CLAIMS

1. An Optical Character Recognition (OCR) system, comprising:

an acquisition device configured to obtain a digital image of a physical document;

an image conversion device configured to convert the digital image of the physical document into corresponding machine-readable text;

a correction device configured to:

evaluate the machine-readable text using a trained Long short-term memory (LSTM) neural network language model to determine whether correction to the machine-readable text is required;

if correction to the machine-readable text is required, determine a most similar text relative to the machine-readable text from a name and address corpus using a modified edit distance technique; and

correct the machine-readable text with the determined most similar text; and

an output device configured to: (i) output the corrected machine-readable text and (ii) add the corrected machine-readable text and non-corrected machine-readable text to the name and address corpus.

2. The OCR system as claimed in claim 1, wherein the system further comprises:

a database device having stored therein a collection of names and addresses associated with a target domain, wherein the database device is configured to construct the name and address corpus based on the stored collection of names and addresses associated with the target domain; and

a training device configured to train the LSTM neural network language model using the name and address corpus.

3. The OCR system as claimed in claim 2, wherein the database device is further configured to re-construct the name and address corpus based on the collection of names and addresses that comprises the corrected and the non-corrected machine-readable text.

4. The OCR system as claimed in claim 3, wherein the training device is further configured to re-train the LSTM neural network language model using the re-constructed name and address corpus.

5. The OCR system as claimed in any one of the preceding claims, wherein the correction device is further configured to evaluate a log(frequency of edit pairs in history) to determine the most similar text relative to the machine-readable text to be corrected.

6. A system for automatic Optical Character Recognition (OCR) correction, the system comprising:

a processor device; and

a memory device including computer program code;

the memory device and the computer program code configured to, with the processor device, cause the system at least to:

evaluate an OCR result using a trained Long short-term memory (LSTM) neural network language model to determine whether correction to the OCR result is required;

if correction to the OCR result is required, determine a most similar text relative to the OCR result from a name and address corpus using a modified edit distance technique;

correct the OCR result with the determined most similar text; and

add a corrected and non-corrected OCR result to the name and address corpus.

7. The system as claimed in claim 6, wherein the system is further caused to:

construct the name and address corpus based on a collection of names and addresses associated with a target domain; and

train the LSTM neural network language model using the name and address corpus.

8. The system as claimed in claim 7, wherein the system is further caused to:

re-construct the name and address corpus based on the collection of names and addresses that comprises the corrected and the non-corrected OCR result; and

re-train the LSTM neural network language model using the re-constructed name and address corpus.

9. The system as claimed in any one of claims 6 to 8, wherein the system is further caused to evaluate a log(frequency of edit pairs in history) to determine the most similar text relative to the OCR result.

10. A computer-implemented method for automatic Optical Character Recognition (OCR) correction, comprising:

evaluating an OCR result using a trained Long short-term memory (LSTM) neural network language model to determine whether correction to the OCR result is required;

if correction to the OCR result is required, determining a most similar text relative to the OCR result from a name and address corpus using a modified edit distance technique;

correcting the OCR result with the determined most similar text; and

adding a corrected and non-corrected OCR result to the name and address corpus.

11. The method as claimed in claim 10, further comprising:

constructing the name and address corpus based on a collection of names and addresses associated with a target domain; and

training the LSTM neural network language model using the name and address corpus.

12. The method as claimed in claim 11, further comprising:

re-constructing the name and address corpus based on the collection of names and addresses that comprises the corrected and the non-corrected OCR result; and

re-training the LSTM neural network language model using the re-constructed name and address corpus.

13. The method as claimed in any one of claims 10 to 12, further comprising evaluating a log(frequency of edit pairs in history) to determine the most similar text relative to the OCR result.