SOSE 2017,CIS,LMU,München

Protokoll zur Sitzung am 15.05.2017:

Präsentation von Tobias Eder:

Tobias's Thema lautet als "Exploiting bilingual word embeddings to establish translational equivalence", seine Bachelorarbeit wird von Dr. Alexander M. Fraser betreut. Am Anfang der Präsentation hat er die Hauptpunkten seiner Motivation erläutert nämlich die Wichtigkeit der Übersetzung ohne Wörterbüchern, Übersetzung auf spezifische Domains und Behandlung mit den unbekannten Wörtern. Dann hat er den Begriff „Word Embeddings" definiert, welcher stellt die Wörter in vektoralem Modell dar, dazu gabt es einige Abbildungen mit unterschiedlicher Dichtigkeit von Wörtern in diesem Modell. Vektorraummodele wurden mit der Hilfe von zwei Toolkits repräsentiert: Word2Vec mit Google(2013), Word-Embedding Toolkit und CBOW Programm Model. FastTex ist sehr identisch zu dem ersten aber er betrachtet die Wörter nicht nur als atomare Einheiten, sondern auch die Wörter aus selben Stammen. Dazu hat er noch Sparsity Problem erwähnt, welches bis jetzt noch nicht lösbar ist aber damit kann man es reduzieren lassen. Als nächster Punkt seiner Präsentation hat er lineare Abbildungen vorgestellt auf dem Beispiel von englischen und spanischen Sprachen, diese Abbildungen wurden durch lineare Regression gemacht.

Für die Arbeit wurde vier unterschiedliche Parallelkorpora verwendet: General (Wikipedia), Medical BG, EMEA (Pharmazie mit chemischen Formeln), TED (gesprochene, transkribierte Texte) . Die Evaluierung

richtet auf die 1000 hochfrequenten Wörter, danach wird man das gleiche mit den 1000 seltensten Wörtern machen.

Für das Experiment war einen kleinen parallelen Korpus mit der Hilfe von Moses Toolkit erstellt, der enthält ungefähr 5000 Wörter.