CIS, LMU München
Fach: Kolloquium Computerlinguistisches Arbeiten SS2017
Studentin: Faridis Alberteris Azar

Protokoll zur Sitzung am 29.05.2017

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

Student: Korbinian Schmidhuber
Betreuerin: Annemarie Friedrich

Die Motivation dieser Bachelorarbeit besteht darin, dass regelbasierte Systeme bei vielen Methoden in der Computerlinguistik nicht umsetzbar sind, da Regel oft zu abstrakt sind. Daher sind Beispiel-basierte Systeme oft leichter umsetzbar, falls Daten zur Verfügung stehen. Außerdem sind hand-annotierte Daten meist sehr aufwendig zu erstellen. Dagegen sind Parallel-Korpora leicht zugänglich und immer verbreiteter. Eine weitere Motivation ist es, dass bei Übersetzungen mehrdeutige Konstruktionen durch den Übersetzer disambiguiert werden müssen.

Das Trainieren eines Klassifikators zur Disambiguierung eines Aspekt-Markers im Japanischen ist das Ziel dieser Arbeit. Die Kategorien der Trainingsdaten sollen, laut Schmidhuber, nicht selbst annotiert werden, sondern der jeweiligen Übersetzung entnommen werden.

Der Student erklärt mit Hilfe von Beispielen, welcher der Hintergrund seiner Arbeit ist. Der Aspekt-Marker „te-iru“ in Japanischen kann je nach Kontext einen unterschiedlichen Aspekt ausdrücken:

1) „Verlauf“
   z.B.: Watasi-ha pan o tabe-te iru
   I-TOP bread-AKK eat-TE IRU-PRES
   engl.: *I'm eating bread*.

2) „Zustand als Folge eines vorangegangenen Ereignisses“
   z.B.: Inu-ha sin-de-iru
   Dog-TOP die-TE IRU-PRES
3) engl.: *The dog is dead* (und nicht: *The dog is dying!*)

Als Daten verwendet Scmidhuber Wikipwdia-Korpus, Basic-Sentences-Korpus und „Wachturm“ Ausgaben im Englisch und Japanisch. Er hat Teilkorpora durch Harausfiltern aller Sätze erstellt, die die „te-iru“ Konstruktion nicht enthalten. Er hat mit einem bereits hands-alignierten Korpus, die Verben in anderen Korpora mithilfe von Online-Würterbücher aligniert. Mit der Hilfer einer Anwendung seiner Tutorin Frau Friedrich, hat er die Zeitform der englischen Verben geparset und bestimmt.

Für die Klassifikation bevorzugt er, verschiedene Algorithmen anzuwenden. Dafür hat er die Daten in Trainigs- und Testingdaten eingeteilt. Die erreichte Genauigkeit wird mithilfe der Testdaten evaluiert.

Einige Probleme wurden gemerkt:

- Die Alignierung mit bekannter Alignierung-Software (z.B.: GIZA ++, fast_align) liefert für Sprach-Paare mit sehr unterschiedlicher Wortreihenfolge mit wenigen Daten nur schlechte

Ergebnisse.

- Die Kategorien für den japanischen Aspekt-Marker sind nicht Deckungsgleich mit englischen Tempora.

Der Student hat während der Präsentation erklärt, dass er momentan mit den Aufgaben der Bachelorarbeit aufgehört hat, deswegen präsentiert er keine konkrete Ergebnisse.