Forschungspräsentation im Promotionsprogramm “Empirische Sprachverarbeitung”

Klaus U. Schulz, Centrum für Informations- und Sprachverarbeitung (CIS), LMU München
Zeit: Donnerstag, 03.12.2009, 16.16 – 17.45 Uhr
Ort: Institut für Phonetik und Sprachverarbeitung, Schellingstr. 3, VG, 80799 München, Raum 226 (Institutsbibliothek)

Sprachtechnologie zur Verbesserung der optischen Charaktererkennung auf historischen Dokumenten

Bibliotheken und Archive haben auf breiter Front damit begonnen, historische Dokumente und Texte vom Papierformat in ein elektronisches Format überzuführen, um sie im Internet zugänglich zu machen. Allerdings liegt gegenwärtig noch die Mehrzahl der digitalisierten Bestände nur als Bild vor, wodurch keine Suche/Recherche in den textuellen Inhalten möglich ist. Eine solche Suche würde voraussetzen, dass die Dokumente zuvor mittels OCR in ein textuelles Format überführt werden. Die OCR-Erfassung historischer Dokumente ist jedoch problematisch und führt oft zu schlechten oder unbrauchbaren Ergebnissen.

Im Vortrag sollen Methoden vorgestellt werden, um mit Sprachtechnologie OCR auf historischen Dokumenten zu verbessern. Wir gehen zunächst auf traditionelle Methoden ein, bei denen Lexika und Sprachmodelle zur Überprüfung/Verbesserung eingesetzt werden. Im Kontext historischer Dokumente ergeben sich hierbei neue Probleme und Lösungsansätze. Im Anschluss zeigen wir, wie Lexika und Sprachmodelle optimiert werden können, indem man Inhalt und Gegenstand des zu erfassenden Dokuments beim Design mitberücksichtigt. Dies stellt einen ersten Ansatz für die allgemeinere Zielrichtung dar, OCR dynamisch und adaptiv dem Eingabedokument anzupassen. Aktuelle Forschungsansätze zu "Feedbackmechanismen" stellen einen weitergehenden Schritt in Richtung Adaptivität dar. Hierbei wird das Ergebnis eines ersten OCR-Durchlaufs global analysiert, um statistisch Rückschlüsse auf sprachliche Eigenheiten des Textes und auf typische Erkennungsfehler zu erhalten. Die so erhaltenen Profile werden dann eingesetzt, um OCR oder Nachkorrektursysteme zu optimieren. Als Abschluss gehen wir kurz auf die derzeit unzureichenden organisatorischen und institutionellen Voraussetzungen ein, um die oben genannten Fragestellungen auf breit fundierter empirischer Basis zu untersuchen.