Mitschrift: Sohail Malih

Referat Korbinian Schmidhuber: Betreuerin Annemarie Friedrich (Montag, den 29.05.2017)

Disambiguierung eines japanischen Apsekt-Markers mithilfe von Parallel-Korpora

Das Thema wurde zuerst mit der Herangehensweise und der Motivation gestartet.
Im Grunde geht es um die Regelbasierte Systeme bei vielen Methoden in der Computerlinguistik nicht umsetzbar sind, da diese Regeln oft zu abstrakt sind.
Daher sind Beispiel-basierte Systeme oft leichter umsetzbar, falls Daten hier verfügbar sind.

Hand annotierte Daten sind meist sehr aufwendig zu erstellen.
Parallel – Korpora sind immer verbreiterter und leichter zugänglich dagegen.
Zudem müssen bei Übersetzungen mehrdeutige Konstruktionen durch den Übersetzer disambiguiert werden.

Als nächstes wurde über den Aspekt gesprochen. Aspekt wird markiert, um nicht übersetzbare Konstruktionen wie progressive im Englischen als Verlauf zu beschreiben. Das Problem hier ist jedoch, dass der Übersetzer sich auf eine Bedeutung festlegen muss.

Im Verlauf wird das Ziel erklärt. Wir wollen das Traineren eines Klassifikators zur Disambiguierung eines Aspekt-Markers im Japanischen. Die Kategorien der Trainingsdaten sollen nicht selbst annotiert werden.

Als Idee wird erwähnt, das im englischen die Verlaufsform durch das Progessive gebildet wird. Ein Zustand kann nicht durch das Progressive ausgedrückt werden.

Es gibt verschiedene Parallelkorpora. Das Wikipedia Korpus besitzt 500,000. Dann haben wir hier noch die Basic Sentences welche nur 5000 hat.

Nun wird die Aufbereitung der Daten erläutert.
Es wird eine Teil Korpora erstellt, durch das herausfiltern aller Sätze die die teiru Konstruktion nicht enthalten. Alignierung der Verben (ein Korpus bereits Hand-aligniert in anderen Korpora mithilfe von Online-Wörterbüchern)
Um zu Parsen und bestimmen der Zeitform der englischen Verben, wird hier mithilfe einer Anwendung von Annemarie Friedrich, durchgeführt.

Es wird eine Einteilung der Daten in Trainings- und Testing Daten sortiert.
Verschiedene Algorithmen werden hier zur Klassifikation angewandt.

Zu ende hin werden die Probleme erwähnt. Und zwar liefert für Sprach Paare der bekannte Alignierungssoftware sehr unterschiedliche Wortreihenfolge mit wenigen Daten nur schlechte Ergebnisse. Bei der Alignierung von 500000 Sätzen ergab nur bei 30% aller Wörter überhaupt eine Zuordnung.