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 von Texten aus spezifischen Domains. Ein Beispiel für so eine Domain ist der Bereich der Medizin. Wichtig bei dieser Übersetzung ist, dass dies ohne ein Wörterbuch geschehen soll. Auch sollen unbekannte Wörter im Text verarbeitet werden können.

Weiterhin hat er das Modell, das dies realisieren soll, näher beschrieben.
Ein Kernproblem bei der Datenverarbeitung ist, dass man nur spärlich Daten zur Verfügung hat, anders bei Audiodateien, wo die Daten sehr „dicht“ sind. Wörter werden hier als atomare Einheiten gesehen, weshalb sie nur in einem Bestimmten Kontext auftauchen. Er verwendet deshalb in seiner Arbeit „word embeddings“, die aus einem Text ein Vektorraum-modell macht. Vereinfacht gesagt: jedes Wort wird dabei von einem Vektor repräsentiert. Anschließend erläuterte er die Distributionshypothese. Diese besagt in etwa, dass wenn der Kontext eines Wortes ähnlich ist, wie von einem anderen Wort, dann können sie semantische oder syntaktische Ähnlichkeiten zueinander haben.

Zwei Beispiele für so ein Vektormodell sind das „Word2Vec“ Verfahren und das „Fast Text“ Verfahren. Das Word2Vec Modell wurde 2013 von Google entwickelt. Es verwendet zwei Modelle, das CBOW-Modell und das Skipgram Modell. Bei dem CBOW Modell versucht man, für einen Kontext, mögliche Wörter vorherzusagen. Bei dem Skipgram Modell versucht man vorherzusagen, in welchem Kontext ein Wort vorgekommen ist.
Das Fast Text Verfahren wurde 2016 von Facebook veröffentlicht. Dieses Verfahren verwendet zusätzlich „Subword-inforamtion“ mit n-grammen. Dabei werden Wörter, die morphologische Unterschiede haben, aber denselben Stamm besitzen, trotzdem im Vektorraum geortet. Zusätzlich lassen sich für Wörter, die nicht im Korpus enthalten waren, trotzdem Vektoren errechnen.

Dann hat er über sich annähernde Lineare Abbildungen gesprochen, bei der verschiedene Vektorräume verglichen werden. Es fällt auf, dass wenn man Objekte linear auf zwei Vektorräume abbildet, dass sie sich Raumübergreifend nah sind. Bsp.: „eins“ und „one“ würde man sehr nah beieinander finden. Um dies einfach zu bewerkstelligen verwendet er die lineare Regression und die L2 Regularisierung um Over Fitting vorzubeugen.

Anschließend hat er einige Korpora vorgestellt, beispielsweise TED Talks, EMEA. Am Schluss hat er über die noch zu erledigenden Schritte gesprochen. Er möchte versuchen noch bessere Abbildungen zu erhalten und bessere Regularisierungsmethoden zu finden. Evaluationen werden momentan mit den Top 5 der Wahrscheinlichkeiten oder nur mit der besten Wahrscheinlichkeit vorgenommen. „Fast Text“ soll mit unbekannten Wörtern getestet werden.