Protokoll zur Sitzung vom 29.05.2017

Korbinian Schmidhuber, Betreuer: Annemarie Friedrich

"Disambiguierung eines japanischen Aspekt-Markers mithilfe von Parallel-Korpora"

Korbinian startet seinen Vortrag mit ein paar Worten zur Motivation für seine leider abgebrochene Bachelorarbeit. Regelbasierte Systeme für das Japanische sind kaum umsetzbar, da die Regeln zu abstrakt sind und somit rücken Beispiel-basierte Systeme als Lösung in den Fokus, falls hinreichende Mengen von Daten für diese Sprache existieren. Von Hand annotierte Daten sind aber rar, da sie aufwendig und teuer sind und ausreichend geschultes Personal erfordern. Die Lösung mit Parallel-Korpora ist im Vergleich deutlich einfacher zugänglich. Dr. Schulz argumentiert, dass Parallel-Korpora alleine nicht ausreichen dürften und eine Annotation vonnöten sei. Mehrdeutige Konstruktionen müssten schließlich genau interpretiert sein. Ziel der Arbeit ist es nun, einen Klassifikator zu trainieren, der japanische Aspekt-Marker disambiguiert. Bei dem von Korbinian gewählten Aspekt-Marker handelt es sich um eine Tempus-Kategorie, die verschiedene Verhältnisse und Situationen morphologisch darstellt. Es ist in einem bestimmten Sinn vergleichbar mit dem englischen Progressive. Diese Kategorie wird aus der Übersetzung entnommen. Beispiele für den Aspekt-Marker "te-iru" werden dargestellt: der Aspekt kann entweder einen Verlauf darstellen ("Ich esse *gerade*") oder einen Zustand als Folge einer vorhergegangenen Handlung (hier nennt Korbinian einen Satz, der sinngemäß übersetzt "Der Hund ist tot" heißt, obwohl das Wort "sterben" als Verb im Einsatz ist, da hier die Folge von "sterben" "tot sein" bedeutet). Im Englischen, der Vergleichssprache, die für das Projekt gewählt wurde, gibt es nur den Verlauf und keinen Zustand als Folge. Die Korpora, die hier in Verwendung sind, sind zum einen ein Wikidump (50.000 Samples) mit Artikeln zu japanischer Kultur, zum anderen ein Basic-Sentences Korpus mit 5000 einzigartigen Satzbeispielen. Außerdem die englische und japanische Version vom Wachturm, da dieser gratis online zur Verfügung steht. Daraus werden Teil-Korpora erstellt, die nur Satzkonstruktionen mit te-iru beinhalten. Jetzt wird versucht, eine Alignierung zu erzielen, das bedeutet, dass ein japanisches Wort mithilfe von Online-Wörterbüchern einem englischen Wort oder Begriff zugeordnet wird. Dann wird mithilfe eines Systems von Annemarie Friedrich die englische Zeitform geparsed, um die Daten für den Klassifikator aufzubereiten. Hier soll anhand des Train-Set mit verschiedenen Algorithmen experimentiert werden, um das Ergebnis anschließend mit dem Test-Set auf seine Accuracy zu testen. Dr. Schulz möchte hier wissen, ob solche Aspekt-Marker wirklich rein objektiv annotiert werden können oder ob da nicht die Gefahr von verschiedenen Interpretationsmöglichkeiten durch verschiedene Menschen gibt. Korbinian erklärt, dass Übersetzungen durch Muttersprachler doch recht eindeutig sind und dass es hierbei gar nicht um die Annotation per Hand gehe, sondern, dass die Arbeit die bereits vorhandene englische Annotierung als Umweg zur japanischen Annotierung sieht. Als nächstes werden Probleme, die bei der Bearbeitung aufgetreten sind, erläutert. Die beiden verwendeten

Tools GIZA++ und fast_align liefern beim Sprachpaar Deutsch und Englisch sehr gute Ergebnisse, bei Japanisch und Englisch jedoch nur schlechte. So findet bei 500.000 Proben nur bei 30% eine Alignierung statt und diese erweist sich oft auch noch als falsch. Dr. Schulz fragt, ob eine Alignierung der Verben für diese Arbeit ausreicht und Korbinian bejaht dies. Als weitere Fehlerquelle nennt er die Tatsache, dass der Aspekt-Marker schließlich nicht deckungsgleich mit dem Englischen ist.