Claims

1. A computer system for extracting structured data from unstructured or semi-structured text in an electronic document, the system comprising:

a graphical user interface configured to present to a user a graphical view of a document for use in training multiple data extraction models for the document, each data extraction model associated with a user defined question;

a user input component configured to enable the user to highlight portions of the document;

the system configured to present in association with each highlighted portion an interactive user entry object which presents a menu of question types to a user in a manner to enable the user to select one of the question types, and a field for receiving from the user a question identifier in the form of human readable text, wherein the question identifier and question type selected by the user are used for selecting a data extraction model, and wherein the highlighted portion of the document associated with the question identifier is used to train the selected data extraction model.

2. The method of claim 1, wherein the question types comprise two or more of: a point extraction question type, a section extraction question type and a table extraction question type.

3. The method of claim 1 or 2, wherein the text is tokenized and features are extracted from the tokenized text in dependence on the selected question type.

4. The method of claim 2 and 3, wherein a common model type is used for said two or more question types but with different features and tokenization.

5. The method of claim 3 or 4, wherein the system is configured to assign, for each user defined question, a label to each token of the tokenized text based on any highlighted portion or portions of the text associated with that question, wherein that label is used to train the data extraction model associated with that question.

6. The method of any preceding claim, wherein the user interface is configured to present graphical views of multiple documents in sequence, to enable the user to highlight portions of each document for use in training the multiple data extraction models.

7. The method of claim 6, wherein the graphical user interface presents a tabulated format comprising multiple rows and columns, each row having a document identifier of one of the documents and each column having associated therewith one of the questions identifiers, whereby each cell in each row corresponds to one document and one question, and is populated with any highlighted portion or portions of that document associated with that question.

8. The method of any preceding claim, comprising using, by a prediction component, the trained data extraction models to extract from unlabelled documents answers to the user defined questions, wherein each answer extracted by each data extraction model comprises or denotes any portion of portions of text determined to be relevant to the question associated with that data extraction model.

9. The method of claim 8 when dependent on claim 7, wherein an updated tabulated format is presented after training, with additional rows corresponding to the documents from which answers have been extracted by the data extraction models, in which the extracted answers are arranged according to question identifier.

10. The method of any preceding claim, wherein the system is configured to enable the user to highlight portions of the text for different questions of different question types in the same document.

11. The method of claims 7 and 10, whereby the row of the tabular format corresponding to that document contains highlighted portions associated with the different questions of the different question types.

12. The method of claim 7 or any claim dependent thereon, wherein in response to a view change instruction received in respect of one of the document identifiers, the graphical user interface is configured to switch to a view of that document to enable the user to highlight further portion(s) of the document.

13. The method of claim 7 or any claim dependent thereon, wherein the highlighted portion of each document in each cell is selectable to switch to a view of that document in which that portion is visible and visibly highlighted on the graphical user interface.

14. The method of any preceding claim, wherein the graphical user interface is configured to output a suggested question type for each highlighted portion based on the highlighted portion.

15. The method of claim 14, wherein the question type is suggested based a length of the highlighted portion.

16. A computer system comprising one or more computers programmed to carry out the method of any preceding method claim.

17. One or more computer programs, stored on transitory or non-transitory media, for programming a computer system to implement the steps or functionality of any of claims 1 to 15.