PROTOKOLLE ZU COMPUTERLINGUISTISCHES ARBEITEN

Ines Röhrer
Centre for Information and Speech Processing, LMU
I.Roehrer@campus.lmu.de

1 Referat

Das erste Referat des Tages hält Tobias Eder, in seiner Bachelorarbeit (”Exploiting Bilingual Word Embeddings to establish translational Equivalence”), betreut von Fabienne Braune und Alexander Fraser, beschäftigt er sich mit dem Thema biliguale Wordembeddings. Das Ziel der Arbeit ist es, ein Übersetzen ohne Wörterbuch zu ermöglichen. Das ist u.a. dann nützlich, wenn man in einem Spezialbereich arbeitet und zu der Domain kein Wörterbuch existiert, wie zum Beispiel im medizinischen Bereich.

Das Referat ist gegliedert in die Vorstellung von Word Embeddings, Vektorbaummodellen, Linerarer Abbildungen und Korpora und dem Experimentaufbau.

Word Embeddings können helfen, das Sparse Problem einzudämmen (auch wenn das Sparse Problem hierdurch nicht beseitigt wird). Da Wörter atomare Ereignisse in einem Korpus sind, gibt es neben deren spezifischen Kontext sehr wenig zusätzliche Daten. Die Abbildung eines Wortes im Vektorraum hingehen hat syntaktische oder semantische Beziehungen zu anderen Wörtern, diese Relationen dienen als zusätzliche Informationen zum Ursprungswort. Aus der größe des Abstandes zwischen zwei Wörtern kann man erkennen, wie stark verwandt beide Wörter sind.

In Tobias' Regularisierung werden große Gewichte bestraft, um ein besseres/glatteres Ergebnis zu erreichen. Er verwendet vier verschiedene Korpora für seine Experimente (auf Englisch und Deutsch) unterschiedlicher Länge und mit verschiedenen Embeddings. Er verwendet einen großen ”General” Korpus, einen medizinischen ”Medical” Korpus, den EMEA Korpus (auch medizinisch) sowie die TED Talks (wobei letztere eher gesprochene Sprache beinhaltet).

Für die Evaluation wird eine Auswahl von 1000 hochfrequenten Wörtern verwendet, sowie deren Abbildung in ein Regressionsmodell.

Als Nächstes sollen niedrigfrequente Wörter untersucht werden, sowie bessere Abbildungen und eine andere Regularisierung.

Zum Schluss werden noch einige Paper vorgestellt, zu ähnlichen Arbeiten oder Informationsquellen.