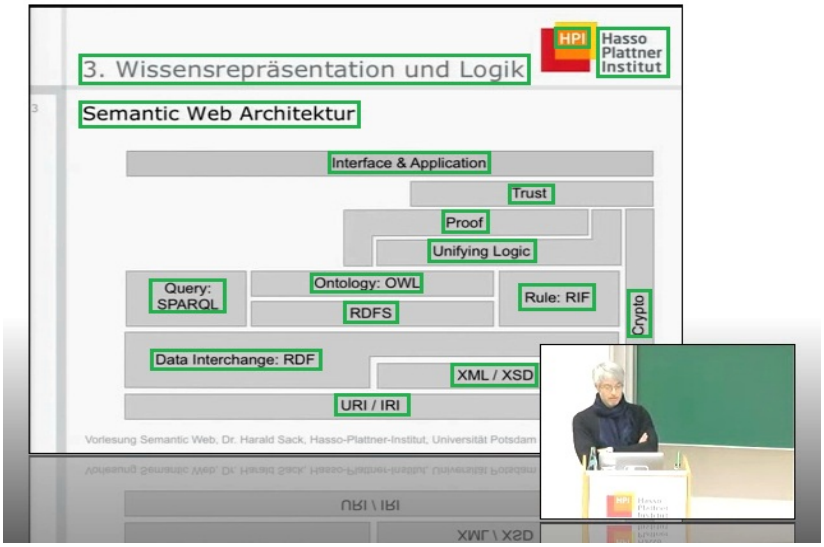
Spellcorrection auf OCR-generiertem Text

Hintergrund

Im Rahmen der automatischen Videoanalyse bei Tele-Task [1] und dem SEMEX-Framework [2] wird eine Texterkennung durchgeführt. Nach einigen Vorverarbeitungsschritten werden zur eigentlichen Texterkennung die OCR-Werkzeuge Tesseract [3] und Ocropus [4] verwendet. Eine OCR extrahiert aus einem Bild die visualisierten Schriftzeichen, eine solche Erkennung



ist jedoch oftmals fehlerbehaftet, was eine Nachverarbeitung notwendig macht.

Zur Korrektur von Text existieren Rechtschreibkorrekturprogramme und -bibliotheken, u.a. auch frei verfügbare, wie z.B. Hunspell [5] oder Aspell. Diese Rechtschreibkorrekturen sind jedoch auf geschriebenen bzw. mit der Tastatur erstellten Text optimiert und korrigieren daher vornehmlich typische (menschliche) Rechtschreibfehler und „Vertipper“. Dieses Verhalten ist für optisch erkannten Text nicht optimal, wo in erster Linie visuell ähnliche Zeichen bei der Erkennung miteinander verwechselt werden. Dies tritt insbesondere bei Video-OCR auf, die auf Material mit geringer Auflösung und zahlreichen Komprimierungsartefakten arbeitet. Eine OCR-Spellcorrection sollte daher also die visuelle Ähnlichkeit von Zeichen berücksichtigen.

Einerseits beinhaltet die Entwicklung einer solchen Spellcorrection ähnliche Anforderungen wie bei einer traditionellen Rechtschreibkorrektur, wie z.B. die Generierung von Wörterbüchern, unter Berücksichtigung eines effizienten Zugriffs, durch optimierte Speicherung und / oder syntaktische Regeln zur Bildung von Flexionsformen und Wortkomposita. Des Weiteren bedarf es der Entwicklung und Implementierung von Algorithmen zum fehlertoleranten Fuzzy-Stringmatching, um sinnvolle Wortkandidaten für eine Korrektur zu ermitteln.

Darüber hinaus müssen für eine OCR-Spellcorrection andere Abstandsmaße zum Stringvergleich entwickelt werden, die dem Ausgangsmaterial OCR-generiertem Text

entsprechen. Zur nachträglichen manuellen Verbesserung nicht erkannter bzw. falsch korrigierter Wörter sollte über Möglichkeiten nachgedacht werden, dieses Feedback für kommende Korrekturen einzubeziehen.

Um die Laufzeiten in einem sinnvollen Rahmen zu begrenzen, ist über Optimierungsmöglichkeiten nachzudenken. Diese können sowohl durch massive Parallelisierung, In-Memory-Technologien als auch über den geeigneten Einsatz von Indizes und effizienten Datenstrukturen erreicht werden.

Zur Entwicklung, Verifikation und Evaluation der OCR-Spellcorrection stehen große Menge an OCR-generiertem Text teilweise mit der dazugehörigen Ground-Truth aus Videomaterial sowie aus gescannten Dokumenten zur Verfügung.

Beschreibung

Im Rahmen des Masterprojekts sollen folgende Leistungen erbracht werden:

1. Einarbeitung in die Grundlagen bestehender Rechtschreibkorrektur-Software und der Besonderheiten der Video-Texterkennung.
2. Entwurf eines Systemmodells für die Implementierung unter Berücksichtigung bestehender Softwarekomponenten
3. Implementierung der projektierten OCR-Spellcorrection
4. Fortlaufende Evaluation der Ergebnisse

Kontakt

Fachgebiet: Internet Technologien und -Systeme

- Fachgebietsleiter: Prof. Dr. Christoph Meinel
- Ansprechpartner: Magnus Knuth, Dr. Harald Sack

Referenzen

[1] Tele-Task: http://www.tele-task.de/

[2] SEMEX (Screencast + Publikationen): http://bit.ly/hpi-semex

[3] Tesseract-OCR: http://code.google.com/p/tesseract-ocr/

[4] Ocropus: http://code.google.com/p/ocropus/

[5] Hunspell: http://hunspell.sourceforge.net/