Segmentierung von historischen Drucken

Eine effiziente Digitalisierung historischer Drucke wie z. B. der zahlreichen Ausgaben von Sebastian Brant´s „Narrenschiff“ ist die Voraussetzung für deren weitere Aufarbeitung. Um die OCR (Optical Character Recognition)-Ergebnisse zu optimieren, ist eine vorangehende Analyse des Layouts einer Seite unabkömmlich. Schwächen in der damaligen Drucktechnik und der stark variierende Aufbau der Seitenlayouts, auch innerhalb einzelner Bücher, erschweren diese Aufgabe erheblich. Selbst State of the Art-Tools sind häufig nicht in der Lage, voll automatisch alle Textblöcke von Bildern, Initialen und Bordüren zu trennen und der korrekten Kategorie (Paragraph, Marginalie, Fußnote, …) zuzuordnen. Eine Korrektur der Ergebnisse gestaltet sich häufig schwierig und zeitaufwändig.

Der im Vortrag vorgestellte Ansatz setzt daher auf eine interaktive Herangehensweise, die simpel aber effektiv arbeitet. Im Mittelpunkt steht dabei ein sehr schneller Connected Components-Ansatz der, basierend auf wenigen Annahmen, in den meisten Fällen bereits sehr gute Ergebnisse erzielt. Nach einer Binärisierung der zu segmentierenden Seite werden zunächst Bilder, Bordüren und Initialen anhand ihrer großflächigen Konturen detektiert, abgespeichert und aus dem Bild entfernt, um die weitere Verarbeitung nicht zu beeinträchtigen. Anschließend folgt eine sogenannte Dilatation, die schwarze Pixel, also den Text, wachsen lässt und nahe Buchstaben, Wörter und Zeilen zu Textblöcken verbindet. Diese können anhand Ihrer Position und Größe klassifiziert werden. Zum jetzigen Zeitpunkt wird dabei zwischen normalen Text (Paragraph), Überschriften, Marginalien und Seitenzahlen unterschieden. Zusammen mit den zuvor detektierten Bildern, Initialen und Bordüren werden die Textblöcke im XML-Format abgespeichert. Um eine Schnittstelle zu anderen Segmentierungs- oder OCR-Tools zu gewährleisten, findet der PageXML-Standard Verwendung. Des Weiteren wird ein Open-Source-Tool vorgestellt, das dem Nutzer eine komfortable, manuelle Nachbearbeitung ermöglicht. Erste Versuche auf verschiedenen Narrenschiff-Versionen lieferten ansprechende Ergebnisse.

Christian Reul

Lehrstuhl für Künstliche Intelligenz und Angewandte Informatik

Julius-Maximilians-Universität Würzburg