Kolloquium, 15. Mai 2017

Dayyan Smith

Exploiting Bilingual Word Embeddings to Establish Translational Equivalence

Vortragender: Tobias Eder

Betreuerin: Fabienne Braune

Tobias Eders Bachelorarbeit beschäftigt sich mit der Frage, ob es möglich ist (in einer bestimmten Domain) auch ohne Wörterbuch zu übersetzen. Dies wäre hilfreich, da nicht für jede Domain und jedes Sprachpaar ein Wörterbuch vorhanden ist.

Word Embeddings könnten helfen dieses Problem zu lösen: Dazu wird für jedes Wort ein Vektor erstellt, der die Kontexte in welchem das Wort vorkommt kodiert. Wörter mit syntaktischen bzw. semantischen Ähnlichkeiten haben ähnliche Vektoren. Zwei beliebte Toolkits zum Erstellen von Word Embeddings sind Word2Vec von Google und fastText von Facebook Research. Sowohl Word2Vec als auch fastText stellen zwei Modelle zum Lernen von Word Embeddings zur Verfügung. Die Embeddings werden gelernt, indem ein "sliding window" über den Text des Korpus' geschoben wird. Dabei wird entweder das Fokuswort von seinen Kontextwörtern gelernt (continues bag of words) oder die Kontextwörter vom Fokuswort (Skipgram). Mit fastText ist es möglich auch subword information (beispielsweise Buchstaben n-Gramme) mit einzubeziehen.

Nachdem für beide Sprachen Word Embeddings erstellt wurden, werden mit linearer Regression lineare Abbildungen berechnet, die vom Vektorraum einer Sprache in den der anderen abbilden. So kann dann der Vektor eines Wortes der Ausgangssprache in den Vektorraum der Zielsprache abgebildet werden. Im Idealfall würde so dann direkt der Vektor der Übersetzung in der Zielsprache gefunden werden, aber in der Realität wird der Vektor ausgewählt, der dem Berechneten am nächsten ist.

Für vier parallele Deutsch/Englisch Korpora, einen General Korpus mit ca. 110M Tokens, den Medical Big Korpus mit ca 50M Tokens, den EMEA (European Medicines Agency) Korpus mit ca. 4M Tokens und den Ted Talks Korpus, mit ca 2M Tokens werden Word Embeddings mit dem CBOW und dem Skipgram Modell trainiert.

Evaluiert wird, wie gut in den verschiedenen Domains mit linearen Abbildungen von dem Vektorraum einer in den einer anderen Sprache abgebildet werden kann, und was für ein Unterschied es macht, wie die Embeddings trainiert werden (CBOW, Skipgram)

Die nächsten Schritte sind, bessere Abbildungen zu finden und andere Regularisierungsmethoden auszuprobieren.