Protokolle zur Sitzung vom 29.05.2017 - Computerlinguistisches Arbeiten

31.05.2017

Drei Studenten stellten in der Sitzung am 29.05.2017 mit Hilfe einer Präsentation mit dem Beamer ihre Bachelorarbeitsthemen vor.

1. Präsentation:

Referentin: Korbinian Schmidhuber

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

Betreuer: Annemarie Friedrich

Ziel der Bachelorarbeit von Korbinian ist das Training eines Klassifikators zur Disambiguierung eines Aspekt-Markers im Japanischen. Dabei verwendet er japanische Korpora, mit Übersetzung in Englisch.
Korbinian beginnt die Vorstellung seiner Bachelorarbeit mit der Motivation. Dabei erklärt er, dass regelbasierte Systeme bei vielen Methoden in der Computerlinguistik nicht umsetzbar sind, da die Regeln oft zu abstrakt sind. Daher sind Beispiel-basierte Systeme oft leichter umsetzbar, falls die nötigen Daten verfügbar sind. Die Erstellung von hand-annotierten Daten ist trotzdem meist sehr zeit- und personalaufwändig, z.B. Wenn man 1000 Sätze im Japanischen annotiert, benötigt dies viel Zeit und geschultes Personal.
Daher ist die Benutzung von Parallel-Korpora sehr vorteilhaft, da diese immer verbreiteter und leicht zugänglich sind.
Danach erläutert Korbinian das genaue Ziel seiner Bachelorarbeit. Hierbei erklärt er zuerst, was in diesem Zusammenhang der Begriff “Aspekt” bedeutet. “Aspekt” ist neben dem Tempus eine Kategorie, die morphologisch eine wiederholte oder abgesschlossene Situation markiert. Dies entspricht bspw. dem Progressive im Englischen. Der Aspekt-Marker “te-iru” kann je nach Kontext einen unterschiedlidchen Aspekt ausdrücken. Korbinian zeigt hier das Beispiel für “Verlauf” und “Zustand” für den Aspekt-Marker anhand zweier japanischen Sätze, die auf englisch übersetzt wurden. Die entstehenden Probleme sind, dass die Kategorien für den japanischen Aspekt-Marker nicht deckungsgleich mit Englischen Tenses sind. Hier wird bspw. der Satz “Mr. Murata is sitting here” gezeigt. Im Japanischen gibt es das Verb “hinsetzen” nicht, sondern wird vom Zustand “sitzt”, als Folge von “sich hinsetzen” verstanden.
Die Verwendeten Daten kommen aus verschiedenen Parallel-Korpora: einem Wikipedia-Korpus mit ca. 50.000 Sätzen, einem basic sentence Korpus mit ca. 5.000 Sätzen und mehreren “Wachturm” Ausgaben in englisch und japanisch. Anschließend werden durch Herausfiltern aller Sätze, die die “te-iru” Konstruktion nicht enthalten, Teil-Korpora erstellt.. Danach erfolgt die Alignierung der Verben. Falls der Korpus nicht bereits hand-aligniert wurde, muss dies mithilfe von Wörterbüchern erfolgen. Im nächsten Schritt erfolgt das Parsen und die Bestimmung der Zeitform der englischen Verben, welches mithilfe einer Anwendung von Betreuerin Annemarie Friedrich erfolgt. Mittels Klassifikator werden die Daten in Trainings- und Test-Daten

eingeteilt. Korbinian wollte hier verschiedene Algorithmen zur Klassifikation ausprobieren, aufgrund des Abbruchs seiner Bachelorarbeit, wurde dies jedoch nicht gemacht.
Zum Schluss stellte Korbinian noch seine Evaluation und die Probleme dabei vor. Die Alignierung mit bekannten Alignierungssoftwares “GIZA++” und “fast_align” liefert mit wenigen Daten nur schlechte Ergebnisse für Sprach-Paare mit sehr unterschiedlicher Wortreihenfolge. Die Alignierung von 500000 Sätzen ergab nur bei 30% aller Wörter überhaupt eine Zuordnung.