EXPLOITING BILINGUAL WORD EMBEDDINGS TO ESTABLISH TRANSLATIONAL EQUIVALENCE

Anton Serjogin
Centre for Information and Speech Processing, LMU
anton.serjogin@gmail.com

15.05.2017

Motivation: übersetzung auf bestimmter Domain und unbekannte Wörter im Text. Es gibt verschiedene Vektorraummodelle, bei denen Semantik und Syntax betrachtet werden können. Die Vektorraummodelle sind hochdimensional, ein großer Nachteil ist aber - wenn einen falschen Satz im Trainingsset gibt, dann sind falsche Ergebnisse geliefert. Word2Vec und FastText sind beide Vektormodelle, der Unterschied ist, daß der Prinzip vom Lernen anders ist. Eine andere Art von Modellen sind Lineare Abbildungen, die produktiv und nicht dimensional sind. Insgesamt war 4 Korpora genommen: General, Medical Big, EMEA, TED Talks. Die Idee: Auswahl an Wörtern aus Korpus und Domänspezifische Testsets erzeugen, damit später die Performances von unterschiedlichen Modellen zu überprüfen.