Referat Jakob Sharab (Montag, den 12.06.2017)

Predicting New Domain Senses in English Medical Texts

Zuerst einmal wurde über die Motivation des Themes gesprochen.
Wörter haben in verschiedenen Domänden eine andere Bedeutung. Beispielsweise das Wort „administration“ In der Allgemeinen Bedeutung entspricht das die „Verwaltung“, welches jedoch in der Medizin die Bedeutung „Verabreichung“ eines Medikaments trägt.

Auch enstehen Fehler bei der Statisitcal Machine Translation (SMT)
Zum Beispiel das Bilden von Wortpaaren. „administration“, welches das Wort aus der ursprünglichen Sprache ist + „Verwaltung“ dessen Übersetzung.
Bei Anwendungen in einer neuen Domäne ist das oben gebildete Wortpaar nicht mehr korrekt. Beispielsweise in der Medizin hat es eine andere Bedeutung, wie wir es vorher erklärt bekommen haben.

Somit bilden wir eine neue Definition für eine Task „Sense Spotting“.
Wir wollen features finden, die Bedeutungsveränderung indizieren. Zudem wollen wir ein Classifier mit diesen features antrainieren.
Uns wurde erklärt, dass eines dieser Features das Topic Model Feature ist.

Das Ziel des Topic Medeling, ist es in großen Textkorpora darin enthaltene Topics zu finden. Diese werden mit Hilfe von Algorithmen erledigt um die einzelne Wörter in den Dokumenten zu analysieren.
Da keine vorhergehende Annotation der Daten nötig ist, ist das ein großer Vorteil.
Welches Nutzen ziehen wir hier raus? Zum einen übersteigt die Menge von Daten das menschliche Kapazität heutzutage. Desweiteren werden große Textarchiven organisiert.

Als nächstes wurde über die Generative Modelle gesprochen.
Hier haben wir ein Klassifizierungsproblem. Beispielsweise möchte zwischen einem Elefanten und einem Hund unterscheiden.
Der diskriminativer Ansatz dafür wäre, das Trainieren eines Klassifikators mit Features, der eine „Linie“ findet, die die Klassen voneinander trennen.
Das heisst je nachdem auf welche Seite der Linie die Werte fallen, wird das Tier als Elefant oder als Hund klassifiziert.

Dann hätten wir noch den generativen Ansatz.
Hier werden zwei Modelle gebaut, diese anaylsieren wie ein Elefant und wie ein Hund ausschaut.

Zuletzt wurde über die Probleme und Ergebnisse gesprochen.
Zum einem war es nicht einfach viele Beispiele für Wörter zu finden, deren Bedeutung sich in der neuen Domäne verändert.
Auch ist es schwierig ohne einen Classifier eine Decision Boundary zu finden. Somit konnten die Ergebnisse nur quantitaiv miteinander verglichen werden.

Und zum Ergebniss haben wir herausgefunden, das die Relative Entrotopie und das Mass aufgrund der gleichen Wörter die besten Resultate erzielen.