Extraktion strukturierter Daten aus Rechnungsdokumenten mittels Maschinellen Lernens

Thema:
Extraktion strukturierter Daten aus Rechnungsdokumenten mittels Maschinellen Lernens
Art:
MA
BetreuerIn:
Christian Wolff
BearbeiterIn:
Felix Wende
ErstgutachterIn:
Christian Wolff
ZweitgutachterIn:
Raphael Wimmer
Status:
abgeschlossen
Stichworte:
Machine Learning, Information Extraction, Invoice Analysis
angelegt:
2021-06-07
Anmeldung:
2021-06-21
Antrittsvortrag:
2021-06-21
Abschlussvortrag:
2021-11-08

Hintergrund

Im Rahmen der Digitalisierung im Arbeitsumfeld werden Rechnungen von der Papierform in die elektronische überführt oder schon rein digital versendet und bearbeitet. Die Ablage erfolgt unter bestimmten Vorschriften, um im Falle einer Prüfung die Revisionssicherheit gewährleisten zu können. Durch diese Digitalisierung können einige Prozesse verbessert, optimiert oder kostengünstiger umgesetzt werden. Das gezielte Durchsuchen der Daten nach bestimmten Informationen bleibt jedoch mühselig, auch wenn keine schweren Ordner mehr gewälzt werden müssen. Ein erster Schritt ist die Texterkennung mittels Optical Character Recognition (OCR) um den Inhalt von Dokumenten, bei denen noch kein Text hinterlegt ist (z.B. eingescannte oder abfotografierte Papierrechnungen), verfügbar zu machen. Eine weitere Verbesserung und Vereinfachung des Suchprozesses, wird durch das automatische extrahieren von relevanten Daten aus den Rechnungen und deren Speicherung in einer strukturierten und durchsuchbaren Form erreicht. Relevante Daten können zum Beispiel die im §14 des Umsatzsteuergesetzes festgelegten Angaben sein, die eine Rechnung beinhalten muss. Dieser Schritt wird bereits von verschiedenen Anbietern mittels Neuronaler Netze umgesetzt. Diese sind jedoch meist nicht Open Source, sondern in kommerziellen Anwendungen integriert. Noch mehr

eingeschränkt ist die Verfügbarkeit von Trainingsdatensätzen, da Rechnungen sehr sensible Informationen enthalten.

Zielsetzung der Arbeit

Ziel dieser Masterarbeit ist die Entwicklung eines Modells zur Extraktion relevanter Daten aus Rechnungsdokumenten, um diese in einer strukturierten Form abzuspeichern und somit effizient durchsuchbar und weiterverarbeitbar zu machen. Dazu sollen Methoden des Maschinellen Lernens, in Kombination mit OCR und Natural Language Processing (NLP), angewandt werden, um vergleichbare Erkennungsleistungen zu erzielen, wie aktuelle Ansätze durch Deep Learning (DL). Als zugrundeliegender Datensatz werden reale Rechnungen zur Verfügung gestellt, die zur Nutzung entsprechend aufbereitet werden müssen. Zum Vergleich des entwickelten Klassifizierers, soll dieser einer Open Source Lösung gegenüber gestellt werden, welche ein Neuronales Netz verwendet (InvoiceNet).

Konkrete Aufgaben

1. Literaturrecherche
2. Toolrecherche (OCR, pdf2python, Annotation)
3. Dokumente digitalisieren (Scannen, OCR, sortieren)
4. Dokumente annotieren
5. Inhalte extrahieren (Text, Position, Größe, etc.) und labeln mit den Annotationen
6. Trainingsdatensatz erstellen
7. Entwicklung des ML-Modells
    - Modell-Auswahl
    - Feature Engineering
    - Feature Selection
    - Parameter Optimization
8. Datenaufbereitung
9. Evaluation und Vergleich mit DL Ansatz

Erwartete Vorkenntnisse

- Machine Learning
- Listenpunktpython

Weiterführende Quellen

- R. B. Palm, F. Laws and O. Winther, „Attend, Copy, Parse End-to-end Information Extraction from Documents,“ 2019 International Conference on Document Analysis and Recognition (ICDAR), Sydney, Australia, 2019, pp. 329-336, doi: 10.1109/ICDAR.2019.00060.
- R. B. Palm, O. Winther and F. Laws, „CloudScan - A Configuration-Free Invoice Analysis System Using Recurrent Neural Networks,“ 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, 2017, pp. 406-413, doi: 10.1109/ICDAR.2017.74.
- D. Schuster et al., „Intellix – End-User Trained Information Extraction for Document Archiving,“ 2013 12th International Conference on Document Analysis and Recognition, Washington, DC,

2013, pp. 101-105, doi: 10.1109/ICDAR.2013.28.
- F. Schulz, M. Ebbecke, M. Gillmann, B. Adrian, S. Agne and A. Dengel, „Seizing the Treasure: Transferring Knowledge in Invoice Analysis," 2009 10th International Conference on Document Analysis and Recognition, Barcelona, 2009, pp. 848-852, doi: 10.1109/ICDAR.2009.47.
- Holt, X., & Chisholm, A. (2018, December). Extracting structured data from invoices. In Proceedings of the Australasian Language Technology Association Workshop 2018 (pp. 53-59).
- Bardelli, C., Rondinelli, A., Vecchio, R., & Figini, S. (2020). Automatic electronic invoice classification using machine learning models. Machine Learning and Knowledge Extraction, 2(4), 617-629.