Repetitorium Statistische Signifikanztests

In der Untersuchung von Machine Learning Systemen kann es zu unterschiedlichen Formen der Evaluation kommen. Die Datenbasis für ML-Experimente wird aufgeteilt in Training Set, Development Set und Test Set. Trainingsdaten dienen als Input des Systems, während Developmentdaten für die justierung von Hyperparametern dienen sollen. Idealerweise wird auf dem Testset nur einmal getestet um endgültige Ergebnisse zu bekommen, allerdings nicht weiter optimiert.

Hat man mehrere Systeme miteinander verglichen, stellt sich die Frage ob eine Abweichung in der Accuracy oder dem F-Score signifikant ist, das heißt, ob die unterschiedlichen Ergebnisse der Systeme repräsentativ für weitere Datensätze sein können. Hierzu wird auf Signifikanztests zurückgegriffen. Beispielsweise gibt es hierbei den Binomialtest oder den paired t-test über der Normalverteilung. Bei diesen Tests wird die Abweichung von einem hohen Prozentsatz der Wahrscheinlichkeitsmasse als Maß genommen.

Korbinian Schmidhuber

Disambiguierung eines japanischen Aspekt-Marker mithilfe von Parallel-Korpora

Betreuerin: Annemarie Friedrich

Die Arbeit beschäftigte sich mit einem Problem aus der maschinellen Übersetzung von morphologisch stark unterschiedlichen Sprachen. In der Übersetzung irregulärer oder ambiger Satzstrukturen sind regelbasierte Systeme schwer umsetzbar, da die Aufschlüsselung durch Sprecher sehr abstrakt und mit viel Intuition passiert. Obwohl beispielbasierte Systeme besser umsetzbar sein können erfordern sie ein hohes Maß an annotierten Daten, die aufwendig zu erstellen sind. Parallel-Korpora, in denen ein Text in zwei oder mehr unterschiedlichen Sprachen vorliegt, sind dafür weiter verbreitet und oft offen zugänglich.

Ziel der Arbeit ist es, einen Klassifikator zur Desambiguierung des Aspekt-Marker „ている“ aus dem Japanischen zu finden. Der Aspekt ist eine sprachliche Kategorie die morphologisch ausgedrückt wird. Als Beispiel nannte Korbinian das Progressive im Englischen. Im Japanischen zeigt das te-iru in unterschiedlichen Kontexten unterschiedliche Grade der Abgeschlossenheit einer Handlung. Beispielsweise steht 犬は死んでいる für eine abgeschlossene Handlung. Der Hund (犬) stirbt mit abgeschlossenem Aspekt, also ist bereits tot. In der Arbeit sollte für diese grammatische Konstruktion aus englischen Sätzen der Verlaufsform die korrekte Anwendung gelernt und angewandt werden.

Die Datenbasis zur Arbeit lieferte ein Parallel-Korpora handübersetzter Daten aus Wikipedia zum Thema japnischer Kultur. Hinzu kommt ein Basic-Sentence Korpus aus einfachen Sätzen im japanischen und deren englischer Übersetzung. Dritte Quelle waren Ausgaben der Zeitschrift Wachturm, die offen zugänglich auf Englisch und Japanisch verfügbar sind. Zur Aufbereitung der Daten wurden Teil-Korpora durch Herausfiltern von Sätzen ohne te-iru Konstruktion erstellt und im zweiten Schritt unter Anwendung der Software GIZA++ und fast_align aligniert. Im dritten Schritt wurden die Sätze im englischen geparst um die Zeitform zu bestimmen.

Für den Klassifikator wurden die Daten in Trainings- und Testset eingeteilt um dann mit unterschiedlichen Algorithmen zu klassifizieren. Da Korbinian seine Bachelorarbeit bereits abgebrochen hat gibt es zu diesen Experimenten keine genaueren Überlegungen mehr.

Einige der bisher aufgetretenen Probleme waren in der Alignierung, die sehr schlechte Ergebnisse für die Basisdaten geliefert hat. Ein zweiter Ansatz bestand darin händisch ein Wörterbuch zur

Alignierung zu verwenden oder die Datensätze, in denen die Alignierung bereits vorgenommen wurde. Ein zweites Problem stellte die unterschiedliche Verwendung der japanischen Aspekt-Marker mit Zeitformen aus dem Englischen dar. So sind die Sätze

村田さんは結婚してる – Mr. Murata is married.

und

村田さんはここに座っている – Mr. Murata is sitting here.

zwar beides Ausprägungen des te-iru Aspekts, drücken aber auf unterschiedliche Art und Weise eine abgeschlossene Handlung aus, die im englischen grammatisch divergiert.

Dayyan Smith

Regularization of Neural Networks for Natural Language Processing

Betreuerin: Katharina Kann

In der Arbeit geht es um den Einsatz von Regularisierung bei Neuralen Netzwerken zur Stance-Erkennung. Das Anwendungsbeispiel zur Durchführung der Experimente findet auf der Fake-News Challenge statt. Dayyan definierte den Begriff Fake News für seine Arbeit mit einem Zitat aus der New York Times als: „A made up story with an intention to deceive“. Mit welchen Methoden können Fake-News entdeckt und gefiltert werden – ein Aufgabengebiet, welches auch für Experten schwierig sein kann.

In der Fake-News-Challenge geht es um die Untersuchung von Methoden der Künstlichen Intelligenz zum Kampf gegen Fake-News. Als Subtask davon ist es von Interesse bestimmte Nachrichten auf ihre Stellung zum Thema einer Schlagzeile zu untersuchen. Im Rahmen der Fake-News-Challenge wurde hierzu ein Datensatz angefertigt, in welchem Schlagzeilen und Artikeltexte und die Stellung des Textes zur Schlagzeile festgehalten werden. Es gibt 4 mögliche Klassifikationen: agree, disagree, discuss und unrelated.

Für die Verarbeitung des Tasks wurde in Dayyans System sowohl die Schlagzeile als auch der Nachrichtentext durch die Verwendung trainierter Word-Embeddings als Menge von Vektoren repräsentiert. Mit einem recurrent neural network, einer gated recurrent unit, ähnlich einem long short term memory Verfahren, wird eine Repräsentation von Schlagzeile und Text konkateniert. Hinzu kommt erneut die Vektor-Repräsentation von Schlagzeile und Text vor dem Output-Layer des NNs. Das hierzu verwendete Framework ist Theano.

Dass es durch diese Methode möglich ist das Datenset zu repräsentierten ist nicht erstaunlich, wichtig ist es viel mehr, die gelernten Parameter so zu verallgemeinern, dass es auf beliebige Daten anwendbar wird. Zu diesem Zweck dient die Regularisierung der Parameter.

Im System werden drei Arten von Regularisierung eingesetzt: L1-Regularisierung der uniformen Bestrafung von Gewichten im Vektor, L2-Regularisierung mit einer quadratischen Bestrafung der Gewichte, die deswegen mehr große Gewichte betrifft und einer Dropout-Regularisierung, bei welcher einzelne Neuronen im Netz, bei unterschiedlichen Epochen ausgeschaltet werden, wodurch mehrere Neuronen gleichzeitig die Parameter unterschiedlicher Datenpunkte lernen können, dies verhindert eine zu starke Spezialisierung einzelner Neuronen in der Architektur.

Die bisherigen Resultate der Tests zeigen, dass sich die Regularisierung weniger positiv auf die Datensätze auswirkt als anfangs angenommen. Nur wenige Modelle können egal ob mit oder ohne Regularisierung die Baseline des Tasks schlagen. Ein Grund für die schlechten Ergebnisse kann in der Größe der Hidden-Layer oder der Anzahl der Trainingsepochen liegen. Hier ist Dayyan durch seine verwendete Hardware eingeschränkt Modelle mit höherer Komplexität zu verwenden. Dayyan wird seine Ergebnisse trotz der schlechten Ergebnisse bei der Fake-News Challenge einreichen.

Thomas Ebert

Corpus based Identification of Text Segments.

Betreuer: Martin Schmitt

In der Arbeit soll die Effektivität von Worten zur Textsegmentierung für die automatische Textverarbeitung untersucht werden. Dabei wird davon ausgegangen, dass für Menschen die Segmentierung von Sätzen und Ausdrücken in Wörter ein adäquates Mittel für das Textverständnis ist, aber deswegen eine andere Segmentierung für maschinelle Verarbeitung eventuell geeigneter sein könnte. Ziel der Arbeit ist es einen Algorithmus zu entwickeln, mit welchem ein Satz oder Text in gute Segmente zur Weiterverarbeitung zerlegt.

Das Vorgehen ist wie folgt: Zunächst wurden aus dem Wikipedia-Korpus N-Gramme der Länge 1 bis 10 extrahiert. Wegen der großen Menge der Daten wurde sich auf die ersten 10.000 Texte des Korpus beschränkt, die auf insgesamt 22.650.880 Zeichen aufsummieren. Danach wurde für die N-Gramme eine Frequenzliste erstellt und die einzelnen Zeichen-N-Gramme mit einem Gütemaß bewertet.

Dieses Maß setzt sich zusammen aus: Länge des N-Gramms n * log(Frequenz des ngrams k) was die Zipfsche Verteilung der N-Gramme berücksichtigt. Professor Schulz kommentiere, dass die Festlegung eines bestimmten Gütemaßes zu den Kernproblemen dieser Methode zählt, da hierbei oft nur unzureichend motivierte Heuristiken verwendet werden können. In einem Testdurchlauf wird ein eingegebener Satz in diejenigen N-Gramme mit dem höchsten Gütemaß zerlegt.

Zu den Problemen in der Verarbeitung der Segmente zählt insbesondere die Größe der Eingabe, mit deren Länge die Laufzeit exponentiell steigt. Zur Lösung des Problems wurde die Größe des N-Gramm Fensters beschränkt. Dadurch ist aber gleichzeitig nicht mehr die Berechnung der höchsten Güte nicht mehr garantiert, aber die Segmentierung gegebenenfalls trotzdem besser als bei einem symbolischen Ansatz.

Evaluation der Methode ist u.A. bei Textsegmenten schwer da die Granularität unklar bleibt. Je nach Anwendung kann deswegen die Segmentierung relevant oder irrelevant sein. Als Beispiel wurde IR genannt, wo wie Korrektheit einzelner Segmente oft vernachlässigt werden kann, solange der Informationsgehalt gesamter Dokumente nicht verloren geht. Andere Anwendungen wie News Boundary Detection sind hingegen auf korrekte Segmentgrenzen angewiesen.

Als Experimentgrundlage für die Untersuchung dient Sentimentanalyse auf Basis von Movie Review Daten. Mit der Hilfe von word2vec werden n-gramm Embeddings als Input für einen LSTM Encoder erstellt. Mit diesem soll dann der Sentimentanalyse-Task bearbeitet werden.
Weitere Überlegungen für die Arbeit waren insbesondere die Untersuchung von Alternativen bei der Extraktion der N-Gramme. Obendrein stellt sich die Frage, ob das Ergebnis der Evaluation auf dem Sentimentanalyse-Task schon aussagekräftig ist, oder ob es auf mehrere angewandt werden sollte.