Exploiting Bilingual Word Embeddings to Establish Translational Equivalence

Der erste Vortrag des Tages wurde von Tobias Eder gehalten, dessen Betreuer Dr. Fraser und Dr. Braune sind. Tobias begann damit, die Motivation seiner Arbeit zu erläutern. Es geht um die Aufgabe der Übersetzung eines Textes aus spezifischen Domains. Zum Beispiel aus dem Bereich der Medizin. Wichtig bei dieser Übersetzung ist, dass dies ohne ein Wörterbuch geschehen soll. Auch sollen unbekannte Wörter im Text verarbeitet werden können.

Dann hat er das Modell, das dies realisieren soll, näher beschrieben. Dabei hat er zwei wichtige Verfahren behandelt. Zum einen das „Word2Vector“ Verfahren und zum anderen das „Fast Text“ Verfahren. Bei der Word2Vektor Methode wird jedes Wort von einem Vektor repräsentiert. Dabei wurde dann die Distributionshypothese aufgestellt. Diese Hypothese besagt, dass es zwischen den Vektoren zu syntaktischen und zu semantischen Ähnlichkeiten kommen kann. Bei dieser Ähnlichkeit kann man Wiederholungen der Richtungen der Vektoren feststellen, also im Grunde stellt dies Beziehungen zwischen den Vektoren da. Er bezeichnete dies zum Teil auch als ein „überraschendes“ Ergebnis. Dazu gibt er noch zwei Beispiele, wo dieses Verfahren auch vorkommt. Nämlich im Google Toolkit, aus dem Jahr 2013 und dem CBOW und Skipgram Modell. Die Fast Text Methode lernt mit Hilfe von Wort Repräsentationen und mit Hilfe von Subwords, also „n-Grammen“. Dazu versucht man noch Wörtern, die nicht vorkommen, ebenfalls einen Wort-Vektor zu geben. Dieses Verfahren wurde auch von Facebook im Jahr 2016 verwendet. Ein Problem ist jedoch, dass wenn die Sätze Fehlinformationen enthalten, das auch die Gewichtung der Vektoren durcheinander kommt.

Dann hat er über sich annähernde Lineare Abbildungen gesprochen, bei der immer zwei Sprachen miteinander Verglichen werden, zum Beispiel Englisch und Deutsch. Es ging auch um einen Vorgang, den er lineare Regression genannt hat. Dabei wird eine Matrix verwendet, die Ähnlichkeiten sucht. Außerdem bestraft man dabei hohe Gewichte mit einer Minimierung. Das hat den Zweck, eine Anpassung auf die Testdaten zu verhindern.

Anschließend hat er einige Korpora vorgestellt, beispielsweise Wikipedia, TED Talks, EMEA und 1000 hochfrequentierte Wörter aus dem Medizinbereich.

Am Schluss hat er noch über die noch zu erledigenden Schritte gesprochen. Zum einen will er noch niedrigfrequentierte Wörter behandeln und er will versuchen bessere Abbildungen zu erhalten. Des Weiteren sucht er nach anderen Methoden für die Regularisierung.