Automatisierte Text- und Strukturerkennung als Grundlagentechnologie für die Digital Humanities

Veranstalter: Netzwerk zur Förderung der Digital Humanities in Jena (DHnet-Jena)
Datum, Ort: 27.09.2016, Jena
Bericht von: Martin Prell, Historisches Institut, Friedrich-Schiller-Universität Jena

Am 27. September 2016 trafen sich auf Einladung des wissenschaftlichen Netzwerks zur Förderung der Digital Humanities in Jena (DHnet-Jena)[1] zahlreiche ForscherInnen und Interessierte aus dem Bereich der digitalen Text- und Strukturerkennung in Jena. Das wissenschaftliche Programm bestand aus Vorträgen und einem anschließenden Workshop zur Handschriftenerkennungssoftware Transkribus.[2]

Nach einer Begrüßung durch ANDREAS CHRISTOPH (Jena) referierte EVA LANG (Passau) zu den Digitalisierungsstrategien in Bibliotheken und Archiven am Beispiel des Archivs des Bistums Passau (ABP). Zur Vorstellung der Digitalisierungsbemühungen des ABP wählte sie als Beispiel die Matrikelbücher der Pfarrarchive mit einem Umfang von insgesamt circa 10.000 lfd. Metern. Aus zeitökonomischen Gründen fokussiert das ABP derzeit auf die Digitalisierung und elektronische Erfassung der Register zum schnelleren Auffinden der Matrikelbucheinträge.[3] Dabei ging die Referentin sowohl auf die technische Infrastruktur (Datenbank, Suchfunktion, Portaleinbindung) ein als auch auf geplante Features, wie beispielsweise die Einbindung von Geoinformationssystemen zur geographischen Visualisierung der Nameneinträge.

Anschließend sprach GÜNTER MÜHLBERGER (Innsbruck) über das im Rahmen des EU-Projektes READ[4] durchgeführte Projekt zur Handschriftenerkennung – Transkribus. Er skizzierte zunächst die Entwicklung des Projektes aus seinen Vorläufern heraus und kam anschließend auf Aspekte wie Konzeption, Funktionsweise, Genauigkeit/Fehlerquote und Nutzungsszenarien der Software zu sprechen. Das derzeitige System liefert eine Fehlerquote von 10 Prozent bei Buchstaben und 20 Prozent bei Wörtern, wobei betont werden muss, dass diese noch abnehmen wird, je mehr Daten dem System zur Verfügung gestellt werden. Transkribus ist daher an der weiteren Zulieferung von Quellenmaterial durch öffentliche Institutionen und private ForscherInnen interessiert und lädt darüber hinaus auch zu einer Partnerschaft (Memorandum of Understanding) ein, der bereits über 20 Partner angehören und mit der verschiedene zusätzliche Services einhergehen. Ziel des bis 2019 finanzierten Projektes ist es, die verschiedenen Akteure aus Archiven, Bibliotheken, Forschungseinrichtungen und Öffentlichkeit auf einer Plattform zusammenzubringen, um das textuelle Kulturerbe aufbereiten und erforschbar machen zu können. Abschließend legte Mühlberger den Fokus auf ausgewählte Features der Software wie beispielsweise die zahlreichen Upload-Möglichkeiten eigenen Quellenmaterials, die Bereitstellung von Programmierschnittstellen und eines eLearning-Tools sowie die in Entwicklung befindliche App zum Erfassen und Hochladen von Bildmaterial mit einem Mobiltelefon.

Den dritten Vortrag hielt FLORIAN KLEBER (Wien) zum Thema Layoutanalyse als grundlegende Voraussetzung für die Text- und Handschriftenerkennung. Dabei ging er auf den Unterschied zwischen „Layout Analyses“ und „Document Understanding“ ein sowie die einzelnen Bestandteile der Computer Vision (Recognition, Segmentation und Reconstruction). Näher beleuchtete der Referent die Layoutanalyse und ihre Arbeitsschritte (Segmentation, Lokalisation, Klassifikation) zur Erkennung von Zeilen und Wör-

[1] Webseite des DHnet-Jena: http://dhnet.uni-jena.de (06.10.2016).

[2] Webseite des Transkribusprojektes: https://transkribus.eu/Transkribus/ (06.10.2016). Transkribus stellt seine Daten frei zugänglich zur Verfügung unter: https://github.com/transkribus (06.10.2016). Eine User-Anleitung zur Verwendung der Software findet sich unter: https://transkribus.eu/wiki/index.php/Main_Page#How_To_Papers (06.10.2016).

[3] Link zum Onlineportal Für Kirchenbücher – Matricula: http://icar-us.eu/cooperation/online-portals/matricula (06.10.2016).

[4] Webseite des EU-Projektes READ (Recognition and Enrichment of Archival Documents): http://read.transkribus.eu/ (06.10.2016).

tern sowie Herausforderungen angesichts der Vielzahl verschiedenartig strukturierter Texte. Laut Evaluation der International Conference on Document Analysis and Recognition des Jahres 2013 (ICDAR 2013) erkennt die Software des Computer Vision Lab[5] bereits 94 bis 98 Prozent korrekte Zeilen. Als Abschluss des Vortrages hob Kleber mögliche Anwendungsgebiete der Layoutanalyse hervor, die von der Layoutanalyse als spezifische Forschungsrichtung über Keyword-Spotting bis hin zur Handschriftenerkennung reichen.

Der abschließende Vortrag des Tagungsteils oblag RAPHAEL UNTERWEGER (Innsbruck). Er sprach zu den Softwareentwicklungen „RuleApplier" und „Structify" und gab Einblicke in die Funktionsweise beider sowie deren Einbindung in einen umfassenderen Workflow. Der „RuleApplier" bietet das in der Programmiersprache JESS entwickelte Framework, innerhalb dessen Structify[6] als Anwendung zur Strukturerkennung und -erstellung von Zeitschriftenartikeln zum Einsatz kommt. Ein wichtiger Vorteil letzterer Software ist insbesondere ihre schnelle Anpassbarkeit bei Anwendung in größeren Projekten, so der Referent. Anhand dreier Projekte (Leipziger Literaturzeitung, Jenaer Volksblatt, Mitteilungen aus Justus Perthes geographischer Anstalt), die in Kooperation mit der Thüringer Universitäts- und Landesbibliothek (ThULB) entstanden, demonstrierte Unterweger Anwendungsmöglichkeiten und Ergebnisse von Structify. Neben der ThULB nutzt auch die Deutsche Nationalbibliothek die Software zur automatischen Erkennung von Dissertations- und Buchtitelseiten.

Nach dem Mittag trafen sich die Teilnehmer zum zweiten Teil der Veranstaltung, dem Transkribus-Workshop, der von GÜNTER MÜHLBERGER (Innsbruck) und EVA LANG (Wien) durchgeführt wurde. Ziel des Workshops war es, die zentralen Funktionen und notwendigen Arbeitsschritte kennenzulernen und den Umgang mit der Transkribussoftware einzuüben. Im Fokus standen daher die zahlreichen Upload-Möglichkeiten (FTP, DFG-Viewer, PDF, Single Document), das Anlegen eigener Sammlungen und deren kollaboratives Bearbeiten, das Definieren von Grundlinien einzelner Zeilen, die manuelle sowie automatische Transkription, der Export von Daten (TRP, TEI, DOCX etc.), die Anzeige im XML-Viewer sowie die Nutzung des Tabellen-Editors und der Suchfunktionen. Alle Aspekte wurden step by step so vorgeführt, dass die Teilnehmer diese auf den eigenen Rechnern problemlos nachvollziehen konnten. Deutlich wurde dabei unter anderem die bereits in den Vorträgen konstatierte zentrale Notwendigkeit der automatischen, möglichst fehlerfreien Segmentierung der Daten vor der automatischen Texterkennung. Gespannt dürfen die NutzerInnen von Transkribus auf den in naher Zukunft bereitzustellenden übergreifenden Pool (neben den schon existierenden Datenpools einzelner Projekte) bereits trainierter Daten sein, mittels dessen eine solide Texterkennung eingespielter Handschriften auch ohne vorheriges intensives Trainieren des Systems möglich sein soll.

Tagung und Workshop haben sehr anschaulich die Notwendigkeit, das Potential, die erforderlichen Voraussetzungen und bereits praktikable Lösungen zentraler DH-Technologien vor Augen geführt, die aus den textbearbeitenden Wissenschaften zukünftig nicht mehr wegzudenken sind. Die Verringerung bestehender Fehlerquoten, der Ausbau zu umfassenden Forschungsumgebungen, der Betrieb und eine beständige Weiterentwicklung setzen allerdings eine verstetigte Finanzierung dieser und ähnlicher Projekte sowie intensivere Kooperationen mit Bibliotheken und Archiven voraus. Damit gehen zugleich auch erforderliche neue Grundsatzüberlegungen zum Zugang historischer Dokumente einher, der durch Reproduktionsgebühren oft erschwert oder gar unmöglich gemacht wird.

Konferenzübersicht:

Andreas Christoph (Jena): Begrüßung

Eva Lang (Passau): Vom Kirchenbuch zur Datenbank. Digitalisierungsstrategien in Bibliotheken und Archiven

Günther Mühlberger (Innsbruck): Transkribus. Eine virtuelle Forschungsumgebung zur automatisierten Texterkennung gedruckter

[5] Webseite des Computer Vision Lab der TU Wien: http://www.caa.tuwien.ac.at/cvl/ (06.10.2016).

[6] Link zum Structify-Tool: http://dbis-halvar.uibk.ac.at/dokuwiki/doku.php?id=main:structify (06.10.2016).

und handschriftlicher Dokumente

Florian Kleber (Wien): Ohne Layoutanalyse keine Text- und Handschriftenerkennung

Raphael Unterweger (Innsbruck): Strukturdaten und Dokumentenerkennung mit RuleApplier und Structify

Günther Mühlberger (Innsbruck), Eva Lang (Passau): Transkribus-Workshop

Tagungsbericht *Automatisierte Text- und Strukturerkennung als Grundlagentechnologie für die Digital Humanities*. 27.09.2016, Jena, in: H-Soz-Kult 17.11.2016.