Vortrag 1 von Tobias Eder (15.05.2017):
Exploiting bilingual word embeddings to establish translational equivalence

Am Anfang hat Tobias die Gliederung zu seinem Thema gezeigt und ist als erstes auf die Erklärung des Themas und die Motivation eingegangen. Konkret ging es bei der Arbeit darum, dass man z.B. gerne eine Übersetzung auch ohne ein Wörterbuch bzw. eine Übersetzung einer Domäne ohne Wörterbuch aber mit einer vergleichbaren Sprache durch einen Abgleich der Wörter erreichen könnte.
Dann ist er auf Bilingual Word Embeddings eingegangen und hat erklärt, dass diese sehr verdichtete Datensätze beinhalten und wie er diese Daten nutzt.
Wenn der Kontext von einem bestimmten Wort in einem Text ähnlich ist zu einem anderen Wort, dann könnte es durchaus auch sein, dass dieses Wort die selbe Bedeutung hat. Danach ist er noch auf Vektorraummodelle eingegangen und hat gezeigt wie er sie in seiner Arbeit benutzt.
Er hat erklärt, dass man in Vektorraummodellen diese Wörter sowohl syntaktisch als auch semantisch abbilden kann. Vom Professor wurde hierzu noch angemerkt, dass bei ganz seltenen Wörtern im Korpus das sparseness Problem durch Vektorraummodelle leider nicht besonders gut gelöst wird. Es ist aber sehr gut um hypothetische Aussagen zu entwickeln.
Tools zum Arbeiten mit Vektorraummodellen sind z.B. Word2vec von Google und fastText von Facebook Research.
Für die linearen Abbildungen werden zwischen 100-1000 Dimensionen im Vektorraum verwendet.
Die Abbildungen werden durch lineare Regression gemacht, dazu wird auch die Ridge-Regression genutzt.
Als Korpora für die Experimente werden 4 unterschiedliche Korpora genutzt:
General (ca. 110M tokens)
Medical big (ca. 50M tokens)
EMEA (ca. 4M tokens)
TED talks (ca. 2M tokens)
Weiter wurden unterschiedliche Embeddings wie z.B. CBOW und Skigram verwendet.
Bei den Experiment wurden 1000 hochfrequente Wörter aus den Korpora, sowie domänenspezifische Testsets verwendet.
Die Ergebnisse der Experimente zeigten sehr unterschiedliche Genauigkeit und Performance der Modelle.
Für die Verbesserung der Experimente könnte man z.B. niedrigfrequentige Wörter mitbenutzen, und prüfen ob man noch bessere Abbildungen bekommt. Dazu könnte man noch andere Regularisierungsmethoden verwenden und die Evaluation mit OOV Wörtern überprüfen.