Zürcher Hochschule
für Angewandte Wissenschaften

School of Engineering

InIT Institut für angewandte Informationstechnologie

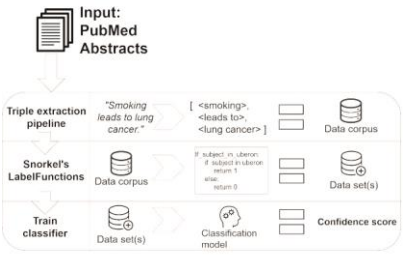
Draining the Data Swamp – Integration of Unstructured and Structured Data

Diplomierende
David Lüscher
Vele Ristovski

Dozierende
Martin Braschler
Kurt Stockinger

Open Data Exploration hat sowohl für die Wissenschaft als auch die Wirtschaft an Bedeutung gewonnen. Information Extraction ist eine vielverbreitete Methode, um unstrukturierte Daten wie Text in strukturierte Daten zu überführen. Die Integration von aus Text gewonnenen Daten mit weiteren strukturierten Datensätzen ist interessant, um Endanwendern die Formulierung umfassenderer Abfragen zu ermöglichen als die einfache Schlagwortsuche. Noise ist jedoch eine zentrale Herausforderung bei Information Extraction. Für den integrierten Zugriff ist es deswegen von Interesse, die Unsicherheit der Information Extraction repräsentieren zu können.

In der vorliegenden Arbeit entwickelten und implementierten wir ein Konzept zur Vorhersage von Class Labels, welche auf die Relevanz eines Informationstripels schliessen lassen. Wir verarbeiteten 38'869 Abstracts aus der biomedizinischen Literatur, die mittels Information Extraction zur Generierung von 511'253 Subjekt-Prädikat-Objekt Aussagen führten. Wir verwendeten Snorkel, ein Weak-Supervision-Framework, um die Subjekt-Prädikat-Objekt Aussagen mit vier Label-Funktionen zu annotieren und erstellten vier unterschiedlich annotierte Datensätze. Anschliessend führten wir acht Experimente durch, um einen Classifier für die Vorhersage von Class Labels zu trainieren. Im Rahmen dieser Arbeit haben wir gezeigt, dass Snorkel das Potenzial ein geeignetes Weak-Supervision-Framework hat, um einen Datenkorpus zu annotieren und basierend darauf einen Classifier zu trainieren, welcher Noise in Subjekt-Prädikat-Objekt Aussagen erkennt. Es ist jedoch weitere Forschung notwendig, um zusätzliche Label-Funktionen zu entwerfen und evaluieren, um die Annotierung des Daten-Korpus zu verbessern.



Konzept zur Vorhersage von Class Labels für Informationstriples zur Darstellung der Unsicherheit der Extraktion

snorkel

https://www.snorkel.org/