Klyueva Olga
CIS, LMU München
12.06.1917

Jakob Sharab: Predicting new domain senses in English medical text

Die Motivation für die Bachelorarbeit besteht darin, dass die Wörter in verschiedenen Domänen eine andere Bedeutung bekommen. Beispielweise während das Wort „administration" generell als Verwaltung auf Deutsch übersetzt wird, wird es im medizinischen Bereich als „Verabreichung" benützt. Wegen dieses Problems wurde ein neuer Task „Sense Spotting" definiert. Mit Hilfe von Sense Spotting wurden die Wörter, die in einer neuen Domain ihre Bedeutung wechseln, erkannt. Dafür wird ein Klassifier mit Features definiert, die die Bedeutungsveränderungen des Wortes indizieren. Der Klassifier kann klassifizieren, ob das Wort in einer neuen Domain eine Bedeutung wechselt oder nicht.
In der Bachelorarbeit geht es um eine von den Features: Topic Modelling Feature.
Das Ziel von Topic Modelling ist, innerhalb eines größeren Textkorpus enthaltene Topic/Themen zu finden. Es wird mit Hilfe von statistischen Algorithmen gemacht. Die einzelnen Wörter werden in den Dokumenten analysiert. Der Vorteil des Topic Modelling ist, dass keine Annotation vorher nötig ist. Die Motivation dafür ist, die großen Archive von Dokumenten zu strukturieren.
Man will einen Elefant von einem Hund unterscheiden. Anhand dieses Tasks wurde der Unterschied zwischen generativen und discriminativen Modellen gezeigt. Mit dem ersten Modell wird ein Klassifier mit verschiedenen Features trainiert. Es gibt Decision Boundary, die verschiedene Klassen voneinander trennt. Je nachdem auf welcher Seite die Decision Boundary am Ende ist, wird das Tier der einen Klasse oder der anderen zugeordnet.
Bei dem generativen Modell wird geschaut aus welchen Bestandteilen die zwei Tiere bestehen. Es wird Features Verteilung für die jeweilige Klasse berechnet und festgestellt, zu welcher Klasse es zugeordnet wird.
Als nächstes geht es um Latent Dirichlet Allocation, eine generatives probabilistisches Modell. In der Bachelorarbeit wurde Latent Dirichlet Allocation benutzt, um die verschiedenen Dokumente in einzelne Topics zu untergliedern. Die Grundidee dafür ist, dass jedes Dokument aus einer Mischung von latenten (versteckten) Topics besteht, die wiederum aus einzelnen Wörtern bestehen. Es wurde auf Annahme von bag of words zugegriffen, wenn die Reihenfolge von Wörtern vernachlässigbar ist, um grob an den Wörtern zu erkennen, über welches Thema es in dem Text geht. Es kann auf die Dokumente innerhalb eines Korpus übertragen werden: dass die Reihenfolge von Dokumenten nicht wichtig ist und die Dokumenten austauschbar sind.
Latent Dirichlet Allocation Modell benötigt folgende Schritte. Als erstes wurde die Anzahl der Wörter festgelegt, die ein Topic enthält. Danach wurde eine Mischung an Topics, die ein Dokument enthält, festgelegt. Beispielweise besteht das Dokument zu 60 Prozent aus Medikamenten Topic und 30 Prozent aus Krankheiten Topic. Der nächste Schritt ist eine Generierung der Wörter. Man wählt das Topic aus dem das Wort stammt. Das Wort wurde mit Hilfe des Topics generiert.
Es wurde gezeigt welche Topics das Dokument enthält. Als Ergebnisse bekommt man für einen Topic die Wörter mit der Wahrscheinlichkeit.
Es wurde eine Formel präsentiert. In der Formel wurden folgende Ähnlichkeitsmassen definiert:
-Kosinus Ähnlichkeit, die man benutzt, um die Ähnlichkeit zwischen Topics zu messen.
-Relative Entropie, die den Abstand zwischen zwei Wahrscheinlichkeitsverteilungen misst. Je höher der Wert ist desto weiter sind zwei Verteilungen auseinander, desto unähnlicher sind die Topics.
-Die Ähnlichkeit aufgrund der Anzahl gleicher Wörter.
Das Ziel der Bachelorarbeit ist verschiedene Ähnlichkeitsmassen zwischen verschiedenen Topics mit einander zu vergleichen.
Als erstes wurden die Korpora tokenisiert und Stopwörter entfernt. Mit Hilfe des Latent Dirichlet Allocation Modell wurden die Dokumente in 100 Topics unterteilt. Es wurde das Gensin Toolkit benutzt um die Wahrscheinlichkeit einzelner Wörter zu bekommen. Es wurde ein Wort ausgesucht, dessen Bedeutung sich mit einem neuen Topic ändert und eine hohe Wahrscheinlichkeit bekommen hat. Für die Kosinus Ähnlichkeit und Relative Entropie wurden alle Wörter aus alten Domänen, in denen das Wort enthalten ist, mit allen anderen Topics aus neuen Domain verglichen und die Wahrscheinlichkeit der gleichen Wörter rausgefiltert und als Input für die Kosinus Ähnlichkeit und Relative Entropie benutzt.

Für die Ähnlichkeit aufgrund der Anzahl gleicher Wörter wurden 2500 Wörter eines Topics angeschaut und festgestellt, wie viele Wörter gleich sind.
Das gleiche wurde gemacht für die Wörter, die die Bedeutung in eine neue Domain nicht ändern, um einen Vergleichswert zu haben.
Zum Schluss werden Ergebnisse erwähnt. Relative Entropie und die Ähnlichkeit aufgrund der Anzahl gleicher Wörter haben die besten Resultate geliefert.