Klyueva Olga
CIS, LMU München
29.05.1917

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

Als erstes wurde über Motivation für die Bachelorarbeit gesprochen. Regelbasierte Systeme sind schwer umsetzbar, weil die Regeln nicht vorhanden sind oder zu abstrakt sind für Anwendung. In der Bachelorarbeit geht es um ein beispielbasiertes System. Man trainiert das System anhand der Beispiele. Das System lernt, welche Bedeutung ein Aspekt-Marker hat.
Parallel Korpora ist Japanisch-Englisch, wo es für jeden japanischen Satz eine Übersetzung ins Englische gibt.
Hand-annotierte Daten sind sehr zeit- und kostenaufwendig. Es ist leichter, wenn man Parallel Korpora hat. Die Grundidee ist, dass man Tenses aus den englischen Sätzen extrahiert. Anhand der Tenses werden die Aspekte rausgelesen und als Annotation benutzt.
Das Übersetzungsprogramm soll mehrdeutige Worte auf eine Bedeutung festlegen. Das Ziel der Bachelorarbeit war einen Klassifier zu trainieren, der eine Disambiguierung eines Aspekt Markers im Japanischen ergibt. Aspekt des Verbes in vielen Sprachen wird morphologisch realisiert und drückt die zeitliche Lage einer Situation aus. Progressive Tense bespielweise drückt den Verlauf aus. Es gibt Aspekte, die ein abgeschlossenes Ereignis ausdrücken.
Im Deutschen wird ein Aspekt morphologisch nicht markiert, sondern mit Hilfe bestimmter Worte, bespielweise mit dem Wort "gerade". Deswegen wurde in der Bachelorarbeit englische Sprache ausgewählt.
In Rahmen der Bachelorarbeit wurde der Aspekt-Marker "te-iru" ausgewählt, wobei -te ein Suffix ist und –iru ein Hilfsverb ist, der normalerweise das Verb "sein" ausdrückt. Es wurden einige Beispiele gezeigt, wo mit unterschiedlichen Kontext ein unterschiedlicher Aspekt ausgedrückt wird: Verlauf oder ein Zustand als Folge eines vorangegangenen Ereignisses.

Im Rahmen der Bachelorarbeit wurden Parallelkorpora benutzt:

- Wikipedia-Korpus (Wikipedia Artikel zum Thema Japanische Kultur) bestehend aus 500 000 Sätzen
- Basic Sentences-Korpus (5 000 Sätze)
- „Wachturm" Ausgaben in Englisch und Japanisch

Aufbereitungen der Daten wurden folgendermaßen gemacht. Zuerst wurde ein Teil-Korpus erstellt, die die ti-iru Konstruktion enthalten.
Zweiter Schritt ist die Alignierung der Verben, eine Zuordnung von Wörtern, die es in Parallel-Korpora gibt: ein Wort auf Japanisch wurde mehr oder weniger wörtlich auf Englisch übersetzt. Es ist nicht immer eindeutig. Manchmal kann ein Wort einer Sprache auf verschiedene Wörter einer anderen Sprache abgebildet werden und andersrum verschiedene Wörter einer Sprache auf ein Wort anderer Sprache abgebildet werden. Es war notwendig die Sätze mit mehrdeutigen Konstruktionen im Japanischen mit der te-iru Konstruktion eindeutig auf Englisch zu bestimmen.
Nächster Schritt war das Parsen und die Bestimmung der Zeitform der englischen Verben. Es wurde mit einer Software gemacht.
Anhand Trainingsdaten wird der Klassifikator trainiert. Es wurde vorgesehen, verschiedene Algorithmen für Klassifikation zu verwenden.
Evaluation der erreichten Genauigkeit wird mithilfe der Testdaten gemacht. Es wird analysiert, wie oft der Klassifier die richtige Entscheidung getroffen hat.

Am Schluss wurden Probleme erwähnt:

- Alignierung mit bekannter Alignierungssoftware. Die Wortstellung im Japanischen und im Englischen ist sehr unterschiedlich, beispielsweise ein Japanisches Prädikat steht immer am Schluss. Von 500 000 Sätzen wurde ein Drittel aligniert, teilweise auch falsch.

- Kategorien für den japanischen Aspekt-Marker sind nicht deckungsgleich mit Englischen Tenses. Es wurden Beispiele gezeigt: ein Zustand als Folge eines vorangegangenen Ereignisses wird nicht mit Progressive Tense übersetzt, wie man es erwartet.