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

Da regelbasierte Systeme in der Computerlinguistik oft wegen ihrer abstrakten Natur nicht umsetzbar sind, ist es leichter ein auf einem Beispiel basierendes System zu verwenden, vorausgesetzt es stehen genügend Daten zur Verfügung. Das ist die Motivation der Arbeit. Von Hand annotierte Daten sind teuer und aufwändig zu erstellen. Die Verwendung von Parallelkorpora wird immer gebräuchlicher, da diese leicht zugänglich sind. Ziel der Arbeit ist es einen Klassifikator zur Disamgiuierung eins Aspekt-Markers im Japanischen zu trainieren. Die Kategorien der Trainingsdaten sollen nicht selbst anmontiert werden, sondern der jeweiligen Übersetzung entnommen werden. Der Hintergrund ist, dass der Aspekt Marker „te iru“ im Japanischen je nach Kontext einen unterschiedlichen Aspekt ausdrücken kann. Zum Beispiel „Watusi ha pan o tabe-te iru“ -> „I am eating bread“.
Im Englischen wird die Verlaufsform durch das Progressive gebildet. Ein Zustand kann im Englischen nicht durch das Progressive ausgedrückt werden. Ein Verlauf wird im Japanischen als - Zustand als Folge eines vorangegangenen Ereignisses - ausgedrückt.
Daten die in der Arbeit verwendet werden sind: verschiedene Parallelkorpora, der Wikipedia Korpus, der Basic Sentences-Korpus und Wachturm Ausgaben in Englisch und Japanisch. Daten werden wie folgt aufbereitet: Man erstellt Teil-Korpora durch das Herausfiltern aller Sätze, die die „te-iru“ Konstruktion nicht enthalten. Verben sind in einem Korpus bereits per Hand alignert. In weiteren Korpora erfolgt die Alignierung der Verben mithilfe von Online-Wörterbüchern. Die Daten werden geparst und deren Tempus bestimmt. Für die Klassifikation werden die Daten zunächst in Trainings- und Testdaten eingeteilt. Danach werden verschiedene Algorithmen zur Klassifikation verwendet. Evaluiert wird die erreichte Genauigkeit mithilfe der Testdaten. Probleme gibt es bei der Alignierung. Bekannte Alignierungssoftware liefert, für Sprach-Paare mit sehr unterschiedlicher Wortreihenfolge und mit wenigen Daten, nur sehr schlechte Ergebnisse. Zum Beispiel gab es bei der Alignierung von 500.000 Sätzen nur für 30% der Wörter eine Zuordnung. Kategorien für den japanischen Aspekt-Marker sind nicht deckungsgleich mit Englischen Tenses.