Sanskrit und Computer

Ein Programm zur Sprachanalyse von indischen Texten mit integriertem OCR-Modul

Inaugural-Dissertation
zur
Erlangung des Doktorgrades
am
Fachbereich Geschichts- und Kulturwissenschaften der FU Berlin

Vorgelegt von
Oliver Hellwig
aus Berlin

Erster Gutachter: Univ. Prof. Dr. Harry Falk
Zweiter Gutachter: Dr. Habil. Gerhard Ehlers

Datum der Disputation: 16.11.2002

Zusammenfassung

Thema der Arbeit ist die digitale Verarbeitung von Sanskrit-Texten. Dazu wurden ein Programm zur Digitalisierung in Devanagari gedruckter Texte (OCR) und ein Programm zum lexikalischen und morphologischen Tagging digitalisierter Texte entworfen und in C++ implementiert.

Die Digitalisierung wird mithilfe von Gruppen neuronaler Backpropagation-Netze durchgeführt, die auf Formbeschreibungen der Nagari-Zeichen trainiert werden. Zusätzlich zu fest installierten Klassifikatoren können trainierbare Klassifikatoren auf Basis des k-Nearest-Neighbours-Algorithmus aktiviert werden. Das OCR-Modul erreicht eine fontabhängige Erkennungsgenauigkeit von ca. 93-95%, wobei ein Grossteil der Fehler im Rahmen der Zeilensegmentierung verursacht wird. Die Möglichkeiten eines sprachbasierten Postprocessings der Daten werden diskutiert und seine Grenzen einer Nachbehandlung aufgrund sprachimmanenter Probleme (Sandhi, Homonymie) aufgezeigt.

Im Rahmen des lexikalischen und morphologischen Taggings werden die Hauptprobleme – Sandhi, Grösse des Wortschatzes und Kompositabildung im Sanskrit – durch einen mehrstufigen rekursiven Auflösungsalgorithmus gelöst, der auf eine fest codierte Sammlung von sprachlichen Regeln und eine umfangreiche Datenbank mit lexikalischen und grammatikalischen Informationen zurückgreift.

Aufbauend auf OCR und Tagging erlaubt die Programmkonstruktion die sukzessive Erstellung einer Datenbank getaggter Sanskrit-Texte, die zum ersten Mal eine effiziente lexikonbasierte Suche in diesen Texten möglich macht.

Inhaltsverzeichnis