Jakob Sharab
CIS, LMU München
Matr.Nr.: 11186993

Kolloquium zu Computerlinguistisches Arbeiten -

Protokoll zur Sitzung vom 29.05.2017

Präsentation 1: Disambiguierung eines japanischen Aspekt-Markers mit Hilfe von Parallel Korpora

Der erste Vortrag wurde von Korbinian Schmidhuber gehalten, dessen Arbeit von Annemarie Friedrich betreut wird. Korbinian merkte gleich zu Beginn an, dass er seine Bachelorarbeit bereits abgebrochen hat.
Seinen Vortrag begann er mit der Motivation hinter seinem Thema, zum einem ist es nämlich so, dass Regelbasierte Systeme bei vielen Methoden in der Computerlinguistik nicht umsetzbar sind, da Regeln oft zu abstrakt sind, oder man nur durch Intuition weiß, welche Regel man verwenden muss. Deshalb sind Beispiel-basierte Systeme oft leichter umsetzbar. Zudem ist es meist sehr aufwendig handannotierte Daten zu erstellen, dafür sind aber Parallel-Korpora immer mehr verbreitet zugänglich.
Das Ziel von Korbinians Arbeit wäre es gewesen, einen Klassifikator zur Disambiguierung eines Aspekt-Markers für die japanische Sprache zu entwickeln. Dabei sollten die Kategorien der Trainingsdaten nicht selbst annotiert werden, sondern aus der jeweiligen Übersetzung entnommen werden. Speziell ging es darum den japanischen Aspekt-Marker „te-iru“, der je nach Kontext einen anderen Aspekt ausdrücken kann, zu erkennen. Beispielsweise eine Verlaufsform, oder einen Zustand als Folge eines vorrangegangenen Ereignisses ausdrücken. Hierfür zeigte Korbinian uns verschiedene Beispiele für den entsprechenden Aspekt.
Als nächstes stellte Korbinian die Daten vor, welche verwendet werden sollten. Dabei handelte es sich um diverse Parallelkorpora, einmal auf Japanisch und einmal auf Englisch, wie z.B. Wikipedia-Artikel, der Basic-Sentences-Korpus und den „Wachturm"-Korpus. Aus diesen Daten hatte er wiederum Teil-Korpora erstellt, welche nur Sätze mit der „te-iru“ Konstruktion enthielten. Die nächsten Schritte waren die Alignierung der Verben, das Parsen und das Bestimmen der Zeitformen der englischen Verben. Wobei der letzte Schritt mit Hilfe einer Anwendung von Frau Friedrich druchgeführt werden sollte.
Danach ging Korbinian genauer auf den Klassifikator ein, wofür er zunächst einmal seine Daten in Test- und Trainingsdaten eingeteilt hatte. Er erklärte weiter, dass er für den Klassifikator verschiedene Algorithmen zur Klassifikation verwenden wollte und anschließend schauen wollte, welche die besten Ergebnisse bringen.
Zuletzt kam Korbinian auf die Probleme zu sprechen, auf die er während seiner Arbeit stieß. Zum einen liefern die meisten bekannten Alignierungssoftwares, wie z.B. „GIZA ++“ oder „fast_align“ für Sprachpaare, die eine stark unterschiedliche Wortreihenfolge in ihrer Syntax besitzen, sehr schlechte Ergebnisse. Bei 500.000 Sätzen beispielsweise gab es nur bei ca. 30% aller Wörter eine Zuordnung. Aber auch, dass die Kategorien für die japanischen Aspekt Marker oft nicht deckungsgleich sind mit den Englischen Zeitformen, war problematisch.

Präsentation 2: Regularization of Neural Networks for Natural Language Processing

Die zweite Präsentation wurde von Dayyan Smith mit Katharina Kann, als Betreuerin, gehalten. Er begann zunächst damit vorzustellen, was er allgemein in seiner Arbeit macht, nämlich die Auswirkungen von Regularisierungen eines Neural Networks für Stance Klassifikation, im Rahmen einer „fake News“- Erkennung, zu untersuchen.
Dafür definierte er erst einmal was „fake News“, also falsche Nachrichten, sind. Dabei merkte Dayyan

an, dass „fake News“ keine Nachrichten sind, mit denen man nicht einverstanden ist, sondern er verwendete die Definition der „New York Times“. Diese besagt, dass „fake News“ erfundene Berichte sind, die die Absicht haben zu täuschen.
Als nächstes ging es um die Frage, wie man „fake News“ erkennt, da dies eine schwierige und komplexe Aufgabe ist, die aber im Rahmen der „Fake News Challenge“ in verschiedene Schritte unterteilt wurde. Der erste Schritt ist, die Stance, also die Haltung oder Übereinstimmung eines Artikels zu einer Überschrift herauszufinden. Dafür ging Dayyan nochmal genauer auf den Begriff „Stance“ ein, diese sagt nämlich aus, ob verschiedene Behauptungen miteinander übereinstimmen, oder nicht. Im Fall der „Fake News Challenge“ geht es bei der „Stance Detection“ darum, ob der Inhalt einer Überschrift und eines Zeitungsartikels miteinander übereinstimmen, sich widersprechen, nichts miteinander zu tun haben, oder es unsicher ist. Hierfür zeigte Dayyan uns mehrere Beispiele für eine Überschrift und einen Artikel, wobei es auch für uns nicht immer einfach war die richtige Stance zuzuordnen.
Anschließend ging Dayyan auf das Encoding seiner Daten ein. Hierfür hat er „word2vec“ auf die einzelnen Überschriften und Artikel angewendet, um die Vektoren für die einzelnen Wörter zu erhalten. Diese hat er wiederum mit „GRU“ weiterverarbeitet, um Vektoren für die einzelnen Sätze zu erhalten. Diese werden wiederum erst durch zwei Hidden Layers und dann durch eine Classification Layer, die die richtige Stance ausgeben soll, verarbeitet, wobei er uns ein Diagramm der Architektur seiner Neural Networks zeigte.
Daraufhin merkte Dayyan an, dass sich erst zeigt, wie gut ein Modell ist, wenn man es auf unbekannte Daten anwendet, da es bei bereits bekannten Daten oft zu overfitting kommt. Daher benutzt man Regularisierung, um dies zu vermeiden und Dayyan fuhr mit verschiedenen Regularisierungsmethoden fort, die er im Laufe seiner Arbeit verwendet hat und deren Auswirkungen auf die Neural Networks er untersuchen will. Die ersten beiden Regularisierungsmethoden, die er nannte, waren „L2“ und „L1“. Bei der „L2“-Methode werden große Gewichte mehr verkleinert als kleine, da der Wert der Gewichte im Quadrat „bestraft“ wird. Im Gegensatz dazu werden bei „L1“ der absolute Wert der Gewichte „bestraft“ und so werden sowohl die großen als auch die kleinen Gewichte verringert. Die dritte Art der Regularisierung ist die „Dropout Regularization“, bei der immer ein Model trainiert wird und anschließend immer andere Neuronen ausgelassen werden.
Abschließend präsentierte Dayyan seine Ergebnisse, wobei er die Accuracy als Bewertungsmaß verwendete. Das Resultat seiner Arbeit war, dass die Accuracy meist besser ist ohne Regularisierung, wobei die „Dropout Regularization“ nicht besser war als die „L1“- und „L2“-Methoden.

Präsentation 3: Corpus Based Identification of Text Segments

Die dritte Präsentation wurde von Thomas Ebert gehalten, der seine Arbeit bei Martin Schmitt schreibt.
Als erstes ging Thomas auf die Frage ein, was mit einem Textsegment gemeint, da je nachdem beispielsweise einzelne Buchstaben, Wörter oder Sätze gemeint sein können. Er meinte, dass für die Textaufbereitung für NLP-Aufgaben, meist eine Tokenisierung verwendet wird, das heißt es wird nach einzelnen Wörtern vorgegangen. Er merkte jedoch an, dass es nicht immer einfach ist zu definieren, was ein einzelnes Wort ist und was nicht, da wir diese Entscheidung oft intuitiv treffen.
Als nächstes ging Thomas auf das Ziel seiner Arbeit ein, nämlich einen Algorithmus zu entwickeln, der einen Satz in seine besten Segmente, nämlich Buchstaben N-Gramme, zerlegt. Sein Vorgehen war dafür erst einmal so, dass er seine Texte in Buchstaben N-Gramme der Länge 1-10 zerlegt hat. Wobei sein Korpus, ein englischsprachiger Wikipediakorpus, unannotierte Rohtexte enthält und er die ersten 10.000 Texte zur Extraktion verwendet hat. Als nächstes hat Thomas eine Frequenzliste für die einzelnen N-Gramme erstellt und diese mit einem Gütemaß nach der Formel $n * \log(freq)$, wobei n die N-Gramm-Länge und $freq$ die Häufigkeit des N-Gramms ist, bewertet. Es wurde jedoch nicht ganz klar inwiefern dieses Gütemaß, etwas über die Qualität des N-Gramms aussagt. Zum Testen wird ein Satz eingegeben, der in N-Gramme, die das höchste Gütemaß besitzen, zerlegt werden.
Als Problem dabei nannte Thomas, dass die Laufzeit exponentiell mit der Größe des Eingabesatzes

erhöht. Zur Lösung dieses Problems nannte er einen heuristischen Ansatz, bei dem die Größe der Windows festgelegt wird, wobei jedoch die Berechnung der höchsten Güte nicht mehr gewährleistet ist.
Als nächstes kam Thomas zum Punkt der Evaluierung der einzelnen Textsegmente, wobei diese jedoch schwierig, da man sich oft uneinig über die Granularität dieser ist und je nach Anwendung einzelne Fehler wichtig oder unwichtig sein können. Daher wird die Auswirkung der Fehler auf die Endanwendung als Maß verwendet. Zur Evaluierung verwendete er erst einmal „word2vec“, um Buchstaben N-Gramm-Embeddings zu erhalten und will diese weiter zur Sentiment Analysis auf Filmkritiken verwenden. Als mögliches Modell hierfür nannte Thomas eine Sigmoidfunktion auf den letzten Zustand eines „Long-Short-Term Memory“- Encoders anzuwenden.
Als letzten Punkt nannte er die Erkenntnisse, die er bis jetzt schon erlang hat und noch weitere offene Fragen. Beispielsweise kann man jetzt schon sehen, dass die Buchstaben N-Gramme eine Zipf’sche Verteilung aufweisen und dass die häufigsten N-Gramme, die Größer als drei sind, hauptsächlich Funktionswörter sind. Offene Fragen sind bis jetzt z.B. noch, ob es andere Möglichkeiten gibt, die N-Gramme zu extrahieren und ob das Ergebniss der Evaluierung schon aussagekräftig ist.