Protokolle zur Sitzung vom 12.06.2017 - Computerlinguistisches Arbeiten

15.06.2017

Drei Studenten stellten in der Sitzung am 12.06.2017 mit Hilfe einer Präsentation mit dem Beamer ihre Bachelorarbeitsthemen vor.

2. Präsentation:

Referent: Jakob Sharab

Titel: Predicting New Domain Senses in English Medical Texts

Betreuer: Fabienne Braune

Als zweites hielt Jakob Sharab seine Präsentation. Er begann als erstes die Motivation hinter seinem Thema zu erklären. Und zwar gibt es Wörter die in verschiedenen Domänen andere Bedeutungen haben, wie z.B. das Wort „administration“, das allgemein eher „Verwaltung“, aber in der medizinischen Domäne „Verabreichung“. Durch solche Wörter kommt es bei Statistical Machine Learning Systems oft zu Fehlern, da diese Wortpaare bilden, nämlich einmal mit dem Wort selbst und mit dessen Übersetzung. Wenn ein solches System nun in einer neuen Domäne eingesetzt wird, stimmen die Wortpaare nicht mehr überein. Daher wurde ein neuer Task, namens „Sense Spotting“ definiert, welches mit Hilfe von verschiedenen Features helfen soll, solche Wörter zu erkennen. Danach erklärte Jakob Topic Models, deren Ziel es ist in großen Textkorpora enthaltene Topics zu finden. Der Vorteil dabei ist, dass keine vorherige Annotation der Daten nötig ist und es wird eingesetzt um große Datenmengen, wie Textarchive, zu organisieren und zu strukturieren.

Als nächstes erklärte er allgemein generative Modelle, an einem Klassifizierungsproblem, bei dem man zwischen einem Hund und einem Elefanten unterscheiden will. Dabei gibt es zwei Ansätze, wobei der erste der diskriminative Ansatz ist. Hier wird ein Klassifikator mit Features trainiert, der dann eine Decision Boundary findet, um die zwei Klassen voneinander zu trennen. Abhängig davon, auf welcher Seite der Wert fällt wird das Tier dem Elefanten oder dem Hund zugeordnet. Beim generativen Ansatz werden als erstes zwei Modelle gebaut, die analysieren aus welchen Bestandteilen das Tier besteht. Dann wird das zu klassifizierende Tier den beiden Modellen übergeben, die dann die Wahrscheinlichkeit $p(x|y)$ für jede Klasse berechnen.

Anschließend ging Jakob auf die Latent Dirichlet Allocation ein, welche auch ein generatives Modell ist und er in seiner Arbeit dafür benutzt hat Dokumente in Topics zu untergliedern. Die Grundidee hierbei ist, dass jedes Dokument aus Topics und diese wiederum aus einer Verteilung über Wörter besteht. Zudem basiert sie auf der „bag-of-words“-Annahme, die besagt, dass die Reihenfolge von Wörtern vernachlässigbar ist, was sich auch auf Dokumente in einem Korpus erweitern lässt. Zudem wird hier eine generative Entstehung von Dokumenten in verschiedenen Schritten angenommen. Als erstes wird die Anzahl der Wörter festgelegt, aus dem ein Topic besteht und danach die Mischung an Topics die in einem Dokument enthalten sind. Danach wird werden die Wörter generiert, wobei erst das Topic bestimmt wird aus dem das Wort stammt und danach wird das Wort selbst mit Hilfe des Topics generiert. Durch Umkehrung des generativen Prozesses, also der Fragestellung welche latente Struktur das vorliegende Dokument generiert hat, wird ein Dokument in Topics unterteilt.

Danach ging es um das Topic Model Feature, welches im Rahmen des „Sense Spotting“ Task definiert wurde. Diese geht davon aus, dass wenn sich die Häufigkeit eines Wortes innerhalb eines Topics

ändert, dies ein Indikator für eine Bedeutungsveränderung ist. Als Beispiel nahm er wieder das Wort „administration“, welches in der medizinischen Domäne mit der Bedeutung „Verabreichung“ auftaucht, was in der neuen Domäne nicht der Fall ist. Das kann man an der Verteilung des Wortes über die Topics sehen, da das Wort in der medizinischen Domäne in Topics, die Wörter wie „medicament“ oder „daily“ enthalten, eine hohe Wahrscheinlichkeit haben. In der neuen Domäne hat das Wort in einem ähnlichen Topic jedoch nur eine geringe Wahrscheinlichkeit. Dies zeigte er auch anhand einer Formel, die, wenn dessen Wert hoch ist, anzeigt, dass die Bedeutung gleichgeblieben ist und umgekehrt, wenn der Wert niedrig ist, dass sich die Bedeutung verändert hat. Hierbei wird die Kosinus-Ähnlichkeit verwendet, um die Ähnlichkeit zwischen Topics zu messen. Ziel seiner Arbeit war es nun verschiedene Ähnlichkeitsmaße miteinander zu vergleichen.

Dafür ging er auf die einzelnen Ähnlichkeitsmaße ein. Das erste war die Kosinus-Ähnlichkeit, die den Winkel zwischen zwei Vektoren berechnet. Je nachdem wie hoch der Wert ist, sind sich die Vektoren entsprechend ähnlich. Das zweite Maß war die Relative Entropie, die den Abstand zwischen zwei Verteilungen berechnet, wobei sich hier zwei Verteilungen ähnlich sind, wenn der Wert klein ist. Dies würde der Intuition hinter dem Topic Model Feature widersprechen, was Jakob dadurch löste, dass er die Relative Entropie durch 1 geteilt hat. Das dritte Maß, basierte auf der Anzahl der gleichen Wörter, wobei die Idee dahinter war, dass je mehr Wörter unter den Top n Wörtern gleich sind, desto ähnlicher sich zwei Topics sind.

Als nächstes stellte er seine Daten vor, welche er in seiner Arbeit verwendet hat. Dabei stammen die Daten aus der medizinischen Domäne aus dem EMEA Korpus und die aus der Nachrichten Domäne aus dem General Korpus.

Danach ging Jakob auf das Vorgehen in seiner Arbeit ein. Als erstes hat er die Daten tokenisiert und Stoppwörter entfernt. Danach hat er sich ein Wort ausgesucht, dass in einem Topic eine hohe Wahrscheinlichkeit hat und dessen Bedeutung sich verändert. Danach hat er für die Kosinus-Ähnlichkeit und die Entropie alle Wahrscheinlichkeiten aller gleichen Wörter zwischen dem Topic aus der alten Domäne und allen anderen Topics herausgefiltert und sie als Input für die zwei Ähnlichkeiten benutzt. Für das dritte Maß hat er die Anzahl der gleichen Wörter, unter den Top 2500 Wörtern, überprüft. Danach hat er die Formel für alle drei Ähnlichkeiten angewendet. Das gleiche hat er für Wörter gemacht die ihre Bedeutung nicht verändern, um einen Vergleichswert zu haben.

Als letztes ging Jakob auf die Probleme und Ergebnisse ein. Zu den Problemen gehörte z.B., dass es nicht einfach war genügend Wörter zu finden, die ihre Bedeutung verändern, da die Wörter immer die Bedingung erfüllen müssen, in einem Topic eine hohe Wahrscheinlichkeit zu haben. Das Ergebnis seiner Arbeit war, dass die Relative Entropie und das Maß aufgrund der Anzahl der gleichen Wörter die Besten Resultate ausgaben.