Zürcher Hochschule
für Angewandte Wissenschaften

School of
Engineering

InIT Institut für angewandte
Informationstechnologie

Automatische Übersetzung von natürlicher Sprache nach SPARQL mit neuronalen Netzen

Diplomierende
Sebastian Drozdz
Nicolas Hoferer

Dozierende
Mark Cieliebak
Kurt Stockinger

Die vorliegende Arbeit setzt sich mit der Anwendung von neuronalen Netzen im Bereich der Sprachübersetzung auseinander. Das Ziel der Arbeit ist es, in natürlicher Sprache gestellte Fragen in entsprechende SPARQL-Anfragen zu übersetzten. SPARQL ist eine graphen-basierte Abfragesprache für RDF, welche eine wichtige Rolle beim Arbeiten mit dem Semantic Web spielt. Damit wäre eine Vielzahl von Datenquellen, wie zum Beispiel das DBPedia, einer breiten Öffentlichkeit zugänglich. Die computergestützte Sprachübersetzung bringt jedoch eine Vielzahl von Problemen mit sich. Die Mehrdeutigkeit der natürlichen Sprache bildet eines der Hauptprobleme, warum es solche Systeme bis anhin nicht zur Marktreife geschafft haben.

Die Autoren verwenden einen RNN (Recurrent Neural Network)-basierenden Ansatz, um natürliche Sprache in SPARQL zu übersetzten, da in den letzten Jahren grosse Fortschritte im Bereich des maschinellen Lernens gemacht wurden und viele Probleme aus der Sprachverarbeitung an das Modell ausgelagert werden. Ausserdem zeigt diese Art von Netzwerken bereits sehr gute Resultate bei klassischen Übersetzungsproblemen.Durch die Anwendung von verschiedenen Massnahmen und Preprocessing-Verfahren konnte ein bestehendes Modell signifikant verbessert werden. Es standen Trainingsdaten aus dem SQA2018 Challenge und eine dazugehörige DBPedia-Kopie zur Verfügung.

Das ursprüngliche System erzielte auf den Daten des SQA2018 Challenge anfänglich keine guten Resultate. Dank Anpassungen des Modells, wie zum Beispiel die Verwendung von Attention und Anpassung des Dropouts, konnte die Präzision gesteigert werden. Um die relativ kleine Anzahl von 5000 Trainingsfragen auszugleichen, wurden verschiedene Named Entity Recognition (NER)-Methoden eingesetzt, welche dem Modell halfen, die Ressourcen besser zu erkennen, da diese durch generische Platzhalter in der ursprünglichen Frage ersetzt wurden. Diese Erweiterung ermöglicht es auch, Entitäten zu erkennen, welche nicht in den Trainingsdaten vorkommen. ImVerlauf der Arbeiten wurden diese Methoden auch auf die Ontologie (Ontology) und deren Eigenschaften (Properties) angewendet, wodurch die Präzision weiterhin gesteigert werden konnte.

Bild klein 1.

Bild klein 2.