29.05.17

Protokoll zur Sitzung vom 29.05.2017 – Computerlinguistisches Arbeiten

1. Referat: Korbinian Schmidhuber, Disambiguierung eines japanischen Aspekt-Markers mithilfe von Parallel-Korpora (Annemarie Friedrich)

Korbinian beginnt damit die Motivation hinter der Arbeit zu erklären. Er erläutert, dass regelbasierte Systeme bei vielen Methoden in der Computerlinguistik nicht umsetzbar sind, da Regeln oft zu abstrakt sind. Beispiel-basierte Systeme hingegen sind oft leichter umsetzbar, falls genug Daten verfügbar sind. Hand-Annotierte Daten sind meist sehr aufwendig zu erstellen. Parallel-Korpora hingegen sind viel verbreiteter und auch leichter zugänglich. Bei Übersetzungen müssen mehrdeutige Konstruktionen durch den Übersetzer disambiguiert werden.

Das Ziel der Arbeit ist es, einen Klassifikator für die Disambiguierung eines Aspekt-Markers im Japanischen zu trainieren. Die Kategorien der Trainingsdaten sollen nicht selbst annotiert werden, sondern aus der jeweiligen Übersetzung entnommen werden.

Der Aspekt-Marker den Korbi wählt ist der ‚te-iru'-Marker, der im Japanischen je nach Kontext einen unterschiedlichen Aspekt ausdrücken kann. Der Marker kann zum Beispiel einen Verlauf oder aber auch einen Zustand als Folge eines vorangegangenen Ereignisses ausdrücken.

Die Daten, die er verwendet hat, war einerseits ein Wikipedia-Korpus, zum anderen ein Basic-Sentences-Korpus, das im Rahmen einer Universität erstellt worden ist, und auch die sogenannten „Wachturm"-Ausgaben in Englisch und Japanisch.

Zur Aufbereitung der Daten erstellt er Teil-Korpora durch herausfiltern aller Sätze, die die „te-iru" Konstruktion nicht enthalten. Als nächstes erfolgt die Alignierung der Verben und das Parsen und bestimmen der Zeitform der englischen Verben. Die Software, die dazu benutzt wurde, wurde von seiner Betreuerin gestellt.

Korbinian erwähnt, dass er seine Bachelorarbeit bereits abgebrochen hat, und erläutert welche Schritte noch zu erledigen wären. Für den Klassifikator sollte die Einteilung der Daten in Trainings- und Testing-Daten erfolgen, um die Anwendung verschiedener Algorithmen zur Klassifikation testen. Die Evaluation sollte die erreichte Genauigkeit mithilfe der Testdaten erfolgen.

Als nächstes beschreibt er die Probleme, die bisher aufgekommen sind. Diese liegen hauptsächlich bei der Alignierung mit der Alignierungssoftware. Das Programm liefert für Sprach-Paare mit sehr unterschiedlicher Wortreihenfolge, mit wenigen Daten, nur schlechte Ergebnisse. Die Alignierung von 500K Sätzen ergab nur bei 30% aller Wörter überhaupt eine Zuordnung.

Ein weiteres Problem ist, dass die Kategorien für den japanischen Aspekt-Marker nicht deckungsgleich sind mit den Englischen Tenses sind.

2. Referat: Dayyan Smith, Regularization of Neural Networks for Natural Language Processing (Katharina Kann)

Dayyan presents his dissertation topic with an example of Fake News and explains what the meaning of fake news. The goal of Dayyan's work is to explore the effect of regularization of a neural network for stance classification. Fake new can have different meanings, however, they are defined as 'a made-up story with an intention to deceive' in his dissertation.

Dayyan ansers the question of how fake news can be detected by giving us the answer that in general, it is very difficult to assess the veracity of a news story, because it is a very complex task. He explains that even for humans this task can be complicated. Exploring how artificial intelligence technologies could be leveraged to combat those fake news is the main subject of the so-called Fake News Challenge. He gives a short overview of the challenge, its goals and tasks.

The first stage in his work is the stance detection, which is finding the stance of any headline in an article. He explains that the stance is the result of, whether two different texts agree with each other or not. He gives some examples to emphasize this. Possible stances are: agree, disagree, discuss and unrelated. Dayyan presents a few examples for each stance, including a case where it is not clear what sort of stance should be chosen as both text fragments neither fully agree nor disagree with each other. In the context of the fake news challenge, stance detection is estimating the relative perspective (or stance) of two pieces of text relative to a topic, claim or issue.

He uses word2vec pre-encodings to initialize word embeddings as input to a Neural Network and uses a Gated Recurrent Unit (GRU) to get the sentence representations. Next he describes the network architecture in a graphic, which includes 2 hidden layers.

Dayyan emphasizes that one should not be impressed when a complex model fits a data set very well, because with enough data every model can fit a given data set.

In the next part he shows some diagrams to visiualize the different states of underfitting and overfitting. Following this, Dayyan presents different types of regularization.

As of the type 'L2 Regularization', big weights are pushed down more than small weights because the square of weights is penalized. Another type of regularization is the Dropout Regularization, which he visualizes with different figures comparing a normal regularization to the dropout regularization. Next he shows his results, including examples using no regularization, L1 and L2 regularization.

3. Referat: Thomas Ebert, Corpus based Identification of Text Segments

(Martin Schmitt)

Thomas beginnt mit der Motivation für seine Arbeit und erläutert zunächst grundlegende Begriffe, wie zum Beispiel die Begriffe „Textsegment“ (bedeutungstragende Einheit), „Morphem“, „Wort“, „Phrase“, „Satz“ und „Topic“.

Die Textaufbereitung für NLP-Aufgaben ist meist wortbasierend (z. B. die Tokenisierung). Das Wort ist nicht eindeutig definiert aber intuitiv. Die Tokenisierung ist im Allgemeinen sehr fehleranfällig, weshalb eine lokale Anpassung nötig ist. Im Zentrum steht die Frage, ob das intuitive Konzept „Wort“ die beste Art ist für einen Computer um einen Text zu segmentieren.

Ziel der Arbeit ist es, einen Algorithmus zu entwickeln, der einen Input-Text in seine besten Segmente zerlegt (z.B. in Buchstaben oder N-Gramme etc.). Es stellt sich die Frage ob, der nicht-symbolische Ansatz besser ist.

Als nächstes beschreibt Thomas sein Vorgehen. Der erste Schritt ist es N-Gramme der Länge 1-10 aus dem Wikipedia Korpus zu extrahieren. Der Korpus enthält unannotierte Rohtexte und keine POS Tags. Die ersten 10K Texte des Korpus werden zum Extrahieren verwendet. Es wird eine Frequenzliste für die N-Gramme erstellt. Die N-Gramme werden dann mit einem Gütemaß bewertet: Gütemaß = n*log(freq), wobei freq die absolute Häufigkeit des N-Gramms ist. Zum Testen wird ein Satz eingegeben.

Als nächstes erläutert er die Probleme in seiner Arbeit. Mit der Größe der Eingabe steigt die Laufzeit exponentiell. Die Lösung hierzu wäre ein heuristischer Ansatz. Auch beim Festlegen der Größe für das Fenster gibt es Schwierigkeiten. Bei der Berechnung der höchsten Güte ist nicht mehr garantiert, aber die Segmentierung ist ggf. noch besser als beim symbolischen Ansatz.

Die Evaluierung von Text Segmenten ist schwierig. Häufig gibt es Uneinigkeit über die Granularität von Segmenten. Je nach Anwendung können Fehler relevant oder irrelevant sein. Thomas betont, dass z.B. bei Information Retrieval die Korrektheit von Segmentgrenzen vernachlässigt werden kann, bei „news boundary“ hingegen, ist sie wichtig. Die Auswirkung auf die Endanwendung (z.B. Sentiment Analysis) wird als Maß verwendet.

Es wird word2vec verwendet um Buchstaben N-Gramm embeddings zu erhalten und für die Evaluation wird die Sentiment Analyase auf Satzebene evaluiert. Auch werden Movie Review Data und der Vergleich mit Word embeddings für die Evaluierung verwendet. Er erläutert ein mögliches Modell für die Evaluierung: das Sigmoid auf dem letzten Zustand eines LSTM-Encoders (LSTM = Long-Short-Term Memory). Die Sigmoidfunktion wird auf die Summe der gewichteten Eingabewerte angewendet um ein Ergebnis zu erhalten.

Sein letzter Punkt beinhaltet Erkenntnisse und offene Fragen. Es lässt sich zusammenfassend sagen, dass auch Buchstaben N-Gramme eine Zipf‘sche Verteilung aufweisen. Häufigste N-Gramme, die größer 3 sind, enthalten Funktionswörter. N-Gramme, die größer 8 sind, enthalten Inhaltswörter. Bisher sind noch keine Ergebnisse für die Evaluierung vorhanden. Weitere Fragen wären, ob es noch andere Möglichkeiten gibt, um N-Gramme zu extrahieren und ob das Ergebnis der Evaluierung schon aussagekräftig ist.