Protokolle zur Sitzung am 3.6.17 (Repetitorium und Kolloquium)

1 Computerlinquistik in Anwendungen der BSB-IT

Im Rahmen des heutigen Repetitoriums wurde der Vortrag über computerlinguistische Anwendungen in der Bayerischen Staatsbibliothek gehalten.

Am Anfang wurde die BSB kurz vorgestellt: gegründet 1558 ist die BSB eine der bedeutendsten europäischen Forschungs- und Universalbibliotheken mit internationalem Rang. Ihr Bestand beläuft sich auf etwa 10,5 Millionen Bände, 10 % davon sind digitalisiert. Die BSB arbeitet sowohl mit gedruckten Objekten, als auch mit Handschriften, Bildern und Musikstücken.

Eins der Ziele der BSB ist es, eine effektive und effiziente Suche in in ihrem Bestand zu ermöglichen. Was die dafür benötigten Technologien angeht, wurden u. A. Java (für backend und frontend), Spring-Framework und REST-Webservices genannt. Information Retrieval Probleme werden mithilfe von *Apache Solr* gelöst und Named Entity Recognition mit einer modifizierten Version von *Stanford NER*.

Am Ende der Sitzung wurde kurz über neuere CL-Anwendungen in BSB (u. A. *optical music recognition, automatische inhaltliche Erschließung*) berichtet und diskutiert.

Zeichen (ohne Leerzeichen): 914

2 Pascal Guldener: Neural Netorks for Named Entity Recognition

Den ersten Vortrag im Kolloquium hat Hr. Guldener gehalten. Sein Vortrag basierte auf dem Paper von Ronan Collobert et al („Natural Language Processing (Almost) from Scratch)".

Es wurde versucht, das NER-Problem (im Paper sind aber auch POS, Chunking und Semantic Role Labeling Probleme erwähnt) mithilfe eines System zu lösen, das über (fast) kein linguistisches Wissen verfügt. Der traditionelle Ansatz benutzt das vorhandene Wissen darüber, wie die Sprache funktioniert, was beispielsweise mithilfe der *feature engineering* erreicht werden kann. Dieser Ansatz ist jedoch in allen Hinsichten teuer.

Es wurde kurz über die allgemeine Architektur (window approach und sentence approach) der neuronalen Netzwerke des SENNA-Projekts (Semantic/syntactic Extraction using a Neural Network Architecture) berichtet: Ein Wort wird als ein d-dimensionaler Vektor dargestellt, die erste Schicht des Netzwerks konvertiert jedes Input-Wort zu einem Index, der einer Look-Up Tabelle übergeben wird.

Das nicht überwachtes Training der Sprachmodelle basierte auf 100 000 häufigsten Wörter aus der englischen Wikipedia und 30 000 häufigsten Wörtern aus Reuters. Während der Trainingsphase wurden alle Hyperparameter vorsichtig angepasst (z. B. learning rate, embedding dimensions, hidden units). Der F1-Score dieses Systems (+Gazetteer) beläuft sich auf 89.59, was um 0.28 größer ist als der entsprechende Score bei einem Benchmarksystem von Ando und Zhang (2005, semi-supervised training).

Unüberwachtes Training kann also sehr mächtige Wort-Repräsentationen produzieren, die dann je nach konkreter Aufgabe beliebig erweitert werden können.

3 Anton Serjogin: Learning String Edit Distance

Den 2. Vortrag hat Hr. Serjogin gehalten. Der Vortrag basierte haupsächlich auf dem Paper von Ristard „Learning String Edit Distance“.

Am Anfang des Vortrags wurde der Begriff „string distance“ grob als „eine Methode zur Quantifizierung der Unterschiedlichkeit von zwei Strings“. definiert. String Distance wird in zahlreichen Bereichen eingesetzt, u. A. in Bioinformatik und NLP.

Es wurde kurz über die „Levenshtein Distance“ berichtet, was eine der häufigsten Metriken sei. Anhand einiger Beispiele wurde gezeigt, wie man die drei Operationen durchführt (deletion, insertion und substitiution).

Wenn man das Problem als „stochastische Transduktion“ betrachtet, ergeben sich 2 String Distances: 1) Viterbi edit distance (negativer Logarithmus der Wahrscheinlichkeit der höchstwahrscheinlichen Editierfolge für das String-Paar) und 2)Stochastic edit distance (negativer Logarithmus der Wahrscheinlichkeit des String-Paars).

Es wurden kurz 3 Varianten der gedächtnislosen stochastischen Transduktors erwähnt: 1) Parameterbindung; 2) finite Mischungen; 3) stochasticher Transduktor mit Gedächtnis.

Es wurde dann über eine Anwendung berichtet, wo String Distances eingesetzt werden: Aussprache der Wörter. Der Input für die Anwendung wurde als ein 6-Tupel ($< W, A, B, L, C, C' >$) beschrieben, wo: W - die Menge der syntaktischen Wörter ist; A - das Alphabet der phonologischen Segmente ist; B - das Alphabet der phonetischen Segmente ist; L - das Aussprachelexikon ist; C - das Training-Korpus von markierten phonetischen Strings ist; C’ - das Korpus der nicht markierten phonetischen Strings. Der Output ist eine Menge der Labels für den Testkorpus C’.

Diese Anwendung wurde in 5 verschiedenen Variationen trainiert und bei Benutzung jeweils der Levenshtein, stochastischen (tied, untied, mixed) und Viterbi (tied, untied, mixed) Distanzen evaluiert. Es wurde im Abschluss kurz darüber diskutiert.

4 Katja Bertholdt: das Zipfsche Gesetz (I. Teil)

Die letzte Vortragende hat angefangen, über das Zipfsche Gesetzt zu berichten. Das Thema wird in der nächsten Sitzung weiter behandelt.

Am Anfang des Vortrags wurde kurz über George Kingsley Zipf gesprochen, der dieses Gesetz entwickelte, das nach ihm benannt wurde.

Das Zipfsche Gesetz beschreibt ein Modell, mit dessen Hilfe es möglich ist, bei bestimmten Größen, die in eine Rangfolge gebracht werden, deren Wert aus ihrem Rang abschätzen zu können. Das Gesetz findet häufige Verwendung in der Linguistik, aber nicht nur da, sondern in vielen Naturwissenschaften.

Für Linguistik kann man das Gesetz so beschreiben: der Rang i eines Wortes ist indirekt proportional zu seiner relativen Häufigkeit; wenn z.B. Wörter nach der Häufigkeit sortiert werden, ist die Wahrscheinlichkeit ihres Auftretens umgekehrt proportional zur Position innerhalb der Reihenfolge:

$$p(n) \sim \tfrac{1}{n}.$$

Daraus kann man schließen, dass die meisten Wörter selten sind.

Anhand des Prinzips der geringsten Anstrengung wurde das Gesetz nochmals erläutert. Die Vortragende hat für Demonstration des Gesetzes das Projekt „Deutscher Wortschatz“ (Uni Leipzig) benutzt. Alle Beispiele, die die Vortragende durchgerechnet hat, entsprachen dem Zipfschen Gesetz, das zusätzlich mithilfe mehrerer Graphiken veranschaulicht wurde.