CIS, LMU München

Studentin: Elena Atanasova

12.06.2017

Protokoll zur Sitzung vom 12.06.2017 – Computerlinguistisches Arbeiten

2. Vortrag: Jakob Sharab, "Predicting new domain senses in english medical texts"

BA-Betreuer: Dr. Fabienne Braune

Den zweiten Vortrag des heutigen Tages hält Jakob Sharab, die von Herr Braune betreut wird. Am Anfang der Präsentation stellte er seine Ideen für die Gliederung seiner Bachelorarbeit vor und danach präsentierte er die Motivation. Die Motivation ist, dass die Wörter in verschiedenen Domänen eine andere Bedeutung haben. z.B. das Wort „administration". Allgemein heißt es Verwaltung und in der Medizin bedeutet es „Verarbeitung" (eines Medikamentes). Fehler bei Statistical Machine Translation System (SMT). Bilden von Wortpaaren mit „administration" (Wort aus der ursprunglichen Sprache + „Verwaltung" dessen Übersetzung). Bei Anwendung in neuer Domäne ist das oben gebildete Wortpaar nicht mehr korrekt, deshalb definiert er eines neues Task „Sense Spotting". Es findet von Features die Bedeutungsveränderung indizieren und trainieren eines Classifiers mit Hilfe dieser Features. Eines dieser Features ist das Topic Model Features.

Das Ziel von Toping Modeling ist es, in großen Textkorpora darin enthaltene Topics zu finden und mit Hilfe von Algorithmen, die einzelne Wörter in den Dokumenten analysieren. Damit wird keine vorhergehende Annotation der Daten nötig, was ein Vorteil ist. Die Menge von Daten übersteigt heutzutage menschliche Kapazitäten und dafür braucht man organisieren von großen Textarchiven. Dazu gehört auch die LDA (Latent Dirichlet Allocation).

Klasifizierungsproblem ist: man möchte zwischen einem Hund und einem Elefanten unterscheiden. Dafür gibt es zwei Ansätze: diskrimitiver und generativer. Diskrimitiver Ansatz ist traineren eines Klassifikators mit Features, der eine Linie findet, die die Klassen voneinander trennen und der generative Ansatz ist bauen zweier Modelle, die analysieren wie ein Hund und wie Elefant ausschaut, übergeben das zu klassifizierenden Tieres an die Modelle und berechnen der Wahrscheinlichkeit p(x|y) für jede Klasse: y(0) = Elefant und y(1)=Hund.

LDA ist ein generatives Wahrscheinlichkeits-Modell, das in dieser Arbeit benutzt wird, um Dokumente in einzelne Topics zu untergliedern. Die Idee dahinter ist es, dass jedes Dokument aus einer zufälligen Mischunglatenter Topics besteht. Jeides Topic seinerseits besteht aus einer Verteiliung über Wörter. Dieses Modell basiert auf der Annahme der „bag-of-words".

Danach ist Sharab wieder zu Topic Model Feature zurückgekommen. Dieses Feature nimmt die Änderung der Häufigkeit eines Wortes innerhalb eines Topics, beim Wechsel in die neue Domäne und interpretiert sie als Indikator für eine Bedeutungsänderung.

Sharab verwendet die „Kosinus-Ähnlichkeit" zum Messen der Ähnlichkeit zwischen Topics. Diese Bachelorarbeit hat dann als Ziel, verschiedene Ähnlichkeitsmaße miteinander zu vergleichen. Andere verwendete Messer für die Ähnlichkeitsmaße sind die „relative Entropie" und die „Ähnlichkeit aufgrund der Anzahl gleicher Wörter".

Sharab hat bisschen mehr über die Kosinus-Ähnlichkeit erklärt. Er nimmt zwei Vektoren und berechnet den zwischen diesen eingeschlossenen Winkel. Danach gibt Sharab einen Wort zwischen 0 und 1 aus und je höher der Wert, desto ähnlicher sind sich zwei Vektoren.

Die Relative Entropie berechnet den Abstand zwischen zwei Wahrscheinlichkeitsverteilungen. Je höher der Wert, desto weiter auseinander sind zwei Verteilungen.

Ähnlichkeit aufgrund der Anzahl gleicher Wörter bedeutet: je mehr gleiche Wörter unter den Top n Wörtern zweier Topics sind, desto ähnlichersind sie sich. Für diese Arbeit wurden zwei parallele Korpora auf Englisch und Deutsch verwendet.

Es wurden Daten für die medizinische Domäne wie zum Beispiel EMEA Korpus verwendet. Für die Nachrichten Domäne wurden ein general Korpus verwendet, das im WMT Shared Task 2016 verwendet wurde.

Sharab hat mit einige Probleme gestoßen. Es war für ihn nicht einfach, viele Beispiele für Wörter zu finden, deren Bedeutung sich in der neuen Domäne verändert, da diese Wörter immer die Bedingung erfüllen müssten, in einem Topic eine hohe Wahrscheinlichkeit zu haben.

Da es schwierig ist, ohne einen Klassifikator eine Decision Boundary zu finden, konnten die Ergebnisse nur quantitativ miteinander verglichen werden. Dabei haben die Relative Entropie und das Maß aufgrund der gleichen Wörter die besten Resultate erzielt.