Disambiguierung eines japanischen Aspekt-Markers mithilfe von Parallel-Korpora

Vortragender: Korbinian Schmidhuber Betreuerin: Annemarie Friedrich

Den ersten Vortrag hält Korbinian Schmidhuber. Er erklärt, dass er vor einigen Wochen seine Bachelorarbeit abgebrochen hat und deswegen nur den Ansatz erklärt und was er bis dahin geschafft hat. Als Hintergrund erklärt er, dass in der Computerlinguistik regelbasierte Systeme für viele Anwendungen schlecht umsetzbar sind und daher beispiel- bzw. statistikbasierte Systeme oft deutlich einfacher sind, falls genügend Daten verfügbar sind. Dies gilt auch für maschinelle Übersetzung, womit sich seine Arbeit beschäftigt. Hand-annotierte Daten sind meist sehr zeit- und kostenaufwendig, während parallele Korpora immer verbreiteter und einfacher zugänglich sind. Bei maschinellen Übersetzungen müssen allerdings mehrdeutige Konstruktionen vor dem Übersetzen disambiguiert werden. Deshalb ist das Ziel von Korbinians Bachelorarbeit das Trainieren eines Klassifikators zur Disambiguierung eines Aspekt-Markers im Japanischen. Die Kategorien sollen also nicht in den Daten annotiert werden, weil annotierte Daten teuer sind, sondern automatisch aus der jeweiligen Übersetzung entnommen werden. Ein Aspekt ist eine grammatikalischer Marker zusätzlich zu Zeit. Zum Beispiel in Englisch wird mit dem progressive markiert dass etwas gerade passiert. Im Deutschen wird Aspekt kaum mit Morphologie markiert, sondern durch Wörter wie „gerade“, weswegen in der Arbeit auch die Übersetzung von Japanisch zu Englisch und umgekehrt betrachtet wird. Im Japanischen gibt es den Aspektmarker te-iru, der je nach Kontext unterschiedliche Aspekte ausdrücken kann, wie zum Beispiel Verlauf oder Zustand als Folge eines vorangegangenen Ereignisses. Die Idee der Arbeit ist, dass im Englisch die Verlaufsform durch das progressive markiert wird, während ein Zustand nicht durch das progressive ausgedrückt werden kann. Deshalb könnte man die parallelen Korpora verwenden, um die te-iru Konstruktion im Japanischen zu disambiguieren. Für die Arbeit werden verschiedene Korpora verwendet: der Wikipediakorpus, der basic sentences Korpus und der Wachturm, weil er frei im Internet in Englisch und Japanisch verfügbar ist. Aus diesen Korpora wurden Teilkorpora erstellt, indem alle Sätze ohne te-iru Konstruktion herausgefiltert wurden. Danach wurden die Verben mithilfe von Online-Wörterbüchern aligniert. Dann wurden die englischen Verben geparst und ihre Zeitform mithilfe einer Anwendung von Frau Friedrich bestimmt. Für den eigentlichen Klassifikator wurden die Daten dann in Trainings- und Testdaten aufgeteilt und verschiedene Algorithmen zur Klassifikation ausprobiert. Die erreichte Genauigkeit wird mithilfe der Testdaten evaluiert. Zu den aufgetretenen Problemen gehört, dass die Alignierung mit den zwei Softwares GIZA++ und fast_align für Sprachpaare mit sehr unterschiedlicher Wortreihenfolge wie Englisch und Japanisch mit nur wenigen Daten sehr schlechte Ergebnisse liefert. Bei der Alignierung von 500.000 Sätzen gab es nur zu 30% der Wörter überhaupt eine Zuordnung. Ein anderes Problem ist, dass die Kategorien für den te-iru-Aspektmarker nicht deckungsgleich mit den englischen Tenses ist.

Regularization of Neural Networks for Natural Language Processing

Vortragender: Dayyan Smith Betreuerin: Katharina Kann

Dayyan’s bachelor thesis is about exploring the effect of regularization of a neural network for stance classification in the context of fake news detection. Fake news, as defined by the New York Times, is a made-up story with an intention to deceive. Dayyan explains that assessing the veracity of a news story is a complex and cumbersome task, even for trained experts. Automatic fake news detecting as also complex but cam be broken down into several stages. The “fake news challenge”, which is a challenge that aims at exploring how artificial intelligence technologies could be leveraged to combat fake news, states that as a first step, it would be helpful to know which news organizations agree with

a given claim. Therefore, the first stage of the challenge is stance detection. Stance detection is the task of finding the stance of an article towards any headline where possible stances are agree, disagree, discuss and unrelated. For this, he uses pretrained word embeddings from word2vec for both headline and body and puts them through a Recurrent Neural Network. He then concatenates the ouput of the hidden layers with the headline and body embeddings. According to Dayyan, one shouldn't be impressed when a complex model fits a data set well because with enough parameters, you can fit any data set, which is why you need a model to perform well on unseen data, which can be done with regularization. The different regularization methods used are L1-Regularization, L2-Regularization and Dropout Regularization. In L2-Regularization, big weights are pushed down more than small weights because the square of weights is penalized while in L1-Regularization, big and small weights are pushed down a little because the absolute value is penalized. Dropout regularization is different because here, you don't change the cost function but rather modify the net itself where you drop some neurons. He then presented the results of his evaluation, which were that only some of his models made it past the baseline. He wants to hand in those models to the fake news challenge.

Corpus based identification of text segments

Vortragender: Thomas Ebert Betreuer: Martin Schmitt

In der Bachelorarbeit von Thomas Ebert geht es um Textsegmente. Ein Textsegment ist eine bedeutungstragende Einheit und kann daher alles von einem Morphem bis zu einem Topic, also einem ganzen Absatz, sein. Die Textrepräsentation in NLP ist meistens wortbasiert, wobei die Einteilung in Wörter Tokenisierung heißt, welche aber sehr fehleranfällig ist. Es sind Anpassungen an die Sprachen nötig, weil verschiedene Sprachen ein sehr unterschiedliches Konzept eines Worters haben. Außerdem ist das Konzept Wort zwar für den Menschen intuitiv, aber nicht eindeutig definiert. Die zentrale Frage der Arbeit ist daher, ob das Konzept eines Wortes die beste Art für einen Computer ist. Deswegen entwickelt er einen Algorithmus, der einen Satz in seine besten Segmente zerlegt.

Zuerst werden N-Gramme der Länge 1 bis 10 aus dem Wikipedia Korpus Englisch extrahiert. Der Korpus besteht aus unannotierten Rohtexten, von denen die ersten 10.000 Texte verwendet wurden, was 22 Mio. Zeichen entspricht. Für die N-Gramme wurde eine Frequenzliste erstellt und danach wurden alle N-Gramme mit einem Gütemaß bewertet. Das Gütemaß ist n* log(freq), wobei n die N-Gramm-Länge ist und freq die absolute Häufigkeit des N-Gramms. Der eingegebene Satz wird dann in die N-Gramme der höchsten Gütemaße zerlegt.

Probleme sind hierbei, dass mit der Größe der Eingabe die Laufzeit exponentiell steigt. Eine mögliche Lösung ist ein heuristischer Ansatz, bei der eine Window-Größe festgelegt wird, mit dem über den Text gegangen wird. Die Berechnung der höchsten Güte ist dann nicht mehr garantiert, aber die Segmentierung ist gegebenenfalls trotzdem besser als ein rein symbolischer Ansatz.

Die Evaluation von Textsegmenten ist sehr schwierig, weil häufig Uneinigkeit über die Granularität von Segmenten herrscht. Je nach Anwendung können Fehler relevant oder irrelevant sein. Beispielsweise kann bei IR die Korrektheit von Segmentgrenzen teilweise vernachlässigt werden, bei news boundary detection nicht. Trotzdem wird hier die Endanwendungen wie IR oder Sentiment Analysis als Maß verwendet. Zur Evaluation benutzt er word2vec, um die Buchstaben-N-Gramm embeddings zu erhalten. Dann wird Sentiment Analysis auf Satzebene zur Evaluation verwendet und mit den Systemen mit normalen Word Embeddings verglichen.

Eine aus der Arbeit gewonnene Erkenntnis ist, dass auch Buchstaben-N-Gramme eine Zipf’sche Verteilung aufweisen. Darüber hinaus enthalten die häufigsten n-Gramm e größer 3 Funktionswörter und die häufigsten N-Gramme größer 8 oft Inhaltswörter. Die Evaluationsergebnisse sind noch nicht vorhanden. Weitere Fragen sind, ob es eine andere Möglichkeit gibt, um N-Gramme zu extrahieren und ob das Ergebnis der Evaluation schon aussagekräftig ist, obwohl die Methode nur auf eine Task angewendet wurde.