Fraunhofer
IAIS

FRAUNHOFER-INSTITUT FÜR INTELLIGENTE ANALYSE- UND INFORMATIONSSYSTEME IAIS

QUALITÄTSGESICHERTE ARTIKELSEPARIERUNG FÜR ZEITUNGSARCHIVE

Die inhaltliche Erschließung von Zeitungsdigitalisaten stellt durch den unterschiedlichen Aufbau und die veränderliche Anordnung von Artikeln eine besondere Herausforderung dar. Die konventionelle Texterkennung ist nicht in der Lage, einzelne Artikel seitenübergreifend zu separieren und ihre Bestandteile wie Überschriften, Autorenangaben, Textkörper, Bilder und Bildunterschriften zu identifizieren. Aber nur mit diesen Metadaten kann ein zeitgemäßes Zeitungsarchiv aufgebaut werden, das die Suche, Referenzierung und Anzeige auf Artikelebene unterstützt – z. B. für die Zweitverwertung auf mobilen Endgeräten oder die semantische Vernetzung.

Da rein maschinelle Ansätze für die Artikelseparierung nach dem heutigen Stand der Forschung nicht immer allen Qualitätsanforderungen in der Praxis genügen, setzt Fraunhofer IAIS auf ein kombiniertes Modell aus automatischen und manuellen Erschließungsmethoden. Zunächst werden die digitalisierten Zeitungsseiten mit layoutbasierten Verfahren segmentiert. Anschließend erfolgt die Klassifikation der resultierenden Layoutelemente und die Zuordnung zu den ursprünglichen Artikeln. Dabei kommen neben probabilistischen Modellen auch regelbasierte Verfahren zum Einsatz, die je nach Layoutformat angepasst und trainiert werden. Erst nach Separierung der Artikel erfolgt die eigentliche Texterkennung, z. B. durch Abbyy FineReader.

Nach der automatischen Erschließung werden die Ergebnisse auf Basis einer spezifischen Erfassungsanweisung manuell annotiert. Hierzu werden Softwarewerkzeuge eingesetzt, die eine schnelle Sichtung und effiziente Nachbearbeitung ermöglichen. Damit können in der Praxis individuell wählbare Qualitätsvorgaben erreicht und manuelle Erschließungsaufwände gezielt an Dienstleister ausgelagert werden.

Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS

Schloss Birlinghoven
53757 Sankt Augustin

Ansprechpartner
Dr. Iuliu Konya
Telefon +49 2241 14-1926
iuliu.konya@iais.fraunhofer.de

www.iais.fraunhofer.de

ArchivInForm
Zentrum für Digitalisierung und Archivleistungen