Claims

1. A computer-implemented method (100) for performing data capture in an unlabelled document based on a labelled document, wherein the labelled document contains one or more elements each labelled with respect to a field, and wherein each of the labelled document and the unlabelled document is an image, or contain image data, or is a PDF document, the method comprising:

determining (S110) one or more anchors, wherein an anchor comprises a pair of location points of words or visual features in the labelled document and the unlabelled document that match, wherein one location point of the pair is in the labelled document and the other location point of the pair is in the unlabelled document, and wherein a match is indicative that a similarity of the words or visual features is over a predetermined threshold;

generating (S120) an overall kernel using the determined one or more anchors, wherein generating the overall kernel comprises:

generating a convolutional kernel for each of the one or more anchors; and

generating the overall kernel by summing the one or more convolution kernels;

determining, using a kernel machine, a field for each of one or more elements in the unlabelled document, wherein the kernel machine is trained using at least a subset of matrix entries for the overall kernel; and

identifying, based on the determined fields, one or more elements to be captured in the unlabelled document, wherein the one or more elements to be captured correspond in field to the one or more labelled elements contained in the labelled document.

2. The method according to claim 1, wherein each convolution kernel measures the similarity of words within and between the labelled document and the unlabelled document, wherein the similarity is a function of a distance of the location point of the respective word from the respective location point of the anchor and a similarity in terms of the location of each respective word relative to the respective location point of the anchor in the respective document.

3. The method according to any one of the preceding claims, wherein the labelled document further comprises the one or more fields indicated by the one or more labels.

4. The method according to any one of the preceding claims, wherein determining a field for each of one or more elements in the unlabelled document comprises grouping the plurality of elements in the unlabelled document into one or more groups, each group corresponding to a field.

5. The method according to claim 4, wherein grouping the plurality of elements in the unlabelled document comprises:

generating a linear text sequence of the plurality of elements in the unlabelled document by ordering the words corresponding to the plurality of elements into a linear text sequence;

determining the argmax classification of each element in the generated linear text sequence; and

grouping every maximal contiguous subsequence of elements in the generated linear text sequence that share an argmax classification into a respective group.

6. The method according to any one of claims 3 to 5, wherein the one or more fields and one or more elements in the labelled document are presented in an information structure, and the method further comprises outputting the identified one or more elements and their corresponding fields in the same information structure.

7. The method according to any one of claims 3 to 6, further comprising, prior to determining (S110) one or more anchors, acquiring a schema that defines the one or more fields contained in the labelled document and whether elements corresponding to each field in the labelled document are presented in a table.

8. The method according to any one of the preceding claims, wherein determining (S110) an anchor comprises:

extracting all the words in the labelled document and the unlabelled document;

generating a first linear text sequence by ordering words contained in the labelled document into a linear text sequence and removing any word that corresponds to a labelled element;

generating a second linear text sequence by ordering words contained in the unlabelled document into a linear text sequence;

identifying a match between the first and second linear text sequences; and

designating the location points of the matched words as the anchor.

9. The method according to any one of the preceding claims, wherein determining (S110) an anchor comprises:

identifying a visual feature that is present in both the labelled document and the unlabelled document; and

designating the location points of the identified visual feature as the anchor.

10. The method according to any one of the preceding claims, wherein determining (S110) an anchor comprises:

placing a plurality of points around the boundary of the unlabelled document and a same number of points around the boundary of the labelled document;

identifying a match between a point in the labelled document and a point in the unlabelled document; and

designating the locations of the matched points as the anchor.

11. The method according to any one of the preceding claims, further comprising, prior to determining (S110) one or more anchors:

acquiring a plurality of candidate labelled documents; and

determining the labelled document from the acquired candidate labelled documents based on a similarity metric between each of the plurality of candidate labelled documents and the unlabelled document.

12. A system (200) for performing data capture in an unlabelled document based on a labelled document, wherein the labelled document contains one or more elements each labelled with respect to a field, and wherein each of the labelled document and the unlabelled document is an image, or contain image data, or is a PDF document, the system comprising:

a determining unit (210) configured to determine one or more anchors, wherein an anchor comprises a pair of location points of words or visual features in the labelled document and the unlabelled document that match, wherein one location point of the pair is in the labelled document and the other location point of the pair is in the unlabelled document, and wherein a match is indicative that a similarity of words or visual features is over a predetermined threshold;

a generating unit (220) configured to generate an overall kernel using the determined one or more anchors by generating a convolution kernel for each of the one or more anchors and generating the overall kernel by summing the one or more convolution kernels; and

an identifying unit (230) configured to determine, using a kernel machine, a field for each of one or more elements in the unlabelled document, wherein the kernel machine is trained using at least a subset of matrix entries for the overall kernel, and to identify, based on the determined fields, one or more elements to be captured in the unlabelled document, wherein the one or more elements to be captured correspond in field to the one or more labelled elements contained in the labelled document.

13. A computer-readable storage medium comprising instructions which, when executed by a computer, cause the computer the carry out the steps of the method of any one of claims 1 to 11.