well each text value is relevant to the difference between the two numerical values.
Figure 8 is a flow diagram showing a mechanism of detecting correlation/reason between one or more text values and the difference between two numerical values.

Detailed description

A computer-implemented method 100 set out in Figure 1 is used for generating a report including correlation/reason between the change of one or more numerical values in an unstructured document and one or more text values in the document. The method, comprising:

(a) Step 101: providing, by an input module of a processing server, a document in an at least partially unstructured format;

(b) Step 102: structuring, by a structuration module of the processing server, the document to produce a fully structured document;

(c) Step 103: using an extraction module of the processing server to:

(i) extract one or more numerical and text values from the structured document;

(ii) predict a context of the structured document based on content of the one or more numerical and text values extracted;

(iii) identify a relationship between the one or more numerical values in the structured document and the context of the structured document; and

(iv) calculate the difference among at least two numerical values having the same relationship to the context of the structured document;

(d) Step 104: detecting, by a detector of the processing server, a correlation/reason between the one or more text values and the difference between the two numerical values; and

(e) Step 105: generating, by a generator of the processing server, a report including correlation/reason between the one or more text values and the difference between the two numerical values.

The present structuring step 102 is intended to convert the unstructured financial documents which are often in PDF format such as annual report, quarterly report, announcement to structured format. In the present context, an unstructured document contains information that either does not have a pre-defined data model or is not organized in a pre-defined manner, insofar as the unstructured document cannot be easily searchable in its current form

by a computer program. Structured data usually resides in relational databases (RDBMS). Fields store length-delineated data such as phone numbers, social security numbers, or ZIP codes. Even text strings of variable length like names are contained in records, making it a simple matter to search. Data may be human- or machine-generated as long as the data is created within an RDBMS structure. This format is eminently searchable both with human generated queries and via algorithms using type of data and field names, such as alphabetical or numeric, currency or date. A structured document refers to one that is structured 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 unstructured format will be converted by the digitizing step 102 to a structured format.

The present structuring step 102 consists of four sub-steps set out in Figure 2.

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

The second sub-step 202 is to identify layout information of the document. The layout information may contain the columns to be placed in the grid, 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 be obtained by extracting the character level rendering information from text readable PDF format documents. Thanks to the PDFBox, an open-source PDF reader engine, the contents in the document as well as the corresponding location information can be obtained. This invention will group characters into words, lines and paragraphs with layout information for further usage. In one embodiment, the structuration module may treat the PDF document provided by the input module as input, and identify the contents such as title, paragraph, table as well as the corresponding layout information of each content.

The third sub-step 203 is to categorize a plurality of said contents into structured contents and unstructured contents.

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

The extracting step 103 is intended to extract one or more data objects from the document structured 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 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 such as one or more numerical and text values according to the layout information, and then use text information to match the target field value. In one embodiment, step 103 involves using the layout information identified by sub-step 202 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 the following sub-steps as illustrated in Figure 3.

The first sub-step 301 is to provide, 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, 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 further comprises providing unstructured content having particular layout information, and training the extraction module to predict the context of the one or more numerical and text values.

In one embodiment, data augmentation method can be used to create more "fake" training data to overcome the lack of training data problem. The detailed data augmentation method is:

(i) there are a set of fields and their corresponding values for each labelled document; and

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

(iii) deleting values for the first document; and

(iv) filling into new values which are corresponding values of second document.

To give an example, assuming there exist two annual reports from different companies, each report has its own layout style. A "fake" data can be created, which has the first report's layout while part of the content (e.g. sales, cost of sales etc.) is from the second report.

The second sub-step 302 is to pre-train the neural network, comprising:

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

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

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-train model. Compared to the original BERT model which is the base pre-train model and LayoutLM model which introduce absolute location embedding, this invention enhances the embedding method by adding relative location embedding. In this way, the pre-training model can not only understand the searchable contents, but also utilize the location 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, and 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 original BERT model which consists of 12-layer Transformer with 768 hidden sizes, and 12 attention heads, which contains about 113M parameters.

The financial documents may contain tables which have different style with plain text. Figure 5 illustrates an example embedding layer for the financial tables. unlike the relative position embedding for plain text, the text in a table cell contains two attributes (in which row and in which column). Therefore, this invention adds row and column index embedding to the embedding layer for tables.

The system is preferred to be deployed on GPU servers due to the computational complexity. Since the system employs complex algorithms to convert the unstructured documents to structured format 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 and the size of documents in a case.

The [CLS] token embedding may be a [CLS] token embedding for the document. Determining the context comprises the respective masked portion 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.

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 major pre-train task is 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, it will become "[CLS] ben [MASK] ##icia ##ry [MASK] is X company [SEP]".

Then the pre-train model will 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 pre-train step can be further enhanced to whole word masking method, which means masking will take effects on a whole word, not a single token, so regarding to the example above, it will become "[CLS] [MASK] [MASK] [MASK] [MASK] name is X company [SEP]" because "beneficiary" is a word. This invention also enhances this pre-train step by adjusting the masked word distribution, which means we count the word frequency first, and then reduce the selected probability for low frequency words.

Pre-train Step	
Model	MLM Accuracy
Original BERT + Whole Word Mask	0.865
Original LayoutLM + Whole Word Mask	0.876
Enhanced Model + Whole Word Mask	0.883

Table 1

Table 1 shows the performance of the pre-train task. According to the experiment results, the enhanced method can achieve better performance than the original models.

The fourth sub-step 304 is to apply the trained neural network to the structured document, to predict the context of the structured document. Figure 6 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 text representation of the specific field (e.g. the sales in 2018, the cost of sales in 2018etc.). 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 (possible/impossible) represents whether the "text_of_one_potential_paragraph" contains the target field value of 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	0.871
Original LayoutLM	0.885
Enhanced Model	0.892

Table 2

Table 2 shows the performance of the fine-tune task. According to the experiment results, the enhanced method can achieve better performance than the original models.

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. annual report) may have mainstream pattern or writing style, so this invention concludes such patterns into templates and uses regular expression matching modules to extract target fields. For example, the net profit of the annual report is usually located in a table which has one cell with pattern 'net\sprofit', then its adjacent cell with pattern '[\d\,\.]+' is the target field value.

The sixth sub-step 306 is to apply the combination of the results from 305 and 306 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).

After predicting the context of the one or more data objects, the present step 103 needs to identify a relationship between the one or more numerical values in the structured document and the context of the structured document, and calculate the difference among at least two numerical values having the same relationship to the context of the structured document.

The detecting step 104 is intended to detect correlation/reason between one or more text values and the difference between two numerical values through the following sub-steps as illustrated in Figure 8.

The first sub-step 401 is to select one or more target text values for detection. In one embodiment, a spam filter model which has exactly the same model structure as fine-tune model can be used to detect whether a paragraph is spam data which may be non-readable or unrelated. The spam filter model can be used to detect whether there is relationship between the one or more text values. In one embodiment, a keyword filter model which applies white list mechanism can be used to filter potential paragraphs based on a predefined keyword list. In particular, the keyword filter model can detect whether the one or more text values containing at least one text value in a predefined text value list. In one embodiment, a neural network may be trained to identify the most similar target text value based on the similarity score calculated by mapping the one or more text values to an embedding representation.

The second sub-step 402 is to predict a numeric score on how well each text value is relevant to the difference between the two numerical values.

First of all, sub-step 402 may provide, to a neural network of the processing server, a training corpus in which a plurality of training text values, and numeric score associated with each training text value, 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. As mentioned before, providing a training corpus also comprises providing a training corpus of documents from the domain, wherein the domain may be a financial domain.

The model structure is exactly the same as sub-step 301-304, the only difference is the training data for this task. As shown in Figure 7, the input features (blue color) are the target financial field and a paragraph from the document, and the outputs (red color) are the binary relevance of the target field with the paragraph and the start/end positions of the selected text values, which are corresponding to fine-tune model structure.

The present detecting step 104 may further comprises a database. In one embodiment, the database contains knowledge about the financial status of an enterprise for the past few years, knowledge about the marketing status of an industry for the past few years.

The generating step 105 is intended to generate a report including correlation between the one or more text values and the difference between the two numerical values.

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.

Statements

1. A computer-implemented method of generating a report including correlation/reason between the change of one or more numerical values in an unstructured document and one or more text values in the document, comprising:

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

(b) structuring, by a structuration module of the processing server, the document to produce a fully structured document;

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

(i) extract one or more numerical and text values from the structured document;

(ii) predict a context of the structured document based on content of the one or more numerical and text values extracted;

(iii) identify a relationship between the one or more numerical values in the structured document and the context of the structured document; and

(iv) calculate the difference among at least two numerical values having the same relationship to the context of the structured document;

(d) detecting, by a detector of the processing server, a correlation/reason between the one or more text values and the difference between the two numerical values; and

(e) generating, by a generator of the processing server, a report including correlation/reason between the one or more text values and the difference between the two numerical values.

2. The method of 1, wherein structuring the document comprises:

identifying a plurality of contents in the document, each content being from a set of desired contents;

identifying layout information of the document;

categorizing a plurality of said contents into structured contents and unstructured contents; and

converting the unstructured contents into structured format.

3. The method of 1 and 2, wherein extracting the one or more numerical and text values comprises locating one or more numerical and text values according to the layout information and context structures.

4. The method of 1 to 3, wherein predicting the context of the structured document comprises 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;

(c2) pre-training the neural network by:

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

making a prediction of each masked portion based on an unmasked portion of the respective context;

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 layout 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 structured document, to predict the context of the structured document;

(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) together to give the final predictions.

5. The method of 4, wherein providing a training corpus further comprises providing unstructured content having particular layout information, and training the extraction module to predict the context of the one or more numerical and text values.

6. The method of 1 to 5, wherein detecting the correlation/reason between the one or more text values and the difference between the two numerical values comprises:

selecting one or more target text values for detection; and

predicting a numeric score on how well each text value is relevant to the difference between the two numerical values.

7. The method of 6, wherein selecting the one or more target text values for detection comprises at least one of:

detecting whether there is relationship between the one or more text values and the two numerical values;

detecting whether the one or more text values containing at least one text value in a predefined text value list; and

training a neural network, the neural network identifying the most similar target text value based on the similarity score calculated by mapping the one or more text values to an embedding representation.

8. The method of 6, wherein predicting the numeric score comprises training the detector to predict the numeric score associated with a text value in a structured document, by:

providing, to a neural network of the processing server, a training corpus in which a plurality of training text values, and numeric score associated with each training text value, have been identified;

pre-training the neural network by:

masking a portion of the numeric scores associated with a plurality of said training text values;

making a prediction of each masked portion based on an unmasked portion of the respective numeric score;

fine-tuning the prediction with an additional prediction layer that receives an embedding for the respective masked numeric score and a [CLS] token embedding for a text value in which the masked portion was associated with;

training the pre-trained neural network on a domain-specific corpus from a domain comprising the structured document, the structured document containing at least one text value; and

applying the trained neural network to the structured document, to predict the numeric score associated with the at least one text value.

9. The method of 4 to 8, wherein the [CLS] token embedding is a [CLS] token embedding for the document.

10. The method of any one of 4 to 9, wherein the domain is a financial domain.

11. A system of generating a report including correlation/reason between the change of one or more numerical values in an unstructured document and one or more text values in the document, comprising:

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

(b) a structuration module of the processing server configure to structure the document to produce a fully structured document;

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

(i) extract one or more numerical and text values from the structured document;

(ii) predict a context of the structured document based on content of the one or more numerical and text values extracted;

(iii) identify a relationship between the one or more numerical values in the structured document and the context of the structured document; and

(iv) calculate the difference among at least two numerical values having the same relationship to the context of the structured document;

(d) a detector of the processing server configured to detect correlation/reason between the one or more text values and the difference between the two numerical values; and

(e) a generator of the processing server configured to generate a report including correlation/reason between the one or more text values and the difference between the two numerical values.

12. The system of 11, wherein structuring the document comprises:

identifying a plurality of contents in the document, each content being from a set of desired contents;

identifying layout information of the document;

categorizing a plurality of said contents into structured contents and unstructured contents; and

converting the unstructured contents into structured format.

13. The system of 11 and 12, wherein extracting the one or more numerical and text values comprises locating one or more numerical and text values according to the layout information and context structures.

14. The system of 11 to 13, wherein predicting the context of the structured document comprises 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;

(c2) pre-training the neural network by:

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

making a prediction of each masked portion based on an unmasked portion of the respective context;

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 layout 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; and

(c4) applying the trained neural network to the structured document, to predict the context of the structured document;

(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) together to give the final predictions.

15. The system of 14, wherein providing a training corpus further comprises providing unstructured content having particular layout information, and training the extraction module to predict the context of the one or more numerical and text values.

16. The system of 11 to 15, wherein detecting the correlation/reason between the one or more text values and the difference between the two numerical values comprises:

selecting one or more target text values for detection; and

predicting a numeric score on how well each text value is relevant to the difference between the two numerical values.

17. The system of 16, wherein selecting the one or more target text values for detection comprises at least one of:

detecting whether there is relationship between the one or more text values and the two numerical values;

detecting whether the one or more text values containing at least one text value in a predefined text value list; and

training a neural network, the neural network identifying the most similar target text value based on the similarity score calculated by mapping the one or more text values to an embedding representation.

18. The system of 16, wherein predicting the numeric score comprises training the detector to predict the numeric score associated with a text value in a structured document, by:

providing, to a neural network of the processing server, a training corpus in which a plurality of training text values, and numeric score associated with each training text value, have been identified;

pre-training the neural network by:

masking a portion of the numeric scores associated with a plurality of said training text values;

making a prediction of each masked portion based on an unmasked portion of the respective numeric score;

fine-tuning the prediction with an additional prediction layer that receives an embedding for the respective masked numeric score and a [CLS] token embedding for a text value in which the masked portion was associated with;

training the pre-trained neural network on a domain-specific corpus from a domain comprising the structured document, the structured document containing at least one text value; and

applying the trained neural network to the structured document, to predict the numeric score associated with the at least one text value.

19. The system of 14 to 18, wherein the [CLS] token embedding is a [CLS] token embedding for the document.

20. The system of any one of 14 to 19, wherein the domain is a financial domain.

Method of Analysing Unstructured Financial Documents

ABSTRACT

Disclosed herein is a computer-implemented method of generating a report including correlation/reason between the change of one or more numerical values in an unstructured document and one or more text values in the document, comprising providing, by an input module of a processing server, a document in an at least partially unstructured format; structuring, by a structuration module of the processing server, the document to produce a fully structured document; using an extraction module of the processing server to extract one or more numerical and text values from the structured document, predict a context of the structured document based on content of the one or more numerical and text values extracted, identify a relationship between the one or more numerical values in the structured document and the context of the structured document, and calculate the difference among at least two numerical values having the same relationship to the context of the structured document; detecting, by a detector of the processing server, a correlation/reason between the one or more text values and the difference between the two numerical values; and generating, by a generator of the processing server, a report including correlation/reason between the one or more text values and the difference between the two numerical values.

[Figure 1]

Claims

1. A computer-implemented method of generating a report including correlation/reason between the change of one or more numerical values in an unstructured document and one or more text values in the document, comprising:

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

(b) structuring, by a structuration module of the processing server, the document to produce a fully structured document;

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

(i) extract one or more numerical and text values from the structured document;

(ii) predict a context of the structured document based on content of the one or more numerical and text values extracted;

(iii) identify a relationship between the one or more numerical values in the structured document and the context of the structured document; and

(iv) calculate the difference among at least two numerical values having the same relationship to the context of the structured document;

(d) detecting, by a detector of the processing server, a correlation/reason between the one or more text values and the difference between the two numerical values; and

(e) generating, by a generator of the processing server, a report including correlation/reason between the one or more text values and the difference between the two numerical values.

2. The method of claim 1, wherein structuring the document comprises:

identifying a plurality of contents in the document, each content being from a set of desired contents;

identifying layout information of the document;
categorizing a plurality of said contents into structured contents and unstructured contents; and
converting the unstructured contents into structured format.

3. The method of claim 1 or 2, wherein extracting the one or more numerical and text values comprises locating one or more numerical and text values according to the layout information and context structures.

4. The method of any one of claims 1 to 3, wherein predicting the context of the structured document comprises 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;

(c2) pre-training the neural network by:

masking a portion of the context associated with a plurality of said training contents;
making a prediction of each masked portion based on an unmasked portion of the respective context;
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 layout 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 structured document, to predict the context of the structured document;

(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) together to give the final predictions.

5. The method of claim 4, wherein providing a training corpus further comprises providing unstructured content having particular layout information, and training the extraction module to predict the context of the one or more numerical and text values.

6. The method of any one of claims 1 to 5, wherein detecting the correlation/reason between the one or more text values and the difference between the two numerical values comprises:

selecting one or more target text values for detection; and

predicting a numeric score on how well each text value is relevant to the difference between the two numerical values.

7. The method of claim 6, wherein selecting the one or more target text values for detection comprises at least one of:

detecting whether there is relationship between the one or more text values and the two numerical values;

detecting whether the one or more text values containing at least one text value in a predefined text value list; and

training a neural network, the neural network identifying the most similar target text value based on the similarity score calculated by mapping the one or more text values to an embedding representation.

8. The method of claim 6, wherein predicting the numeric score comprises training the detector to predict the numeric score associated with a text value in a structured document, by:

providing, to a neural network of the processing server, a training corpus in which a plurality of training text values, and numeric score associated with each training text value, have been identified;

pre-training the neural network by:

masking a portion of the numeric scores associated with a plurality of said training text values;

making a prediction of each masked portion based on an unmasked portion of the respective numeric score;

fine-tuning the prediction with an additional prediction layer that receives an embedding for the respective masked numeric score and a [CLS] token embedding for a text value in which the masked portion was associated with;

training the pre-trained neural network on a domain-specific corpus from a domain comprising the structured document, the structured document containing at least one text value; and

applying the trained neural network to the structured document, to predict the numeric score associated with the at least one text value.

9. The method of any one of claims 4 to 8, wherein the [CLS] token embedding is a [CLS] token embedding for the document.

10. The method of any one of claims 4 to 9, wherein the domain is a financial domain.

11. A system of generating a report including correlation/reason between the change of one or more numerical values in an unstructured document and one or more text values in the document, comprising:

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

(b) a structuration module of the processing server configure to structure the document to produce a fully structured document;

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

(i) extract one or more numerical and text values from the structured document;

(ii) predict a context of the structured document based on content of the one or more numerical and text values extracted;

(iii) identify a relationship between the one or more numerical values in the structured document and the context of the structured document; and

(iv) calculate the difference among at least two numerical values having the same relationship to the context of the structured document;

(d) a detector of the processing server configured to detect correlation/reason between the one or more text values and the difference between the two numerical values; and

(e) a generator of the processing server configured to generate a report including correlation/reason between the one or more text values and the difference between the two numerical values.

12. The system of claim 11, wherein structuring the document comprises:

identifying a plurality of contents in the document, each content being from a set of desired contents;

identifying layout information of the document;

categorizing a plurality of said contents into structured contents and unstructured contents; and

converting the unstructured contents into structured format.

13. The system of claim 11 or 12, wherein extracting the one or more numerical and text values comprises locating one or more numerical and text values according to the layout information and context structures.

14. The system of any one of claims 11 to 13, wherein predicting the context of the structured document comprises 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;

(c2) pre-training the neural network by:

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

making a prediction of each masked portion based on an unmasked portion of the respective context;

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 layout 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; and

(c4) applying the trained neural network to the structured document, to predict the context of the structured document;

(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) together to give the final predictions.

15. The system of claim 14, wherein providing a training corpus further comprises providing unstructured content having particular layout information, and training the extraction module to predict the context of the one or more numerical and text values.

16. The system of any one of claims 11 to 15, wherein detecting the correlation/reason between the one or more text values and the difference between the two numerical values comprises:

selecting one or more target text values for detection; and

predicting a numeric score on how well each text value is relevant to the difference between the two numerical values.

17. The system of claim 16, wherein selecting the one or more target text values for detection comprises at least one of:

detecting whether there is relationship between the one or more text values and the two numerical values;

detecting whether the one or more text values containing at least one text value in a predefined text value list; and

training a neural network, the neural network identifying the most similar target text value based on the similarity score calculated by mapping the one or more text values to an embedding representation.

18. The system of claim 16, wherein predicting the numeric score comprises training the detector to predict the numeric score associated with a text value in a structured document, by:

providing, to a neural network of the processing server, a training corpus in which a plurality of training text values, and numeric score associated with each training text value, have been identified;

pre-training the neural network by:

masking a portion of the numeric scores associated with a plurality of said training text values;

making a prediction of each masked portion based on an unmasked portion of the respective numeric score;

fine-tuning the prediction with an additional prediction layer that receives an embedding for the respective masked numeric score and a [CLS] token embedding for a text value in which the masked portion was associated with;

training the pre-trained neural network on a domain-specific corpus from a domain comprising the structured document, the structured document containing at least one text value; and

applying the trained neural network to the structured document, to predict the numeric score associated with the at least one text value.

19. The system of any one of claims 14 to 18, wherein the [CLS] token embedding is a [CLS] token embedding for the document.

20. The system of any one of claims 14 to 19, wherein the domain is a financial domain.