CIS, LMU München

Studentin: Elena Atanasova

15.05.2017

Protokoll zur Sitzung vom 15.05.2017 – Computerlinguistisches Arbeiten

1. Vortrag: Tobias Eder , "Exploiting Bilingual Word Embeddings to Establish Translational Equivalence"

BA-Betreuer: Dr. Alexander M. Fraser

Den ersten Vortrag des Tages hält Tobias Eder, der von Dr. Fraser betreut wird. Er begann mit einem Überblick und die Wichtigkeit seines Themas für die machinelle Übersetzung und stellte seine Ideen für die Gliederung seiner Bachelorarbeit vor.

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. Je kleiner die Distanz zwischen Wörter, desto grösser ist die Ähnlichkeit.

Weiterhin stellte Tobias die Vektoren vor. Word2Vec Modellist von Google(2013) zur Verfügung gestellt und fastText von Facebook Research(2016).

Die nehmen Wörter von Texten als Input und machen Vektoren damit. Als Ergebnis bekommt man lineare Abbildungen, 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.

Als nächstes hat Eder die Korpora und der Evaluierung vorgestellt.

Für seine Bachelorarbeit werden vier unterschiedliche parallele Korpora benutzt, und zwar: General(ca. 110M Tokens), Medical Big (ca. 50M Tokens), EMEA (ca. 4M Tokens) und TED Talks (ca. 2M Tokens), auch unterschiedliche Embeddings (CBOW, Skipgram).

Für den Experiment wurde einen kleinen parallelen Koprus mit Hilfe von Moses Toolkit erstellt, der circa 5000 Wörter umfasst. Die Evaluierung erfolgt auf die 1000 hochfrequenten Wörter. Später will man das gleiche mit den 1000 niedrigfrequente Wörter machen und auf OOV(out of vocabulary)-Wörtern in fastText. Für die Übersetzung sind auch die bessere Abbldungen und die Regularisierungsmethoden, vor allem wurde festgestellt, dass die Jahreszahlen falsch übersetzt wurden.