LAREX – Ein Werkzeug zur Layout-Analyse und Segmentierung von frühen Buchdrucken

In den letzten Jahren konnten v. A. durch die vermehrte Anwendung von zeilenbasierten OCR-Verfahren unter Verwendung von neuronalen Netzen große Fortschritte bei der automatischen Text-Erkennung früher Buchdrucke erzielt werden. Neben der Nachkorrektur des OCR-Ergebnisses stellt die Segmentierung der gescannten Seiten häufig die letzte große Hürde in einem OCR-Workflow dar. Gerade bei Inkunabeln ist dieser Vorverarbeitungsschritt meist unerlässlich, da komplexe Layouts mit Bordüren und Zierinitialen zu fehlerhaft segmentierten Zeilen und damit zu schweren OCR-Fehlern führen können. Des Weiteren reicht eine reine Text/Nicht-Text-Trennung für viele Anwendungen nicht aus und häufig soll bereits bei der Segmentierung eine semantische Klassifikation (Bild, Fließtext, Marginalie, Seitenzahl, …) erfolgen.

Im Vortrag wird das semi-automatische Open Source Tool *LAREX* (Layout Analysis and Region EXtraction) vorgestellt. Ausgehend von einem Connected Components-Ansatz ermöglicht *LAREX* eine schnelle, intuitive und für jedermann nachvollziehbare Segmentierung und semantische Auszeichnung von gescannten Dokumenten. Während des Segmentierungsalgorithmus werden benachbarte Vordergrundpixel verwachsen, sodass sich Zeichen zu Wörtern, Wörter zu Zeilen und Zeilen zu Textblöcken verbinden. Die so entstandenen Blöcke werden anhand ihrer Attribute und weniger benutzerdefinierter Regeln semantisch klassifiziert. Die Kriterien, wie z. B. erwartete Größe und Position eines Blocks, können komfortabel an das jeweilige Layout angepasst werden. Ziel ist es, zügig ein Parameterset zu definieren, das das vorliegende Layout gut abbildet, sodass nachträglich von Seite des Nutzers keine, oder nur noch minimale Korrekturen vorgenommen werden müssen, um ein optimales Segmentierungsergebnis zu erreichen. Die Verwendung des weit verbreiteten PageXML-Formats ermöglicht eine simple Einbindung in bestehende Workflows.

LAREX kam zuletzt u. a. im Rahmen des *Narragonien digital*-Projekts bei der Digitalisierung dreier Ausgaben von Sebastian Brants *Narrenschiff* zum Einsatz. Trotz des anspruchsvollen Layouts und den einhergehenden detaillierten Nutzeranforderungen konnte der gesamte Segmentierungsprozess im Schnitt nach nur etwa vier Stunden abgeschlossen werden. Die hohe Qualität des Segmentierungs-ergebnisses spiegelte sich in OCR-Zeichengenauigkeiten von durchgängig über 98% wider.

Christian Reul

Lehrstuhl für Künstliche Intelligenz und Angewandte Informatik

Julius-Maximilians-Universität Würzburg