CIS, LMU München

Studentin: Elena Atanasova

29.05.2017

Protokoll zur Sitzung vom 29.05.2017 – Computerlinguistisches Arbeiten

1. Vortrag: Korbinian Schmidhuber , "Disambiguierung eines japanischen Aspekt-Markers mithilfe von Parallel-Korpora"

BA-Betreuer: M. Sc. Annemarie Friedrich

Den erstenen Vortrag des Tages hält Korbinian Schmidhuber, der von Frau Friedrich betreut war. Schmidhuber hat vor zwei Wochen seiner Bachelorarbeit abgebrochen, deswegen presentierte nur was er bis dahin gemacht hatte und welche Probleme aufgetaucht waren. Am Anfang der Presentation stellte er die Motivation dar und erklärte, dass die regelbasierte Systeme bei vielen Methoden in der Computerlinguistik nicht umsetzbar sind, weil die Regeln oft zu abstrakt sind. Daher sind Beispiel-basierte Systeme oft leichter umsetzbar, falls Daten zur Verfügung stehen. Die Hand-Annotierte Daten sind meist sehr aufwendig zu erstellen, die Parallel-Korpora ist immer verbreiteter und auch leicht zugänglich. Bei Übersetzungen müssen mehrdeutige Konstruktionen durch den Übersetzer disambiguiert werden.

Sein Ziel war trainieren eines Klassifikators zur Disambiguieren eines Aspekt-Markers im Japanischen. Die Kategorien der Trainingsdaten sollen nicht selbst annotiert werden, sondern der jeweiligen Übersetzung entnommen werden.

Weiterhin ist Schmidhuber aufs Aspekt Begriff eingegangen. Aspekt ist eine grammatische Kategorie des Verbs, die die zeitliche Lage einer Situation ausdrückt. Im Japanischen ist das Aspekt Marker „"te-iru". Dieser Marker kann aber je nach Kontext unterschiedliche Aspekt ausdrücken: Verlauf oder Zustand als folge vorangegangenen Ereignisses. Im Englischen wird das Verlaufsform durch Progressive ausgedrückt, Zustand aber nicht.

Schmidhuber benutzte verschiedene Korpora und zwar: Wikipedia-Korpus (500.000 Sätze), Basic-Sentence-Korpus (5.000 Sätze) und „Wachturm" Ausgaben in Englisch und Japanisch.

Die Aufbereitung der Daten bestand aus Erstellung von Teilkorpora durch herausfiltern aller Sätze, die die „te-iru" Konstruktion nicht enthalten, Alignierung der Verben (ein Korpus bereits Hand-aligniert, in anderen Korpora mithilfe von Online-Wörterbüchern und parsen und bestimmen der Zeitform der englischen Verben (mithilfe einer Anwendung von Annemarie Friedrich).

Weiterhin teilte Schmidhuber die Daten in Trainings- und Testing-daten ein und wendete verschiedener Algorithmen zur Klassifikation an. Mithilfe der Testdaten hat er Evaluation der erreichten Genauigkeit bekommen.

Im Laufe der Arbeit sind weitere Probleme aufgetaucht: die Alignierung mit bekannter Alignierungssoftware (GIZA++, fast_align) lieferte für Sprach-Paare mit sehr unterschiedlicher Wortreihenfolge mit wenigen Daten nur sehr schlechte Ergebnisse (Alignierung von 500.000 Sätzen ergab nur bei 30% aller Wörter überhaupt eine Zuordnung). Als Alternativlösung wurden Wörterbücher benutzt, um die Zuordnung der Verben zu gewährleisten. Das zweite Problem war es, dass die Kategorien für den japanischen Aspekt-Marker nicht deckungsgleich mit Englischen Tenses waren.