Vortrag 1: Tobias Eder „Exploiting bilingual word embeddings to establish translational equivalence“. Betreuer: Dr. Fraser

Tobias begann mit der Wichtigkeit seines Themas für die maschinelle Übersetzung. Bei der maschinelle Übersetzung ist man oft auf die Wörterbücher und spezifische Domänen angewiesen, daraus entstehen Probleme bei der Behandlung von unbekannten Wörtern. Mit der Hilfe von Vektormodellen (Word2Vec und fastText), die die Word Embeddings repräsentieren, will man versuchen unabhängig davon zu sein. Dazu kommt noch Sparsity Problem, das obwohl bis jetzt nicht lösbar ist, lässt sich aber dadurch mildern.

Man geht davon aus, dass wenn der Wörterkontext ähnlich ist, stehen die Wörter in Verbindung zu einander. Es betrifft sowohl syntaktische, als auch semantische Ähnlichkeit. Die Daten bilden Klaster. Je kleiner die Distanz zwischen Wörter, desto grösser ist die Ähnlichkeit.

Weiterhin stellte Tobias die Vektoren vor. Word2Vec Modellist von Google zur Verfügung gestellt und fasstText von Facebook.

Die nehmen Texte als Input und generieren Vektoren. Als Ergebnis könnte man lineare Abbildungen kriegen, um zu sehen wie nah die Wörter zweier Sprachen zueinander liegen und somit zur Übersetzung genutzt werden. Tobias hat als Beispiel zwei lineare Abbildungen von Englischen und Spanischen Wörter vorgestellt.

Für die Bachelorarbeit werden vier Parallelekorpora benutzt: General(Wikipedia), Medical BG, EMEA (Pharmazie mit vielen chemischen Formeln), TED (gesprochene, transkribierte Texte).

Für den Experiment wurde einen kleinen parallelen Koprus mit Hilfe von Moses Toolkit erstellt. Der umfasst circa 5000 Wörter.

Die Evaluierung erfolgt auf die 1000 hochfrequenten Wörter. Später will man das gleiche mit den 1000 seltensten Wörter machen und out of vocabulary words.

Wichtig sind für die Übersetzung auch die Regularisierungsmethoden, vor allem wurde festgestellt, dass die Jahreszahlen falsch übersetzt wurde.