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 ~~questions~~ 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, ~~configured to use~~ 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 for extracting structured data from unstructured or semi-structured text in an electronic document, the system comprising:~~

~~a plurality of data extraction models, each model configured to receive training data relating to a user defined question, wherein the user defined question conforms to one of multiple question types;~~

~~a training data generator configured to:~~

~~tokenize each electronic document of a document cohort as a sequence of tokens;~~

~~receive, for each of multiple electronic documents of the document cohort, at least one highlighted portion of text for at least one user defined question, wherein the highlighted portion of text has been graphically highlighted by a user as being relevant to that user defined question in a document view of a graphical user interface, and~~

~~provide for each sequenced token of each document a feature value for each of a set of features characterising the token, and a label for each question which indicates if the token is relevant or not relevant to that question based on the at least one~~

graphically highlighted portion of text, the labels being sequenced according to the tokens,

wherein each model is trained using the feature values and the sequenced labels for a specific user defined question as the training data.

17. The method of claim 16, wherein for at least one document, at least two portions are graphically highlighted in respect of at least two of the user defined questions.

18. The method of claim 17, wherein the at least two questions are of different question types.

19. The method of any of claims 16 to 18, wherein each token is a word token.

20. The method of any of claims 16 to 18, wherein each token is a section token.

21. The method of any of claims 16 to 18, wherein at least one of the features is dependent on a candidate label assigned to the token and a candidate label assigned to at least one other token in the sequence.

22. The method of claim 21, wherein at least another of the features is dependent on an observation computed for the token and the candidate label of the token but independent of any label of any other token.

23. The method or system of any preceding claim, wherein the or each data extraction model is a conditional random field.

24. The method of any preceding claim, the questions types comprise multiple section question types for different section lengths.

~~25. The method of any preceding claim, wherein multiple models are trained for at least one of the user defined questions, and evaluated in order to select one of the multiple models for that question.~~

~~26. The method of claim 25, wherein the multiple models have different tokenization, different features, and/or different labelling schemes.~~

~~27. A method of training a data extraction model to identify an extracted point from unstructured or semi-structured text of an electronic document, the method comprising:~~

~~providing a training data structure for training a data extraction model associated with a single user defined question by converting each of a cohort of unstructured or semi-structured documents into word tokens representing words of the document,~~

~~generating, for each document, for each word token of the document, feature values of a set of features derived from the word tokens, the features including at least one feature associated with the relative position of the word token in a sequence of tokens representing each document,~~

~~providing, for each document, a label for each token word, the label being a label which indicates whether the word token is relevant to the user defined question, or whether the word token is not relevant to the user defined question, the labels being sequenced according to the sequence of tokens,~~

~~training the document extraction field model over the cohort of documents using the feature values and the sequenced labels.~~

~~28. The method of claim 27, wherein the at least one feature is dependent on the position of the word token relative to the start of the document.~~

~~29. The method of claim 28, wherein the at least one feature is expressed in proportion to a length of the document.~~

30. The method of any of claims 27 to 29, wherein each of at least some of the features is a binary feature indicating whether the token matches a defined word.

31. The method of claim 30, wherein each of said binary features is independent of the capitalization of the word and one or more additional binary features are provided to indicate a capitalization style.

32. The method of any of claims 27 to 31, wherein each of at least some of the features is a binary feature indicating whether a shape of the word matches a defined word shape.

33. The method of any of claims 30 to 32, wherein each defined word and/or word shape is derived from the cohort of documents.

34. The method of any of claims 27 to 33, wherein at least one of the features is dependent on one or more feature values of one or more other word tokens within a window of the word token.

35. A method of training a data extraction model for identifying whether chunks of structured or unstructured text are relevant to a single user defined question, the method comprising:

dividing a document into predetermined chunks of text, and

for each chunk:

generating feature values for each feature in a feature set of the data extraction model, wherein the feature values are determined by tokenising unstructured text of the chunk and assessing each token based on its properties and its relationship to other tokens within the chunk,

applying a structured data training set to a conditional random field model, wherein the training data set comprises feature values of a set of features for each chunk of the document, with a label as to relevance or non-relevance of that chunk, the chunks in a predetermined sequence in the document.

~~36. A computer-implemented method of extracting structured data from unstructured or semi-structured text in an electronic document, the method comprising:~~

~~tokenizing the text as a token sequence of section tokens, wherein each section token corresponds to a portion of the text;~~

~~for each of multiple candidate label sequences, wherein each label sequence of the multiple candidate label sequences assigns a label to each section token: extracting a set of feature values for each section token, wherein at least one feature value of the set of feature values depends on (i) the label assigned to that section token by that label sequence and (ii) the label assigned to at least one other section token in the token sequence by that label sequence;~~

~~applying at least one data extraction model to each label sequence of the multiple candidate label sequences and the set of feature values determined for that label sequence for each section token, thereby computing a score for each label sequence;~~

~~selecting a label sequence for the token sequence based on the computed scores; and~~

~~providing the extracted structured data, in the form of one or more extracted section tokens of the token sequence, based on the selected label sequence.~~

~~37. The method of claim 36, wherein at least some features are extracted for each section token by determining a distribution across topics for that section token, wherein each of said at least some features is, or is derived from, a probabilistic topic prominence value for a single one of the topics and that section token.~~

~~38. The method of claim 37, wherein each of the topics is determined for each section token as a distribution over a topic word set.~~

~~39. The method of any of claims 36 to 38, wherein at least one feature value is extracted by computing a count of occurrences of any words of a topic word set within~~

~~the portion of text corresponding to the section token, and extracting the feature value as a topic prominence value for the topic word set.~~

~~40. The method of claim 39, wherein the topic prominence value expresses the count in proportion to a total word count of that portion of text.~~

~~41. The method of any of claims 36 to 40, wherein multiple topics having different topic word sets are defined, and one or more features are extracted for each topic.~~

~~42. The method of claim 36, wherein, in determining at least one feature value, a topic distribution may be determined for the portion of text corresponding to the section token.~~

~~43. The method of claim 42, wherein the topic distribution of each section token is compared to an average topic distribution to extract a feature value for that section token as a divergence between those distributions.~~

~~44. The method of claim 43, wherein the average topic distribution is determined from sections of training data labelled as relevant, excluding non-relevant sections, the training data used to train the data extraction model.~~

~~45. The method of any of claims 36 to 44, wherein at least some of the features are learned from extraction training documents used to train the data extraction model.~~

~~46. The method of claim 45, wherein each document of the extraction training documents may be tokenized and each section token thereof may be labelled as relevant or non-relevant to a question which the document extraction model is trained to answer.~~

~~47. The method of claim 45 or 46, wherein features are only extracted from the section tokens labelled as relevant.~~

~~48. The method of any of claims 36 to 47, wherein at least some of the features are learned from documents which have not been used to train the data extraction model(s).~~

~~49. The method of any of claims 36 to 48, wherein at least one of the features is extracted by, for each of multiple known terms, computing a term frequency-inverse document frequency (tf-idf) value based on a count of occurrences of that term in the portion of text and an idf weight.~~

~~50. The method of claim 49, wherein the idf weight has been learned from a set of tokenized training documents, being the inverse of the number of section tokens of those documents in which the term appears.~~

~~51. The method of any of claims 36 to 50, wherein a feature value is computed for each section token based on at least one feature value extracted for at least one other token in the token sequence.~~

~~52. The method of any of claims 36 to 51, wherein the data extraction model has been trained using ground truth labels assigned to tokenized training documents, and the ground truth labels have also been used to select a subset of section tokens of the tokenized training document, and learn at least one property of the subset of section tokens, wherein a feature value is extracted for each section token based on the learned property.~~

~~53. A computer-implemented method of extracting structured data from unstructured or semi-structured text in an electronic document, the method comprising:~~

~~tokenizing the text as a token sequence;~~

~~extracting features for each token of the token sequence;~~

~~applying a data extraction model to the extracted features to determine a highest-probability label sequence for the token sequence, wherein the label sequence assigns a label to each token, which indicates if that token is relevant to a question;~~

wherein if the highest-probability label sequence indicates that at least one token is relevant to the question, then that token(s) is provided as an answer to the question, the extracted structured data comprising the answer in that event;

wherein if highest-probability label sequence indicates that no token is relevant to the question, then an answer forcing process is applied by:

determining, for each token of the token sequence, a marginal probability of that token being relevant, wherein the at least one of the marginal probabilities is used to determine a forced answer.

54. The method of claim 53, wherein, if the highest marginal probability meets a probability threshold, the forced answer comprises at least the token having the highest marginal probability, and if not, the forced answer is a null result.

55. The method of claim 53 or 5545, wherein dynamic programming is used to determine the label sequence having the highest computed probability, without computing the probability of every possible label sequence, and to compute the marginal probability for each token, without requiring the probability of every possible label sequence to be computed.

56. The method of any of claims 53 to 55, wherein a first dynamic programming algorithm may be used to determine the highest probability label sequence, and a second dynamic programming algorithm may be used to determine the marginal probability of each token.

57. The method of claim 56, wherein the first dynamical programming algorithm is. the Viterbi algorithm, and the second dynamic programming algorithm is the forward-backward algorithm.

58. The method of any of claims 53 to 55, wherein the data extraction model is applied to the extracted features to compute a probability of each label sequence of multiple candidate label sequences, wherein each label sequence assigns a label to each token, which indicates if that token is relevant to a question;

~~wherein, for each token of the token sequence, a marginal probability of that token being relevant is computed as a sum of the probabilities computed for the subset of all candidate label sequences for which that token is relevant, wherein the at least one of the marginal probabilities is used to determine a forced answer.~~

~~59~~16. A computer system comprising one or more computers programmed to carry out the method of any preceding method claim.

~~60~~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 ~~58~~15.

~~61. One or more data extraction models trained in accordance with any of claims 1 to 35 for execution on one or more computers to extract data from unstructured or semi-structed text documents.~~