Katja Bertholdt
CIS, LMU München
12.06.17

Protokoll zur Sitzung vom 12.06.17 – Computerlinguistik Kolloquium

2. Vortrag: Jakob Sharab, „Predicting New Domain Senses in English Medical Texts“

Den zweiten Vortrag hält Jakob Sharab, in dem er seine Bachelorarbeit mit dem Thema „Predicting New Domain Senses in English Medical Texts“ vorstellt. Seine Arbeit wird von Fabienne Braune betreut. Die Motivation seiner Arbeit liegt darin, dass Wörter in unterschiedlichen Domänen verschiedene Bedeutungen haben können. Dieses Phänomen kann bei Statistical Machine Translation Systemen (SMT) zu Fehlern führen. So kann das Wort „administration“ im allgemeinen mit „Verwaltung“ übersetzt werden, jedoch ändert sich seine Bedeutung im medizinischen Kontext zu „Verarbreichung“ (von Medikamenten). Aufgrund dieses Problems wurde ein neues Task namens „Sense Spotting“ definiert mit der Aufgabe, Wörter zu erkennen, die in einer neuen Domäne ihre Bedeutung ändern. Dazu wurden Features festgelegt, die diese Bedeutungsänderung indizieren. Diese wurden dann einem Klassifier übergeben, der festlegt, ob das entsprechende Wort seine Bedeutung in einer anderen Domäne ändert oder nicht. In seiner Arbeit hat sich Sharab mit einem dieser Features beschäftigt: mit dem sog. „Topic Model Feature“.
Das Ziel des Topic Modellings ist es, innerhalb großer Textkorpora darin enthaltene Topics zu finden. Dies passiert mithilfe von statistischen Algorithmen, die einzelne Wörter in den Dokumenten analysieren. Ein Vorteil des Topic Modellings ist es, dass keine vorhergehende Annotation der Daten notwendig ist und es ist nützlich, um große Textmengen zu strukturieren und zu organisieren.
Sharab stellt das „Latenet Dirichlet Allocation“ (LDA) Modell vor, welches zu den generativen Wahrscheinlichkeitsmodellen gehört. In seiner Arbeit hat er das LDA verwendet, um Dokumente in einzelne Topics zu untergliedern. Die Grundidee besteht dabei darin, dass jedes Dokument aus einer Mischung von zufälligen latenten (versteckten) Topics besteht. Diese Topics bestehen wiederum aus einzelnen Wörtern. Dabei hat Sharab auf die Bag of Words Annahme zurückgegriffen, die besagt dass die Reihenfolge von Wörtern vernachlässigbar ist und Wörter gegenseitig austauschbar sind, denn trotz vertauschter Wörter kann ein Text einem bestimmten Themengebiet zugeordnet werden. In LDA Modellen wird die Annahme vertreten, dass die Erstellung eines Dokuments in mehreren Schritten abläuft. Zunächst wird die Anzahl der Wörter festgelegt, die ein Topic enthält. Danach wird eine Mischung an Topics festgesetzt, die ein Dokument enthält (bspw. 60% Topic „Medikamente“ und 30% Topic „Krankheiten“). Im letzten Schritt werden dann die Wörter generiert. Dazu wird zunächst das Topic ausgewählt, aus dem das Wort stammt. Anschließend wird das Wort mithilfe des Topics generiert. Um in einem Dokument die enthaltenen Topics zu erkennen, wird in Sharabs Arbeit eine Umkehrung dieses generativen Prozesses betrachtet. Es soll also anhand der Wörter auf ein Topic geschlossen werden.
Sharab geht nun genauer auf das Topic Modelling Feature ein, das im Rahmen des Sense Spotting Tasks definiert wird. Dabei wird davon ausgegangen, dass sich die Häufigkeit eines Wortes innerhalb eines Topics ändert, sobald es sich in einer neuen Domäne befindet. Dies wird als Indikator für eine Bedeutungsänderung gesehen. In einem Paper, auf das sich Sharab in seiner Arbeit bezieht, wurde diese Überlegung als Formel definiert. Deren Wert (und somit auch der Wert des Topic Modelling Features) ist hoch, wenn eine Bedeutungsänderung wahrscheinlich ist. In dieser Formel wurde die Kosinusähnlichkeit verwendet, um die Ähnlichkeit zwischen Topics zu berechnen. Sharab hat in seiner Arbeit verschiedene Ähnlichkeitsmaße herangezogen und verglichen. Er hat sich dabei mit der Kosinusähnlichkeit (berechnet den eingeschlossenen Winkel zwischen zwei Vekotren), der Relativen Entropie (berechnet den Abstand zwischen zwei Wahrscheinlichkeitsverteilungen) und der Ähnlichkeit aufgrund der Anzahl gleicher Wörter (Idee:

Je mehr gleiche Wörter unter den top n Wörtern zweier Topi sind, desto ähnlicher sind sie sich) beschäftigt.
Für die Daten seiner Arbeit verwendet er zwei parallele Korpora in Englisch und Deutsch. Die Daten für die medizinische Domäne stammen aus dem EMEA Korpus (Prüfberichte von Medikamenten von der Pharmabehörde EMEA) und die Daten für die Nachrichten Domäne aus dem General Korpus, der aus Daten verschiedener Korpora besteht.
Für seine Durchführung hat Sharab zunächst die Korpora tokenisiert und Stoppwörter entfernt, da sie keine Informationen über ein bestimmtes Topic liefern. Anschließend wurde mithilfe des LDA Modells die Dokumente in 100 verschiedene Topics aufgeteilt. Sharab hat dann jeweils ein Wort betrachtet, das seine Bedeutung in einem neuen Topic ändert und dem eine hohe Wahrscheinlichkeit in diesem Topic zugewiesen wurde. Daraufhin hat er alle gleichen Wörter aus dem Topic der alten Domäne herausgefiltert und mit allen anderen Topics der neuen Domäne verglichen. Diese Wahrscheinlichkeiten der gleichen Wörter wurde als Input für die Kosinusähnlichkeit und die Relative Entropie verwendet, um die Ähnlichkeit zu berechnen. Für die Ähnlichkeit aufgrund der Anzahl gleicher Wörter hat Sharab die häufigsten 2500 Wörter eines Topics gewählt, um deren Verteilung zu betrachten.
Für einen besseren Vergleich seiner Ergebnisse hat Sharab außerdem zusätzlich Werte von Wörtern berechnet, die ihre Bedeutung nicht ändern.
In seiner Arbeit ist Sharab auf das Problem gestoßen, dass nur wenige Wörter ihre Bedeutung zwischen Domänen ändern und zusätzlich eine hohe Wahrscheinlichkeit innerhalb eines bestimmten Topic haben.
Da Sharab keinen Klassifier zur Verfügung hatte, der anhand bestimmter Features besagt, ob ein Wort die Bedeutung geändert hat oder nicht, hat er in seinen Ergebnissen verglichen, inwieweit die Werte der Wörter, die ihre Bedeutung geändert haben, unterhalb der Werte der Wörter liegen, die ihre Bedeutung nicht geändert haben.