Gemischte OCR-Modelle für die Erkennung gedruckter Texte seit Gutenberg

Der Durchbruch in der OCR-Erkennung historischer Drucke aufgrund trainierbarer neuronaler Netze gibt Anlass zur Hoffnung, das gedruckte Erbe in naher Zukunft in großen Teilen in einer maschinenverarbeitbaren Form zur Verfügung zu haben und damit den Traum der universellen Bibliothek, der ansatzweise in der Bibliothek von Alexandria für die Alte Welt einmal existiert hat, erneut in unvergleichlich größerem Umfang zu verwirklichen. Das spezifisch Neue an dieser zukünftigen Universalbibliothek liegt dann in der Möglichkeit, die Texte nicht nur als Bilder wahrnehmen zu müssen, sondern sie mit maschinellen Verfahren verarbeiten zu können (Suche, Indizierung, Annotation), was angesichts der Menge der Daten eine angemessene Weiterentwicklung traditioneller Textkritik darstellt.

Die besten Ergebnisse mit über 98% Zeichengenauigkeit (über 90% korrekt erkannte Wörter) werden mit Modellen erreicht, die individuell auf eine spezifische Typographie trainiert wurden, die durch den gleichen Vorrat an Lettern, gleiche Zeilenbreiten und ähnliche Wortabstände definiert ist, was etwa für eine Reihe von Büchern einer bestimmten Offizin zutrifft. Auf anderen Typographien funktionieren diese Modelle meist wesentlich schlechter (70-90% Erkennungsrate). Zur großflächigen OCR-Erkennung mit einer akzeptablen Erkennungsrate (>90%) werden daher generalisierte Modelle benötigt, wie sie etwa die Polyfont- oder Omnifontmodelle kommerzieller OCR-Engines für moderne Drucke liefern, die auf einigen hundert Schriftarten trainiert sind. Da es allein in der Inkunabelzeit mehr als 2000 Offizinen mit jeweils mehreren selbst erstellten Schriften gab, können derzeit wegen der sehr begrenzten Verfügbarkeit diplomatischer Transkriptionen nur gemischte Modelle für einige wenige Schriften trainiert werden.

Der Vortrag berichtet über Experimente mit dem Training gemischter Modelle für Antiqua- und Frakturschriften über einen Zeitraum von 400 Jahren Druckgeschichte sowie für Frakturdrucke des 19. Jahrhunderts und zeigt, dass die auf wenigen Drucken trainierten gemischten Modelle tatsächlich akzeptable Erkennungsraten liefern und im Falle der Frakturschriften des 19. Jahrhunderts die Fehlerrate von ABBYY und Tesseract im Mittel um 30% verringern.

Dr. Uwe Springmann

Centrum für Informations- und Sprachverarbeitung (CIS)

Ludwig-Maximilians-Universität München