Referat Tobias Eder „Exploiting Bilingual Word Embeddings to Establish Translational Equivalence”

Die Arbeit wird von Fabienne Braune und Dr. Alexander Fraser betreut.

Das Ziel von Tobias‘ Arbeit ist eine Übersetzung ohne Wörterbuch zu ermöglichen. Diese wird z.B. benötigt wenn seltene, domainabhängige Wörter übersetzt werden sollten oder einfach ein Word nicht im Wörterbuch vorhanden ist. Anders als Audio oder Bilddateien sind Textrepräsentationen sehr sparse. Dies versuchen Word Embeddings zu lösen. Die Idee beruft sich auf die Distributionshypothese, die besagt dass die Bedeutung eines Wortes seine Distribution, d.h. seinem Kontext entspricht. Daher repräsentieren Word Embeddings als Vektor im hochdimensionalen Vektorraum aller Kontextwörter. Mit der cosine similarity zwischen zwei Vektoren kann man in diesem Raum die semantische Ähnlichkeit zweier Wörter berechnen. Hierfür werden zwei verschiedene Toolkits vorgestellt. Einmal existiert Word2Vec, welches 2013 von Google Research vorgestellt wurde, dass die CBOW und Skipgram Modelle implementiert. Außerdem wird fastText vorgestellt, dass 2016 von Facebook Research herausgebracht wurde. Die Besonderheit hier ist dass die word embeddings auch mit Subwortinformation, also n-Grammen gelernt wird. Die Verbindung zu Übersetzung ohne Wörterbuch ist, dass Annäherungen an lineare Abbildungen gelernt werden können, die den Vektorraum einer Sprache auf den einer anderen Sprache abbilden, und dabei die Bedeutung der Wörter erhalten. Die linearen Abbildungen werden mit linearer Regression gelernt, wobei zusätzliche L2 Regularisierung benutzt wird, um große Gewichte zu bestrafen.

In seiner Bachelorarbeit benutzt Tobias den folgenden Experimentaufbau: er hat vier unterschiedliche parallele Korpora:

- General (ca. 110M Tokens)
- Medical Big (ca. 50M Tokens)
- EMEA (ca. 4m Tokens)
- TED Talks (ca. 2M Tokens)

Darauf trainiert er seine Word embeddings, und testet sie auf einem kleinen, selbst erstellten parallelen Korpus für deutsch Englisch, der ca. 5000 Wörter umfasst. Außerdem will er noch mit domänen-spezifischen Sets testen, um die Performance unterschiedlicher Modelle vergleichen. Als weitere Schritte möchte er u.a. versuchen für niedrig frequente Wörter bessere Lösungen zu finden und nach anderen Regularisierungsmethoden suchen.