Christoph
Flemmig
CIS, LMU München

1.06.2017

Protokoll für das Kolloquium vom 29.5.2017

In dieser Sitzung wurde drei Bachelorarbeiten von den Studenten in Form von kurzen Vorträgen mithilfe einer kleinen Präsentation am Beamer vorgestellt.

1. Vortrag:

Korbinian Schmidhuber, Ba-Betreuerin: Annemarie Friedrich
Thema: Disambiguierung eines japanischen Aspekt-Markers mithilfe von Parallel-Korpora

Zu Beginn seines Vortrags sprach Korbinian über die Motivation für seine Arbeit. Er sprach darüber dass Regelbasierte Systeme bei vielen Methoden der Computerlinguistik nicht umsetzbar sind da Regeln oft zu Abstrakt sind. Daher sind Beispiel-basierte Systeme oft leichter umsetzbar falls ausreichend Daten verfügbar sind. Ein weiteres Problem dass er ansprach war dass Hand annotierte Daten meist sehr aufwendig zu erstellen sind. Da Parallel-Korpora leichter zugänglich und weiter verbreitet sind konzentriert sich seine Arbeit auf diese. Ein weiteres Problem auf das er sich konzentriert ist dass Disambiguierungen für mehrdeutige Konstruktionen oft durch den Übersetzter gemacht werden müssen.
So beschrieb Korbinian das Ziel seiner Arbeit als das Trainieren eines Aspekt-Markers im Japanischen.
Als nächstes erläuterte Korbinian den Hintergrund des sprachlichen Problems auf das sich seine Arbeit konzentriert. Er erklärte dass obwohl viele Anwendungen des japanischen Aspekt-Markers „te-iru“ oft mit dem englischen Progressive übersetzt werden kann, dies nicht in allen Fällen möglich ist. So können damit auch Zustände ausgedrückt werden was mit den Progressive nicht möglich ist.
Danach ging Korbinian auf die für seine Arbeit relevanten Parallel-Korpora ein. Er verwendet hier den Wikipedia-Korpus, den Basic Sentences Korpus und Wachturm Ausgaben in Englisch und Japanisch.
Der nächste Teil seines Vortrages konzentrierte sich auf die Aufbereitung der Daten. Er erstellte Teil-Korpora durch das Herausfiltern von Sätzen die die „te-iru“-Konstruktion nicht enthalten. Danach konzentrierte er sich auf die Alignierung der Verben dabei war einer seiner Korpora bereits handaligniert, bei den anderen Korpora arbeitete er mit Online-Wörterbüchern. Der letzte Schritt war das Parsen und bestimmen der Zeitform der englischen Verben mithilfe einer Anwendung von Annemarie Friedrich.
Da er seine Bachelorarbeit abgebrochen hatte gab Korbinian nur noch kurze Angaben über seinen Klassifikator und die Evaluation. Bei der Evaluation sollte es um die erreichte Genauigkeit mithilfe der Testdaten gehen.
Zum Schluss sprach er noch über einige Probleme und seine Lösungsansätze. So erzielte er bei der Alignierung mit bekannter Alignierungssoftware mit wenigen Daten nur unzufriedenstellende Ergebnisse. Ein Ansatz das Problem zu lösen war die Verben erst ins englische zu übersetzen und dann nach passenden Sätzen zu suchen jedoch war ein Problem dabei dass Kategorien für den japanischen Aspekt-Marker nicht deckungsgleich mit den englischen Tenses waren.

2. Vortrag:
Dayyan Smith, BA-Betreuerin: Katharina Kann
Thema: Regularization of Neural Networks for NLP

At the beginning Dayyan spoke about the main Focus of his work. Exploring the effect of regularization of a neural network for stance classification in the context of fake news detection. To explain that further he started with a closer look at the term „fake news“ and how he plans to detect fake news. For that he quoted The New York Times. Fake news according to the Times are „a made-up story with an intention to deceive.“. Since Automatic Fake News Detection is a complex and cumbersome task, the way he approached it is broken down into several stages. He did this with the intent to take part in the Fake News Challenge.
The first stage described was stance detection which is the important one for his work. In this step he tries to find the stance of a given article towards it's headline. The possible stances consist of agree, disagree, discuss and unrelated. After that he gave several examples of headlines, bodies of text and their possible stances towards each other.
What followed was a description of the encoding process he uses. He uses word2vec word embeddings to get vectors for each word. Then he uses GRU to get sentence representations. After that he uses a sequence of word vectors to produce sentence embeddings and then concatenates them. Following that they are put trough two hidden layers. Finally, the embeddings are concatenated with the outputs of both hidden layers for both headline and boty and put through the classification layer.
The next part of the presentation was about regularization. This is the second important part of Dayyans work. He uses three different kinds of regularization, L2 Regularization, L1 Regularization and Dropout Regularization. In L2 Regularization big weights are pushed down more than small weights because the sqare of weights is penalized. L1 Regularization pushes down both big and small weights by a smaller amount because the absolute value is penalized. Dropout Regularization works by changing the model itself. By dropping neurons from the neural network the goal is to find which neurons are more important.
At the end Dayyan mentioned some of his problems and presented his results so far. According to him Regularization causes big fluctuations at this point of his work and the training takes a lot of time. So far his results with L1 and L2 Regularization offer only slight improvements over the results he gets without regularization.

3. Vortrag:
Thomas Ebert, BA-Betreuer: Martin Schmitt
Thema: Corpus based identificatio of text segments

Auch Thomas begann seinen Vortrag mit der Motivation für seine Arbeit. Da die Textaufbereitung für NLP-Aufgaben meist wortbasiert ist ergibt sich eine gewisse Problematik da Wörter oft nicht eindeutig defniert sind, für den Menschen jedoch intuitiv. Dadurch ist die Tekoenisierung jedoch sehr fehleranfällig. Deshalb ergibt sich die Frage ob das intuitive Konzept Wort die beste Art für einen Computer ist einen Text zu segmentieren.
Das Ziel seiner Arbeit ist die Entwicklung eines Algorithmus der einen Satz oder Text in seine besten Segmente zerlegt.
Sein Vorgehen ist dabei das extrahieren von N-Grammen der Länge 1-10 Zeichen aus dem englischen Wikipedia-Korpus. Dieses Korpus enthält unanotierte Rohtexte und er verwendet die ersten 10000 Texte für des Korpus für seine Arbeit. Für die N-Gramme wird eine Frequenzliste erstellt und sie werden danach mit einem Gütemaß versehen. Zum testen wird dann ein Satz eingegeben der in die N-Gramme mit dem höchsten Gütemaß zerlegt werden soll. Hier präsentierte Thomas schon einige Probleme mit dem Ansatz. So stieg die Laufzeit exponentiell mit der Größe der Eingabe. Seine Lösung hierfür war ein heuristischer Ansatz. Ein weiteres Problem war die Größe des Fensters festzulegen.
Als nächstes gab Thomas eine Ausführung über die Evaluierung. Er hatte sie bis zu diesem Zeitpunkt noch nicht durchgeführt, jedoch nannte er schon einige mögliche Probleme. So sei die Evaluierung von Textsegmenten schwierig da es auch häufig uneinigkeit über die Granularität der

Segmente gibt. Des weiteren können je nach Anwendung Fehler relevant oder irreleveant sein. Es sei in Betracht zu ziehen die Auswirkung auf die Endanwendung als Maß zu verwenden. Eine weitere Möglichkeit sei die Verwendung von word2vec um Buchstaben N-Gramm embeddings zu erhalten. Eine andere Möglichkeit ist auch die Sentiment Analyse auf der Satzebene zur Evaluation. Dafür würde er Movie Review Data verwnden und sie mit den Word embeddings vergleichen. Als letztes nannte er noch die Sigmoidfunktion für eine mögliche Evaluation.
Am Ende seines Vortrages nannte Thomas noch einige Erkenntnisse und offene Fragen seiner Bachelorarbeit. So erhielt er die Erkenntnis dass auch Buchstaben N-Gramme eine Zipfsche Verteilung aufweisen und stellte fest dass man an der größe der N-Gramme sowohl Funktionswörter als auch Inhaltswörter erkennen konnte. Als Fragen stellten sich ihm noch ob es noch andere Möglichkeiten gäbe die N-Gramme zu extrahieren und ob das Ergebnis einer Evaluierung seiner bisherigen Arbeit schon aussagekräftig sei.