Protokoll zur Sitzung vom 29.05.2017

Kurzvorstellung von der Bachelorarbeit von Korbinian Schmidhuber

BA-Betreuer: Annemarie Friedrich

Thema: Disambiguierung eines japanischen Aspekt-Markers mithilfe von Parallel-Korpora

Korbinian hat in der heutigen Sitzung seine Bachelorarbeit vorgestellt, in der es um Disambiguierung eines japanischen Aspekt-Markers mithilfe von Parallel-Korpora geht. Vor ein paar Wochen hat der Student seine Arbeit abgebrochen, deswegen bleiben noch einige Fragen offen. Bei vielen Methoden in der Computerlinguistik verwendet man Regelbasierte Systeme, die nicht immer umsetzbar sind, da Regeln oft zu abstrakt sind. Beispielbasierte Systeme sind daher oft leichter umsetzbar, falls genügend Daten in Verfügung stehen. Korbinian verwendet in seiner Arbeit Parallel-Korpora, weil sie verbreiteter und leicht zugänglich sind, als hand-annotierte Daten, die ihrerseits aufwendig zu erstellen sind und brauchen viel Zeit und Ressourcen. Eine wichtige Voraussetzung dabei ist, dass die ambigen, oder mehrdeutigen, Konstruktionen durch den Übersetzer disambiguiert werden müssen. Das Ziel Korbinians Bachelorarbeit ist, einen Klassifikator zur Disambiguierung eines Aspekt-Markers im Japanischen zu trainieren. Die Kategorien von Trainingsdaten werden nicht selbst annotiert, sondern der jeweiligen Übersetzung aus dem Parallel-Korpus entnommen. Unter Aspekt versteht man in der Linguistik eine grammatische Kategorie des Verbs, die die zeitliche Lage einer Situation ausdrückt (die vom Verb beschrieben wird). Der japanische Aspekt-Marker „te-iru“ kann zum Beispiel je nach Kontext unterschiedliche Aspekte ausdrücken: Verlauf oder Zustand als Folge eines vorangegangenen Ereignisses. Die Verlaufsform wird im Englischen durch das Continuous gebildet, ein Zustand aber nicht. Darin besteht der Hintergedanke der Arbeit, dass eine japanische Konstruktion nicht immer mit gleichen englischen Strukturen übersetzt wird.

Für das Experiment werden verschiedene Parallel-Korpora verwendet: Wikipedia-Korpus (Artikel über japanische Kultur), Basic-Sentences-Korpus (simple Sätze) und “Wachturm”-Ausgaben. Alle Daten sollten erst einmal aufbereitet werden. Als Erstes hat Korbinian Teil-Korpora erstellt durch Herausfiltern aller Sätze, die die „te-iru“-Konstruktion nicht enthalten. Danach hat der Student die Verben aligniert, d.h. die Wortordnung festgestellt, bei der ein einzelnes Wort genau mit einem Wort aus der anderen Sprache übersetzt wird. Ein Korpus war bereits hand-aligniert, bei den anderen hat Korbinian die Verben mithilfe von Online-Wörterbüchern übersetzt. Letztendlich hat der Student Parsen und Bestimmen der Zeitform der englischen Verben mithilfe einer Anwendung seiner Betreuerin durchgeführt. Die ganzen Daten wurden in Training und Testing Set aufgeteilt. Zur Klassifikation hat Korbinian verschiedene Algorithmen angewendet, die er in der Sitzung nicht erwähnt hat. Die erreichten Genauigkeiten hat der Student mithilfe der Testdaten evaluiert, hat aber keine konkreten Ergebnisse präsentiert.

Einige Probleme sind Korbinian bei der Arbeit aufgetreten. Alignierung mit bekannter Alignierungssoftware GIZA++ und fast_align haben für das Sprach-Paar Japanisch-Englisch sehr schlechte Ergebnisse geliefert: Alignierung von 500.000 Sätzen hat nur bei 30% aller Wörter überhaupt eine Zuordnung ergeben. Korbinians Meinung nach, es liegt daran, dass dieses Sprach-Paar sehr unterschiedliche Wortreihenfolge hat und der Student nur wenig Daten verwendet hat. Dafür spricht, dass Korbinian für Englisch-Deutsch mithilfe von GIZA++ sogar mit wenigen Daten sehr gute Ereignisse bekommen hat. Das andere Problem ist, dass die Kategorien für den japanischen Aspekt-Marker oft nicht deckungsgleich mit Englischen Tenses sind.