PROTOKOLLE ZU COMPUTERLINGUISTISCHES ARBEITEN

Ines Röhrer
Centre for Information and Speech Processing, LMU
I.Roehrer@campus.lmu.de

1 Referat

Korbinians Bachelorarbeit mit dem Titel ”Disambiguierung eines japanischen Aspektmarkers mithilfe von Parallelkorpora”, betreut von Annemarie Friedrich, hat er leider schon abgebrochen, wodurch die arbeiten unvollständig bleiben werden.
Die ursprüngliche Motivation ist, dass regelbasierte Systeme in der Computerlinguistik oft schwer umsetzbar sind, da sie zu abstrakt vorgehen. Beispielbasierte Systeme sind oft leicher umsetzbar, falls für eine solche Vorgehensweise Daten vorhanden sind. Außerdem sind handannotierte Daten sehr aufwendig zu erstellen, und Parallelkorpora, in welchen für jeden Satz der ersten Sprache ein entsprechender Satz im Korpus der zweiten Sprache vorliegt, immer weiter verbreitet und leicht zugänglich.
Die Idee ist, dass Im Englischen die Verlaufsform durch das Progressive gebildet wird, und ein Zustand nicht durch das Progressive ausgedrückt werden kann.
Die Korpora, auf denen Korbinian gearbeitet hat, sind ein von Hand annotierter Wikipedia Korpus, jeweils auf Englisch und Japanisch, sowie ein Basic Sentences Korpus und der ”Wachturm” in der englischen und japanischen Ausgabe.
Seine Arbeit besteht aus mehreren Schritten, zuerst hat er aus den Korpora Teilkorpora erstellt, aus den Teilen, welche eine ”te iru” Konstruktion beinhalten. Mit diesen Teilkorpora hat er im weiteren Verlauf gearbeitet. Dann folgte die lignierung der Verben, wobei einer von drei Korpora bereits aligniert war, die anderen hat er selbst mit Online-Wörterbüchern aligniert, da eine bekannte Sofware nur sehr schlechte Ergebnisse bei einer automatischen Alignierung erzielt hatte.
UIm Anschluss parste er seine Daten mit Annemaries Software und wendete verschiedene Klassifikatoren an.
Die Evalurierung misst die erreichte Genauigkeit mithilfe der Testdaten.
Probleme hatte er vor allem mit der Alignierung, wegen den nicht verwendbaren Softwareergebnissen. Zusätzlich hat er noch Ausnahmen gefunden, wo sein ”te iru” Aspektmarker doch als Progressivform im Englischen ausgedrückt wird, wodurch sein System auf solche Fälle nicht mehr anwendbar ist.