CIS, LMU München
29.05.2017
Michael Strohmayer
11137111
michael.strohmayer@campus.lmu.de

Protokoll zur Sitzung 29.05.2017 – Computerlinguistisches Arbeiten

Im Laufe der Sitzung wurden von den Studenten Korbinian Schmidhuber, Dayyan Smith und Thomas Ebert jeweils das Thema ihrer Bachelorarbeit vorgestellt.

Disambiguierung eines japanischen Aspekt-Markers mithilfe von Parallel-Korpora

Als erster Referent der Sitzung stellt Korbinian Schmidhuber sein bearbeitetes Bachelorthema vor. Die Problemstellung hier war, dass regelbasierte Systeme bei vielen Methoden in der Computerlinguistik nicht umsetzbar, da die Regeln oft zu abstrakt sind. Daher sind beispielbasierte Systeme oft sehr viel leichter umsetzbar. Die Voraussetzung hierfür ist, dass ausreichend Daten verfügbar sind. Per Hand annotierte Daten sind meist sehr aufwendig zu erstellen, vor allem weil im Japanischen manche Wörter oft mehrere ähnliche Bedeutungen haben. Bei Übersetzungen müssen mehrdeutige Konstruktionen durch den Übersetzer disambiguiert werden.
Das Ziel seine Bachelorarbeit war es, einen Klassifikator zur Disambiguierung eines Aspekt-Markers im Japanischen zu programmieren. Die Kategorien der Trainingsdaten sollen hierbei nicht selbst annotiert, sondern der jeweiligen Übersetzung entnommen werden. Nach der Erklärung der Ziele seiner Arbeit, fügt Korbinian noch eine kurze Definition von Aspekt an. Aspekt ist neben Tempus eine Kategorie in Sprachen, welche Ereignisse klassifizieren kann. Im Deutschen ist Aspekt nicht annotiert, darum benutzt er zum Vergleich die Englische Sprache. Hier ist Aspekt vergleichbar mit dem Progressive. Der Aspekt Marker "te-iru" im Japanischen kann je nach Kontext einen unterschiedlichen Aspekt ausdrücken. Er kann zum Beispiel einen Verlauf darstellen (im Englischen vergleichbar mit z.B. "I'm eating bread") oder als Folge ("The dog is dead", nicht aber "The dog is dying") auftreten.
Die Daten die Korbinian hier verwendet hat, waren Wikipedia-Korporas, ein Basic Sentences Korpus und diverse Wachturm Ausgaben in Englisch und Japanisch.
Er bereitete die Daten die Daten auf, indem er zuerst alle Sätze, welche nicht die "te-iru" Konstruktion enthielten, aus dem Korpus herausfilterte. Anschließend alignierte er die Verben (teilweise von Hand) in einem Korpus, und die restlichen mit Hilfe von Online Wörterbüchern. Das darauf folgende Parsen der Zeitform der englischen Verben wurde mit einer von Annemarie Friedrich zur Verfügung gestellten Anwendung erledigt.
Für den Klassifikator, welcher nicht mehr bearbeitet wurde, wäre eine Einteilung der Daten in Training und Test Daten und die Anwendung verschiedener Algorithmen zur Klassifikation nötig. Außerdem fehlte hier noch ein Evaluierungsverfahren.
Seine größten Probleme beschreibt Korbinian als die unterschiedliche Satzstellung zwischen den verglichenen Sprachen. Die Alignierung wurde mit GIZA++ und fast_align gemacht. Diese performen bei Englisch-Deutsch bereits mit wenigen Daten sehr gut, bei Japanisch-Englisch sehr schlecht. Korbinian erklärt dies mit der Position des Verbs im Japanischen, welches immer am Ende eines Satzes steht.

CIS, LMU München
29.05.2017
Michael Strohmayer
11137111
michael.strohmayer@campus.lmu.de

Protokoll zur Sitzung 29.05.2017 – Computerlinguistisches Arbeiten

Im Laufe der Sitzung wurden von den Studenten Korbinian Schmidhuber, Dayyan Smith und Thomas Ebert jeweils das Thema ihrer Bachelorarbeit vorgestellt.

Regularization of Neural Networks for Natural Language Processing

Der zweite Vortrag war von Dayyan Smith, betreut von Katharina Kann. Sein Thema war hier die Regularisierung von Neuronalen Netzwerken für die Verarbeitung von natürlicher Sprache.
Dayyan beginnt den Vortrag mit einer kurzen Einführung was er eigentlich tut. In seiner Erklärung fallen die Begriffe „stance classification“ und „fake news detection“. Diese versucht er nach einem kurzen Überblick über all seine Vortragspunkte erst einmal zu definieren. Hierbei beginnt er mit Fake News. Er erklärt, dass diese selbst für Experten teilweise schwer zu verifizieren sind. Im Rahmen der Fake News Challenge entstand so die Idee, moderne Technologie zu benutzen um Fake News effektiv zu bekämpfen. Der erste Schritt ist hierbei die Stance Detection. Hierbei wird überprüft, ob zwischen der Headline und dem Body eine Verbindung besteht. Die möglichen Haltungen sind hier „agree“, „disagree“, „discuss“ und „unrelated“. Um die Unterschiede der Haltungen deutlich zu machen, führt er einige Beispiele auf und fragt zu Zuhörer nach ihrer Meinung zu den jeweiligen Fällen.
Anschließend erklärte er, wie das Encoding funktioniert. Mit Hilfe von „Word2Vec“ werden Headline und Body getrennt von einander initialisiert und Word Embeddings erstellt. Danach wird GRU benutzt um die Satzrepräsentationen zu erstellen. Dieser werden dann von Head und Body konkateniert. Innerhalb der Neuronalen Netzwerke werden diese dann in Hidden Layers unterteilt und je nach Zustimmungsgrad klassifiziert.
Dayyan erklärt nach dem internen Ablauf auch noch die Regularisierungsmethoden die er verwendet hat. Dazu zählen die L2 Regularisierung, bei der große Gewichte stärker bestraft werden als kleine, die L1 Regularisierung, bei der große und kleine Gewichte gleich stark heruntergeogen werden, sowie die Dropout Regularisierung.
Nach einer kurzen Fragerunde präsentierte er noch seine Ergebnisse der verschiedenen Regularisierungsmethoden und seine Referenzen.

CIS, LMU München
29.05.2017
Michael Strohmayer
11137111
michael.strohmayer@campus.lmu.de

Protokoll zur Sitzung 29.05.2017 – Computerlinguistisches Arbeiten

Im Laufe der Sitzung wurden von den Studenten Korbinian Schmidhuber, Dayyan Smith und Thomas Ebert jeweils das Thema ihrer Bachelorarbeit vorgestellt.

Corpus based Identification of text segements

Als letzter Referent stellt Thomas Ebert das Thema seiner Bachelorarbeit vor. Er verfolgte das Ziel der Entwicklung eines Algorithmus, welcher einen eingegebenen Satz oder Text in seine besten bedeutungstragende Einheiten zerlegt. Da bedeutungstragende Einheiten in der Sprache nur schwer zu definieren sind (Morphem, Wort, Phrase, Satz...) verwendet er für die Zerlegung Buchstaben oder Ngramme. In dieser Arbeit soll außerdem untersucht werden, ob ein symbolischer Ansatz nicht doch besser ist und welche Chancen und Risiken dieser bietet.
Um dies zu testen extrahierte Thomas zuerst Ngramme der Länge 1-10 aus einem Englischen Wikipedia Korpus. Dieser enthält unannotierte Rohtexte. Die ersten 10.000 Texte (entsprechen etwa 22 Millionen Zeichen) des Korpus wurden zum Extrahieren verwendet und daraus eine Frequenzliste mit Ngrammen erstellt. Diesen Ngrammen wurden auf Basis ihrer Frequenz Gütemaße berechnet. Das Gütemaß ist hier ein Wert, welcher eine Aussage über die Wichtigkeit des Ngrams für den Text trifft. Hierbei entspricht ein hoher Wert einer hohen Wichtigkeit.
Thomas stieß bei seiner Bearbeitung auf das Problem, dass mit größer werdender Eingabe die Laufzeit exponentiell steigt. Seine Lösung hierfür war ein heuristischer Ansatz, bei dem er die Größe eines Windows festlegen konnte. Hierbei ist jedoch die Berechnung der höchsten Güte nicht mehr garantiert. Hier evaluiert er momentan noch, ob die Segmentierung vielleicht noch bessere Ergebnisse liefert als der symbolische Ansatz.
Die Evaluierung gestaltet sich bei Textsegmenten generell schwierig, denn je nach Anwendung können Fehler relevant oder irrelevant sein. Dies hat dann z.B. Auswirkungen auf die Endanwendung.
Um Ngramm embeddings zu erhalten wurde word2vec und zur Evaluierung auf Satzebene Sentiment Analse verwendet.
Zuletzt stellte Thomas noch seine Erkenntnisse und offenen Fragen vor. Diese beschränkten sich im wesentlichen auf die Zipfsche Verteilung von Buchstaben Ngrammen. Zusätzlich möchte er noch andere Möglichkeiten finden um Ngramme zu extrahieren und überprüfen, ob sein Ergebnis eigentlich so wie es momentan ist schon aussagekräftig ist.