Jakob Sharab

Predicting New Domain Senses in English Medical Texts

Betreuer: Fabienne Braune

Jakobs Arbeit beschäftigt sich mit dem Problem der Erkennung des domänespezifischen Sinn eines Wortes aus dem Kontext in welchem es vorkommt. Das Problem tritt beispielsweise für das englische Wort „administration“ in den beiden Kontexten Verwaltung und Medizin auf. Damit wird die Domänendesambiguierung zu einem zentralen Problem von maschineller Übersetzung. Ein Ansatz zur Bewältung dieser Problematik ist das sogenannte „Sense Spotting“ – das Finden von Features in einem Text die eine Bedeutungsveränderung von Worten anzeigen. Eine Art dieser Features sind die Topic Model Features, mit denen sich die Arbeit näher beschäftigt hat.

Bei Topic Model Features handelt es sich um das Clustering von Themengebieten innerhalb großer Textkorpora durch die Analyse einzelner Wörter im Text. Dies kann ohne vorherige Annotation erfolgen und ist deswegen ideal zur Verarbeitung größerer Textmengen. Es basiert auf Latent Dirichlet Allocation (LDA) einem generativen Modell zur Untergliederung von Texten in Sub-topics. Dabei ist der zentrale Ansatz, dass ein Dokument aus einer zufälligen Verteilung latenter Topics besteht. Ein Topic besitzt eine Verteilung über Wörter, welche als Bag-of-Words Modell repräsentiert sind.

Das Topic-Wort-Modell nimmt innerhalb dieser angenommenen latenten Topics die Veränderung der Häufigkeit eines Wortes als Anzeichen für einen Domänenwechsel. Ziel der Arbeit ist es, unterschiedliche Ähnlichkeitsmaße für die Analyse von Bedeutungsveränderung zu finden.

Zu den Untersuchten Möglichkeiten zählen:

(1) Cosinus Ähnlichkeit
(2) Relative Entropie
(3) Zählen von Worthäufigkeiten im Kontext des Wortes

Hierzu wurden zwei Korpora verwendet: Aus der Nachrichtendomäne des WMT 2016 Shared Task, und das EMEA Medizin-Korpus. Beide Dokumente wurden Tokenisiert und Stoppwörter entfernt. Danach wurden die Dokumente unter Verwendung des Python Toolkits gensim in latente Topics mittels LDA geclustert. Es wurden gezielt diese Wörter untersucht, deren Bedeutung sich zwischen beiden Domänen ändern sollte.

In der Analyse stellten sich mehrere Probleme: Zum einen erwies es sich als schwierig genügend Wörter zu finden, deren Bedeutung sich zwischen den Domänen unterscheiden werden um eine

quantitative Analyse durchführbar zu machen. Gleichzeitig ist es schwierig eine klare Grenze für Bedeutungen ohne Klassifizierungsprozess zu finden, welche die Ergebnisse objektiv qualitativ überprüfbar machen würden. Deswegen wurde nur eine quantitative Analyse auf den Daten durchgeführt. Hierbei ergab sich, dass die Kosinus-Distanz und die Relative Entropie beide gute Maße zur Unterscheidung von Bedeutung sind, während reine Kontexthäufigkeit deutlich schlechtere Ergebnisse erzielt.