Signifikanz Tests für Experimente in NLP

Eine wichtige Frage bei neu entwickelten Systemen und Modellen ist: „Wie funktioniert das System auf unbekannten Daten“.
Wie in den meisten Tasks, dieser Art, benötigt man verschiedene Sets. Ein Dev-, Train-, und Test-Set. Das Test-Set stellt uns hier die neuen Daten bereit. Mit diesen lässt sich die Performanz eines Systems feststellen. Oft hat man mehrere Methoden die man vergleichen möchte. Dazu wurden mehrere Testarten vorgestellt: Zum einen der Paired t-Test, der die „Scores“ vergleicht. Der Sign Test, der mit hilfe eines Binominaltests durchgeführt wird und der Randomized Test. Zusätzlich gibt es noch verschiedene Metriken, die man verwenden kann. Zum Beispiel die „Accuracy“ oder die Mean Average Precision (MAP).

Disambiguierung eines japanischen Aspekt-Markers mithilfe von Parallelen Korpora

Der erste Vortrag des Tages kam von Korbinian Schmidhuber.
Die Betreuerin der Arbeit ist Annemarie Friedrich. Er begann damit zu erläutern, das Regelbasierte Systeme in der Computerlinguistik oft schwer umsetzbar sind, da die Regeln oft viel zu abstrakt oder nicht intuitiv ersichtlich sind. Beispiel-basierte Systeme sind dagegen besser umzusetzen. In seiner Arbeit wurden Parallel Korpora verwendet, da diese recht weit verbreitet sind und nicht so aufwendig zu erstellen sind, wie von Hand annotierte Korpora. Das Ziel war es einen Klassifikator, zur Disambiguierung eines Aspekt-Markers im Japanischen, zu trainieren. Der Aspekt ist in manchen Sprachen, neben dem Tempus eine Kategorie, die morphologische Verhältnisse in einer Situation markiert. Der Aspekt-Marker in diesem Fall ist „te-iru“. Man könnte sagen, es drückt einen Zustand, als Folge eines Vorangegangenen Ereignisses aus. Die Daten kamen einmal von einmal von Wikipedia Korpora, einem „Basic Sentences Korpus“ und dem Wachturm Ausgaben im Englischen und Japanischen. Alle Sätze, die den Aspekt-Marker nicht enthalten werden herausgefiltert. Des Weiteren wurden die Verben aligniert. Im Nächsten Schritt kam dann das Parsen und die Bestimmung der Zeitformen der englischen Verben. Die Software dafür kam von der Tutorin. Die Daten sollten dann in Trainings und Testdaten aufgeteilt werden. Mit den Trainingsdaten konnte dann der Klassifikator trainiert werden. Es sollten auch verschiedene Algorithmen angewendet werden. Ebenso steht auch noch eine Evaluation der Resultate an, jedoch hatte Korbinian die Arbeit bereits abgebrochen. Probleme die während der Arbeit auftraten, waren zum Teil die

Schlechten Ergebnisse der Alignierung der Verben. Zudem sind die Kategorien der Aspekte des Japanischen nicht deckungsgleich mit den Englischen Zeitformen.

Regularistaion of Neural Networks for Natural Language Processing

Der zweite Vortrag an diesem Tag wurde von Dayan Smith gehalten. Seine Betreuerin ist Katharina Kann. In seiner Arbeit geht es um die „Stance“-Klassifikation, im Kontext mit „Fake-News“. Fake-News sind in der heutigen Zeit ein großes Problem, da diese falschen Informationen meist frei verbreitet werden können. Die Erkennung von Fake-News ist sehr schwierig und anstrengend, sogar für Experten auf dem Gebiet. Um den Fake-News entgegen zu wirken, gibt es eine „Fake-News Challenge“. Die Fake-News Challenge versucht Technicken, zum Beispiel Küntliche Intelligenz, zu entwickeln mit denen man die Fake-News bekämpfen kann. Dayan sagte, dass es hilfreich wäre zu wissen, welche Organisationen einer Nachricht zustimmen. In der automatischen Erkennung von Fake-News gibt es mehrere Schritte, die durchgeführt werden. Der erste Schritt ist die „Stance“ Erkennung. Stance steht für die Haltung zu etwas. In diesem Fall, wie ist die Haltung eines Artikels zu einer „Headline“. Die Haltungen können: agree, disagree, discuss, unrelated sein. Er verwendet vorher trainierte „Word2Vec Word Embeddings“ um die Vektoren für jedes Wort zu initialisieren (Alle Wörter in der Headline und in dem Artikel). Die Vektoren werden mithilfe eines „Recurrent Neural Net“ Repräsentiert, um sie später Klassifizieren zu können. Die Repräsentationen der Headline und des Artikels werden Verbunden. Dann werden diese Embeddings durch zwei „Hidden Layers“ geführt und der Output, der jeweils entsteht wird wieder angehängt. Dieses Konstrukt wird dann vom Klassifikator behandelt. Um das System auch auf ungesehene Daten vorzubereiten, verwendet er Regularisierung. In diesem Fall eine

Methode, die „Drop Out“ genannt wird. Bei dieser Methode werden einzelne Neuronen eines Neuronalen Netzes herausgenommen. Dies wird in verschiedenen Kombinationen durchgeführt. Die Resultate waren in allen Fällen sehr ähnlich.

Corpus Based Identification of Text Segments

Der letzte Vortrag des Tages war von Thomas Ebert. Sein Betreuer ist Martin Schmidt. Als erstes hatte Thomas über Textsegmente gesprochen, und darüber, wie diese aussehen können. Je nachdem wie man die Granularität ansetzt, können solche Textsegmente nur die Größe eines Morphems haben, oder so groß wie ein ganzer Satz sein. Es geht um die Textaufbereitung für Aufgaben aus dem Bereich der Natural Language Processing und die ist meistens Wortbasierend. Die Tokenesierung ist auch meistens sehr Fehleranfällig und es sind lokale Anpassungen nötig. Zudem ist es streitbar, ob Wörter die beste Segmentierungseinheit sind. In dieser Arbeit will man also einen Text in passende Segmente zerlegen. Der Text kann also in unterschiedlich große Buchstaben n-gramme zerlegt werden. Aus dem Text werden also alle möglichen N-Gramme der Größen 1-10 extrahiert. Das verwendete Korpus ist ein englisches Wikipedia Korpus. Aus den N-Grammen werden dann Frequenzlisten erstellt. Die N-Gramme werden dann mit einem Gütemaß bewertet. Dieses setzt sich aus der Länge der N-Gramme *(mal) dem Logarithmus, von der absoluten Häufigkeit des Grammes, zusammen. Wenn also ein Satz eingegeben wird, so wird er in die N-Gramme zerlegt, die das höchste Gütemaß ergeben. Ein großes Problem ist, das die Laufzeit, mit der Größe des Textes exponentiell ansteigt. Die Evaluierung ist ein wenig schwierig, da man sich nicht über die Granularität der Segmente einig ist. Es wurde „Word2Vec“ verwendet um Buchstaben N-Gramme embeddings zu erhalten und es wurde eine „Sentiment Analyse“ auf Satzebene durchgeführt? Die Ergebnisse sind jedoch noch nicht vorhanden. Interessant ist jedoch, dass n-Gramme, die

größer als 3 sind, oft Funktionswörter sind. N-Gramme, die größer als 8 sind, sind meistens Inhaltswörter.