CIS, LMU München
Fach: Kolloquium Computerlinguistisches Arbeiten SS2017
Studentin: Faridis Alberteris Azar

Protokoll zur Sitzung am 12.06.17

Thema: „Predicting New Domain Senses in English Medical Texts“

Student: Jakob Sharab
Betreuer: Fabienne Braune

Einige Wörter haben in verschiedenen Domänen eine andere Bedeutung. Das führt zu Fehlern bei Statistical Machine Translation Systemen. Mit dieser Motivation wird für diese Bachelorarbeit ein neues Task mit dem Namen „Sense Spotting“ definiert. Hier werden Features gefunden, die die Bedeutungsänderung indizieren. Ein Klassifikator wird dann mit Hilfer dieser Features trainiert. Eines dieser Features ist das „Topic Model Feature“.

Das Ziel vom „Topic Modelling“ besteht darin, in großen Textkorpora die enthaltenen Topics zu identifizieren. Dafür werden mithilfe von Algorithmen, die einzelnen Wörter in den Dokumenten analysiert. Diese Methode hat als Vorteil, dass eine vorgehende Annotation der Daten nicht nötig ist, was heutzutage eine riesige Bedeutung hat.

Es werden für diese Bachelorarbeit die Klassifikatoren verwendet, die mit generativen Modellen arbeiten. Sharab wendet hier die „Latent Dirichlet Allocation“ (LDA) an. Es ist ein generatives Wahrscheinlichkeits-Modell, die in dieser Arbeit benutzt wird, um Dokumente iin einzelne Topics zu untergliedern. Die Idee dahinter ist es, dass jedes Dokument aus einer zufälligen Mischung latenter Topics besteht. Jeides Topic seinerseits besteht aus einer Verteiliung über Wörter. Diese Modell basiert auf der Annahme der „bag-of-words“.

Im Nachstehenden nimmt der Student wieder das Topic Model Feature auf. 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“ für 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“.

1) Kosinus-Ähnlichkeit:

    - Nimmt zwei Vektoren und berechnet den zwischen diesen eingeschlossenen Winkel.
    - Gibt einen Wort zwischen 0 und 1.
    - Je höher der Wert, desto ähnlicher sind zwei Vektoren

2) relative Entropie:

    - Berechnet den Abstand zwischen zwei Wahrscheinlichkeitsverteilungen
    - Je höher der Wert, desto weiter auseinander sind zwei Verteilungen.

3) Ähnlichkeit aufgrund der Anzahl gleicher Wörter:

- Je mehr gleiche Wörter unter den Top n-Wörtern zweier Topics sind, desto ähnlicher sind sie.

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.