Protokoll zur Sitzung am 29.5.17
(Kolloquium)

Korbinian Schmidhuber „Disambiguierung eines japanischen Aspekt-Markers mithilfe von Parallel-Korpora“ Betreuerin: Annemarie Friedrich

Der Vortragende hat seine Bachelorarbeit präsentiert, die er allerdings erst in den nächten Semester zu Ende schreibt.

Zu Beginn der Präsentation hat der Vortragende über die Motivation zur Themawahl gesprochen: regelbasierte Systeme sind bei vielen Fragestellungen in der CL schwer umsetzbar, weil die Regeln entweder nicht ersichtlich sind (sondern nur nach Intuition verwendbar), oder die sind zu abstrakt. Aus diesem Grund sind oft beispielbasierte Systeme gefragt, die nicht mithilfe der Regeln, sondern mithilfe der Beispielen traniert sind. Es ist außerdem aufwendig, Daten per Hand zu annotieren, deswegen bieten sich die Parallel-Korpora an.

Das Ziel der Arbeit war es, einen Klassifikator zu tranieren, der zu solcher Disambiguerung dient. Der Aspektmarker wurde vom Vortragenden als eine morphologisch markierende Kategorie in manchen Sprachen definiert. Im Englischen gibt es z.B. solche Aspekte wie Verlauf, abgeschlossenes Ereignis, wiederholte Ereignisse.

In der Arbeit sollten die Kategorien der Trainingsdaten nicht selbst annotiert werden, sondern der jeweiligen englischen Übersetzung entnommen werden. Für seine Arbeit hat der Vortragende den Aspektmarker „*te-iru*“ gewählt. Je nach Kontext drückt er einen unterscheidlichen Aspekt aus: *Verlauf* (eng. I’m eating bread.) oder *Zustand als Folge vorangegangenen Ereignisses* (eng. The dog is dead.). Die Idee ist, dass die Verlaufsform in der englischen Übersetzung durch das Progressive gebildet wird und ein Zustand nicht.

Was die Parallel-Korpora angeht, wurden im Rahmen der Arbeit die folgenden benutzt: Wikipedia-Korpus (ca. 500 000 Sätze), Basic-Sentences-Korpus (ca. 5000 Sätze) und die Ausgaben des „Wachturm“-Magazins. Es wurden zuerst durch Herausfiltern der nicht relevanten Sätzen (ohne „*te-iru*“) Teilkorpora erstellt und dann eine Alignierung der Verben durchgeführt. Danach wurden die englischen Verben so geparst, dass für jedes Verb seine Zeitform bestimmt wurde (mithilfe der Anwendung von A. Friedrich). Es war geplant, anhand dieser Daten verschiedene Algorithmen zur Klassifikation anzuwenden.

Am Ende des Vortrages wurde über die Probleme, die während der Arbeit aufgetaucht sind, berichtet. Die Hauptschwierigkeit bestand darin, Verben zu alignieren: bekannte Alignierungssoftware (z.B. Giza++, fast_align) liefern für Sprachpaare mit sehr unterschiedlicher Wortreihenfolge (und das ist beim Engisch und Japanisch genau der Fall) nur sehr schlechte Ergebnisse (nur ca. 30 % aller Wörter wurden zugeordnet). Außerdem sind die Kategorien für den Aspekt-Marker nicht deckungsgleich mit den englischen Tenses. An diesen Problemen soll in nächsten Semestern gearbeitet werden.