Vortrag: „Disambiguierung eines japanischen Aspekt-Markers mithilfe von Parallel-Korpora “ Korbinian Schmidhuber Betreuerin : Annemarie Friedrich

Die Motivation der Arbeit besteht darin, dass die regelbasierte Systeme nicht so gut sind und die Regelschreibung oft intuitiver Prozess ist, da die Regel nicht vorhanden sind oder zu abstrakt sind zu umsetzen.

Als Alternative wird ein Beispiel-basiertes System vorgeschlagen, was leicht umsetzbar wäre, falls die Daten verfügbar sind, um das System zu trainieren. Dafür spricht auch, dass die handannotierte Daten sehr aufwendig sind zu erstellen, da man erstens die Leute braucht, die in dem Bereich gut sind und sich auskennen und zweitens solche Daten nicht fehlerfrei sind. Andererseits werden die Parallel-Korpora immer verbreiteter und zugänglicher.
Die Ergebnisse der Arbeit können gut im Bereich der Übersetzung verwendet werden. Da bei der Übersetzung die mehrdeutige Konstruktionen durch der Übersetzer disambiguiert werden müssen. Es können somit nicht mehrere Möglichkeiten existieren, sondern muss nur eins ausgewählt werden.
Das Ziel der Arbeit ist einen Klassifikator zur Disambiguierung eines Aspekt-Markers in Japanischen auszuarbeiten. Dabei sollen die Kategorien der Trainingsdaten nicht selbst annotiert werden, sondern aus der Übersetzung (in dem Fall Parallel-Korpus) entnommen werden.

Weiterhin ist Korbinian aufs Begriff des Aspekts eingegangen. Aspekt ist eine grammatische Kategorie des Verbs, die die zeitliche Lage einer Situation ausdrückt. Im Japanischen ist das Aspekt Marker „“te-iru“.
Dieser Marker kann aber je nach Kontext unterschiedlichen Aspekt ausdrücken: Verlauf oder Zustand als Folge vorangegangenen Ereignisses.
Im Englischen wird das Verlaufsform durch Progressive ausgedrückt, Zustand aber nicht.

Für die Arbeit wurden folgende drei Parallel-Korpora verwendet: Wikipedia Korpus (500000 Sätze), Basic-Sentences-Korpus (5000 einfache Sätze) und englische und japanische Ausgaben von „Wachstum“.

Die Ausarbeitung besteht aus folgenden Schritten: Erstellung von Teilkorpora durch Herausfiltern aller Sätze ohne „te-iru“; Alignierung der Verben; Parsen und Bestimmen der Zeitform der englischen Verben. Weiterhin probiert man unterschiedliche Algorithmen aus und evaluiert mit Hilfe von Testdaten.

Im Laufe der Arbeit sind weitere Probleme entstanden: die Alignierung mit GIZA++, fastalign liefert für Sprachpaare mit sehr unterschiedlicher Wortreihenfolge mit wenigen Daten nur sehr schlechte Ergebnisse (von 5000000 Sätzen ergab nur bei 30% aller Wörter überhaupt eine Zuordnung). Als Alternativlösung wurden Wörterbücher benutzt, um die Zuordnung der Verben zu gewährleisten.
Das zweite Problem ist, dass die Kategorien für den japanischen Aspekt Marker nicht deckungsgleich mit Englischen sind.
Die Bachelorarbeit wurde vor zwei Wochen abgebrochen, somit gibt es keine Ergebnisse.

Vortrag: „Regularization of Neural Networks for Natural language Processing“ Dayyan Smith Betreuerin : Katharina Kann

The goal of this thesis is exploring the effect of regularization of a neural network for stance classification of news articles.
Dayyan presented first Fake News Challenge that exploring how artificial intelligence technologies could be leveraged to combat fake news.

Fake news means a made-up story with an intention to deceive. The detection of fake news is a complex and cumbersome task, that why Automatic Fake News Detection can be broken down into stages.

The first stage is Stance detection. The possible stances are agree, disagree, discuss and unrelated.

The vectors for each words should be constructed with the help of word2vec. Headlines and bodies go separated through two hidden layers, than are concatenated and go through classification layer to get the label.

Regularization of a data is import to get better results. It is alternative, if it is not possible to increase data you have. This is a technique used in an attempt to solve the overfitting problem in statistical models.

There are some different ways to regular the data. The first one is L2 regularization. Here the big weights are pushed down more than small weights because the square of weights is penalized. In the other one (L1) both big and small weights are pushed down a little because the absolute value is penalized. This often drives small weights to zero. The third method is dropout regularization: to drop out certain number of neurons (circa 50%), than restore neurons and drop out other neurons. In this way neurons learn to detect what is matter. This should help cut the noise out.

The goal is with the help of regularization get better score than baseline score. (Baseline system is SVM system).

The results of evaluation is almost the same for all the types of regularization.

Vortrag: „Corpus based identification of text sequences“ Thomas Ebert Betreuer : Martin Schmidt

Motivation der Arbeit besteht darin, dass die Textsegmentierung zentrale Rolle für NLP Aufgaben spielt. Unter Textsegmentierung versteht man Textteilung nach bedeutungstragenden Einheiten, was wiederum unterschiedlich interpretiert sein kann: auf Wort- , Morphem-, Satz-, Phrase- oder Topicebene. Die meistverbreitende Version ist Tokenisierung.

Der Tokenisierungsprozess ist an sich sehr fehleranfällig, da Definition des Wortes intuitiv ist. Es werden somit die lokale Anpassungen nötig.

Die Frage, die in dieser Arbeit gestellt wird, ist ob das intuitive Konzept „Wort“ die beste Art für einen Computer einen Text zu segmentieren ist.
Das Ziel der Arbeit ist ein Algorithmus zu entwickeln, der einen eingegebenen Satz oder Text in seine „beste“ Segmente (Buchstaben, N-Gramme) zerlegt. Dabei will man versuchen die Frage zu beantworten, ob der nicht-symbolische Ansatz besser als der wortbasierte Ansatz ist und welche Risiken und Chancen der nicht symbolische Ansatz mit sich bringt.

Erstens extrahiert man die N-Grammen der Länge 1 bis 10 aus dem Wikipediakorpus (10000 unannotierte englische Texte, die 22650880 Zeichen ergeben). Als zweiter Schritt wurde die Frequenzliste erstellt und mit Gütemaß bewertet (N-grammlänge * log(freq)). Zum Testen wird ein Satz eingegeben, der in der N-gramme mit den höchsten Gütemaß zerlegt wird.

Im Laufe der Arbeit tauchen aber einige Probleme auf: mit der Größe der Eingabe steigt die Laufzeit exponentiell. Die vorgeschlagene Lösung ist die Größe des Fensters zu beschränken und somit festzulegen (heuristischer Ansatz). Anderes Problem besteht darin, dass durch die Festlegung der Fenstergröße die Berechnung der höchsten Güte nicht mehr garantiert werden kann. Trotzdem zeigen die Ergebnisse, dass die Segmentierung besser als bei symbolischem Ansatz ist.

Evaluierung des Systems ist etwas schwer, da keine Goldstandart für Textsegmente zur Verfügung steht. Es entstehen häufig Uneinigkeit über die Granularität von Segmenten. Je nach Anwendung können Fehler relevant oder irrelevant sein. Deswegen versucht man bei der Evaluierung die Auswirkung auf die Endanwendung zu messen (z. B. IR, Sentiment Analysis). Dafür verwendet man word2vec um Buchstaben N-Gramm embeddings zu erhalten. Die Evaluation wurde auf Satzebene durchgeführt mit der Verwendung von Movie Review Data mit dem Vergleich mit Word embeddings.

Die Evaluationsergebnisse sind leider noch nicht vorhanden. Als offene Frage bleiben noch: ob es andere Möglichkeiten gibt, die N-gramme zu extrahieren, als Beispiel wurde das Programm von H. Schütze erwähnt.