Evaluation within the Scope of OCR Workflows

Stefan Pletschacher
PRImA Lab, School of Computing, Science and Engineering, University of Salford

ABSTRACT. *OCR (Optical Character Recognition) has become more and more a synonym for intricate document analysis systems going far beyond the core task of classifying pixel patterns. Depending on the nature of the source material it is common practise to employ a whole range of loosely and/or tightly coupled pre-processing, recognition, and post-processing methods in order to achieve good overall results. Historical documents, for instance, which often suffer from physical deterioration and thus artefacts in the scanned images can greatly benefit from image enhancement as part of the pre-processing stage. For modern business documents, at the other end of the spectrum, it is easier to obtain scans of reasonable quality but the textual content might prove difficult to recognise unless background knowledge (for example from customer databases or accounting systems) is incorporated during recognition and post-processing. Evaluation of complex OCR workflows should therefore not only measure the accuracy of the final output but also target all intermediate stages. Not only can this help to reveal bottlenecks and to identify further optimisation potential of the overall system, it can also provide detailed data for making informed decisions if, especially in mass digitisation, project specific requirements regarding cost, quality and time are to be met. Objective and reproducible evaluation of OCR workflows depends on a number of factors. One important aspect is the selection of datasets which have to be representative of the collection and of the problem that is to be examined. In order to measure the success of a method on a given sample it is necessary to have the corresponding ground truth (the true and/or expected result for this particular task) ready. Unfortunately, the creation of ground truth is typically a manual task and therefore extremely time-consuming and costly. As a consequence, it is very difficult to produce datasets of an acceptable size especially if there are only limited resources available. Semi-automated tools for ground truth production can play an important role in overcoming this problem. As for storing ground truth and processing results it is crucial to rely on mature formats which allow a very accurate representation (such as polygons instead of simple bounding boxes for region outlines) and which can be processed by automated evaluation tools. Workflow frameworks can then be used as experimental environments by setting up whole processing chains including evaluation points for individual methods and the overall system. In order to arrive at a practical interpretation of the results it is important to put the measured facts (metrics) into the context of use scenarios (weights) reflecting the needs of concrete digitisation projects.*

CORIA 2012, pp. 3–3, Bordeaux, 21-23 mars 2012