15.05.17

Protokoll zur Sitzung vom 15.05.2017 – Computerlinguistisches Arbeiten

1. Referat: Tobias Eder, Exploiting bilingual word embeddings to establish translational equivalence (Alexander Fraser, Fabienne Braun)

Der Vortrag beginnt mit einem allgemeinen Überblick über das Thema. Die Motivation für diese Arbeit beruht darauf, dass Übersetzungen auf bestimmten Domänen sich oft als eine schwierige Aufgabe herausstellen, da auch beispielsweise unbekannte Wörter im Text auftreten können, bzw. sehr fachspezifische Wörter nicht korrekt übersetzt werden. Ziel der Arbeit ist es eine Übersetzung ohne Wörterbücher zu ermöglichen und insbesondere auch domain-abhängige Übersetzungen zu verbessern.

Als Einführung erklärt Tobias zunächst was Word Embeddings allgemein sind. Er definiert sie als Repräsentationen von Wörtern in einem Vektorraum. D.h. dass syntaktische oder semantische Eigenschaften von Wörtern als Vektoren dargestellt werden. Die sogenannte Cosine Similarity misst die Ähnlichkeit der Wörter im Vektorraum anhand ihrer Abstände. Eine wichtige Annahme bildet die Distributionshypthese.

Für die Umsetzung seines Experiments benutzt er verschiedene Tools. Zum einen das Word Embedding Toolkit Word2Vec, das u.a. ein Skipgram-Modell enthält. Ein weiteres Tool, das er in seiner praktischen Ausführung benutzt ist fastText, welches auf Methoden des Word-Representation Learning mit Subword-Informationen zurückgreift. Er veranschaulicht sein Vorgehen durch verschiedene Graphiken, u.a. auch die Textklassifikation mit einem linearen Modell. Auch ein Regressionsmodell zur Veranschaulichung der Daten wurde herangezogen.

Im Anschluss präsentiert er die unterschiedlichen Korpora, die er für die Arbeit benutzt und beschreibt sein Experimentaufbau. Für das Experiment werden vier unterschiedliche Korpora benötigt: ein allgemeiner Korpus (General, mit ca. 100M Tokens), ein Korpus aus dem medizinischen Bereich (Medical Big, mit ca. 50 M Tokens), EMEA mit ca. 4M Tokens und ein Korpus mit natürlicher Sprache (Ted Talks, mit ca. 2M Tokens). Es wurden unterschiedliche Embeddings benutzt, z.B. CBOW und Skipgram. Als Übersetzungssprache benutzt er Englisch, das ins Deutsche übersetzt wird. Für jede Domäne wird ein eigenes Testset erstellt.

Zum Schluss gibt Tobias Auskunft über seinen aktuellen Stand. Das System muss noch für niedrig-frequentierte Wörter getestet werden. Weitere Regularisierungsmethoden müssen noch erforscht werden und das System muss zudem noch mit OOV (Out-of-Vocabulary)-Wörtern evaluiert werden.