

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 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/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.