CARL VON OSSIETZKY
universität OLDENBURG
FAKULTÄT II
INFORMATIK, WIRTSCHAFTS-
UND RECHTSWISSENSCHAFTEN

CARL VON OSSIETZKY UNIVERSITÄT OLDENBURG · 26111 OLDENBURG

Identifikation, Interpretation und Überführung heterogener digitaler Dokumentinhalte in eine strukturierte Datenform

Masterarbeit

Die Digitalisierung verändert die heutige Geschäftsform und wirkt sich auf Unternehmen aller Branchen und Verbraucher auf der ganzen Welt aus. Eine wesentliche Folge der Digitalisierung liegt dabei in der stetig wachsenden Menge digitaler Geschäftsdokumente.

Im Gegensatz zu den in den relationalen Datenbanken verwalteten strukturierten Daten liegt eine große Herausforderung in der Verarbeitung der unstrukturierten Dokumente. Werden konkrete Informationen aus den Dokumenten benötigt, so lassen sich diese nicht über klassische Datenbankabfragen extrahieren. Dies gilt insbesondere dann, wenn es sich um gescannte Dokumente handelt. Während die Informationsentnahme in der derzeitigen Unternehmenspraxis im Falle kleiner Losgrößen noch manuell durch Mitarbeiter vorgenommen wird, so wird dieses Ziel in der Forschung inzwischen unter dem neuartigen Zweig der Document AI oder auch Document Intelligence adressiert. Im Gegensatz zu manuell erstellten heuristik- bzw. regelbasierten Workflows, wird dabei unter Nutzung von Deep-Learning-Ansätzen versucht, die Gesamtheit aller Dokumente samt ihrer Inhalte automatisiert zu klassifizieren, zu interpretieren und schließlich zu extrahieren.

Aufgrund der hohen Heterogenität von Geschäftsdokumenten sowie der begrenzten Menge ausreichend annotierter Trainingsdaten ist eine verlässliche Automatisierung der Informationsüberführung in eine strukturierte Datenform allerdings noch nicht vollständig gegeben.

Im Rahmen dieser Masterarbeit sollen die aktuellen Herausforderungen der Informationsextraktion aus (gescannten) Geschäftsdokumenten mit Hilfe moderner Verfahren aus dem Bereich der Optical Character Recognition sowie des Deep Learnings untersucht werden. Zu den Kernthemen der Arbeit zählen:

- Identifikation und Darlegung aktueller An- bzw. Herausforderungen hinsichtlich der Identifikation von Dokumentsegmenten sowie ihrer Interpretation (Erkennung von Header-, Footer-, Textbereichen; Interpretation heterogener tabellarischer Inhalte etc.)
- Datenerhebung und Verarbeitung zum Zwecke von Trainingsprozessen
- Konzeptionierung, Modellierung und Test eines Prototyps
- Evaluation der Ergebnisse

Kontakt:
Gerrit Schumann
Raum: A4-3-318
Tel: 0151 64646465
gerrit.schumann@uol.de

DEPARTMENT FÜR INFORMATIK

ABTEILUNG
WIRTSCHAFTSINFORMATIK I
VERY LARGE BUSINESS APPLICATIONS

PROF. DR. JORGE MARX GÓMEZ

TELEFONDURCHWAHL
(0441) 7 98 – 4470
Sekretariat – 4478

FAX
(0441) 7 98 – 4472

EMAIL
jorge.marx.gomez@uni-oldenburg.de

GEBÄUDE A4
Uhlhornsweg 84 – Raum A4 3-315

OLDENBURG
25.05.2020

VERY LARGE
BUSINESS
APPLICATIONS
Carl von Ossietzky
Universität Oldenburg

POSTANSCHRIFT
D-26111 Oldenburg

PAKETANSCHRIFT
Ammerländer Heerstraße 114 - 118
D-26129 Oldenburg

TELEFONZENTRALE
(0441) 7 98 – 0

BANKVERBINDUNG
Landessparkasse zu Oldenburg

IBAN: DE46 2805 0100 0001 9881 12
BIC: SLZODE22