Claims

1. A computer-implemented method for 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,

wherein digitizing the document comprises:

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,

wherein using the extraction module comprises:

locating one or more data objects according to the layout information; and

predicting the context of the one or more data objects, and

wherein predicting the context 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 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) together to give the final predictions.

2. The method of claim 1, wherein providing a training corpus further comprises providing non-searchable content having particular layout

information, and training the extraction module to predict the context of the one or more data objects.

3. The method of claim 1 or 2, wherein the [CLS] token embedding is a [CLS] token embedding for the document.

4. The method of any one of claims 1 to 3, wherein providing a training corpus comprises providing a training corpus of documents from the domain.

5. The method of any one of claims 1 to 4, wherein the domain is a financial domain.

6. The method of any one of claims 1 to 5, comprising determining the context comprising 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.

7. The method of claim 6, wherein the context is further determined based on an embedding for a location at which the masked portion appears in the document.

8. A system of 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;

(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,

wherein digitizing the document comprises:

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,

wherein using the extraction module comprises:

locating one or more data objects according to the layout information; and

predicting the context of the one or more data objects, and

wherein predicting the context 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 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) together to give the final predictions.

9. The system of claim 8, wherein providing a training corpus further comprises providing non-searchable content having particular layout information, and training the extraction module to predict the context of the one or more data objects.

10. The system of claim 8 or 9, wherein the [CLS] token embedding is a [CLS] token embedding for the document.

11. The system of any one of claims 8 to 10, wherein providing a training corpus comprises providing a training corpus of documents from the domain.

12. The system of any one of claims 8 to 11, wherein the domain is a financial domain.

13. The system of any one of claims 8 to 12, comprising determining the context comprising 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.

14. The system of claim 13, wherein the context is further determined based on an embedding for a location at which the masked portion appears in the document.