Kolloquium - Computerlinguistisches Arbeiten. Protokoll der Sitzung am 15.5.17

Tobias Eder „Exploiting bilingual word embeddings to establish translational equivalence “ Betreuer: Dr. Fraser

Am Anfang der Präsentation hat der Vortragende die Motivation der Wahl des Themas seiner Bachelorarbeit anhand eines Beispiels erläutert: Im medizinischen Bereich gibt es zu übersetzende Texte, die zahlreiche Begriffe enthalten, die eventuell keinen Eintrag in dem Wörterbuch haben. In diesem Fall bietet es sich an, eine „Übersetzung ohne Wörterbuch“ durchzuführen. Um das Prinzip dieses Ansatzes zu erklären, hat der Vortragende über Word Embeddings berichtet.

Im Gegensatz zu sehr „dichten“ Audio- bzw. Image-Daten, sind die Textdaten „spärlich“ (sparse). Um das Problem teilweise zu lösen werden Word Embeddings eingesetzt. Das sind Präsentationen von Wörtern aus Texten in einem hochdimensionalen Vektorraum, wobei es verschiedene Möglichkeiten gibt, die Anzahl an Dimensionen zu reduzieren, um das Model anschaulicher zu machen. Es gibt Annahmen dazu, die wichtigste von denen ist die s.g. Distributionshypothese, laut der eine Abbildung von einem bestimmten Wort in einem Vektorraum zu anderen Vektoren jeweils semantische und, was noch wichtiger scheint, syntaktische Zusammenhänge aufrechterhält. In solchem Vektorraum lassen sich verschiedene Konsistenzen beobachten, z.B. Beziehungen zwischen Ländern und Hauptstädten. Man kann feststellen, dass semantisch ähnliche Wörter Clustern bilden.

Es folgte eine Anmerkung von Herrn Prof.Schulze: Ob man das Sparseness-Problem auf diese Weise löst sei fraglich - extrem viele Wörter können ja nur einmal im Korpus auftreten, weswegen der Vektorraum nicht genügend statistische Evidenz hat. Wenn es außerdem „falsche“ Sätze (z.B. „Rom ist die Hauptstadt von Spanien“) in dem Trainingskorpus zu finden sind, kann es sogar zu falschen Ergebnissen führen.

Danach hat der Vortragende 2 Vektorraummodelle präsentiert - *word2vec*

und *fastText*. Das erste hat 2 Untermodelle - *Continuous Bag of Words* und *Skipgram*. Das CBOW-Modell nimmt einen bestimmten Kontext und versucht vorherzusagen, welches Wort an der Stelle fehlen könnte. Das *Skipgram* macht genau das Gegenteil: es nimmt ein bestimmtes Wort, um zu sagen, in welchem Kontext das Wort auftritt. Das *fastText-Modell* benutzt zusätzlich sogenannte „Subword-Information“ (Buchstaben n-Gramme), was insbesondere dazu dient, morphologische Differenzen zwischen Wörtern in Betracht zu ziehen. Es ist außerdem möglich, für Wörter, die im Trainingskorpus nicht vorkamen, Vektoren auszurechnen.

Anhand eines Beispiels mit 2 Vektorräumen (Englisch und Spanisch) wurde nochmals betont, dass semantische Beziehungen zwischen Wörtern in einer Sprache (einem Korpus) sich auch in dem anderen Korpus auffinden lassen. Wenn man lineare Abbildungen (oder wie von Herrn Schulze gemerkt, eher „Annäherungen“ an lineare Abbildungen) zwischen den Räumen findet, befinden sich semantisch ähnliche Wörter relativ nah voneinander. Die Abbildungen werden mithilfe linearer Regression mit L-2 Regularisierung (Bestrafung großer Gewichte) gemacht.

Es werden 4 unterschiedliche parallele Korpora verwendet (Englisch - Deutsch): generelles Korpus (z.B. Texte von Wikipedia), medizinisches, pharmazeitisches (EMEA) Korpus und ein kleines TED-Talks Korpus. Dabei sind ca. 5000 Wörter maschinell übersetzt (davor stichprobenartig auf Richtigkeit geprüft). Es wird versucht, mithilfe der oben beschriebenen Modelle ca. 1000 hochfrequente Wörter (je ein Korpus), die nicht in den parallelen Korpora auftreten, zu übersetzen. Dabei wird die Performance der Modelle verglichen.

Abschließend hat der Vortragende kurz über weitere Schritte (Übersetzung niedrigfrequenter Wörter, bessere Abbildungen, andere Regularisierungsmethoden, Evaluation auf Out-Of-Vocabulary-Wörtern) berichtet, die er in seiner Arbeit noch realisieren möchte. Es wurde auch auf weiterführende Literatur hingewiesen, insbesondere auf *Linguistic Regularities in Continuous Space Word Representations* von T. Mikolov.