Christoph
Flemmig
CIS, LMU München

15.06.2017

Protokoll für das Kolloquium vom 12.6.2017

In dieser Sitzung wurde drei Bachelorarbeiten von den Studenten in Form von kurzen Vorträgen mithilfe einer kleinen Präsentation am Beamer vorgestellt.

1. Vortrag:

Tobias Ramoser, BA-Betreuer: M.Sc. Martin Schmitt
Thema: Phonologically Enhanced Character Embeddings

Als erstes stellte Tobias die Motivation für seine Arbeit vor. Er betonte dass Anwendungen der maschinellen Sprachverarbeitung im heutigen Alltag sehr wichtig sind, diese sich jedoch hauptsächlich mit Wörtern beschäftigen. Deshalb konzentriert sich seine Arbeit auf die Beziehungen und Zusammenhänge im zwischen Buchstaben im Vektorraum. Speziell konzentriert er sich dabei auf Phonologische Vektorraumrepräsentationen von Buchstaben. Das Ziel seiner Arbeit ist da bei die Erstellung von verschiedenen Vektorrepräsentationen von Buchstaben mit phonologischen Features um diese dann mit Zufallsvektoren bei der Transkription in SAMPA zu vergleichen.
Im nächsten Teil seine Vortrags stellte Tobias die Thematischen Hintergründe seiner Arbeit vor. Als erstes gab Tobias eine Einführung in die Aspekte der Phonetik die für seine Arbeit relevant sind. Dafür erklärte er die Begriffe Artikulationsart, Artikulationsort und Stimmhaftigkeit. Dann gab er einen kurzen Überblick über Phonologie. Im Anschluss sprach er kurz über Word2Vec. Wichtig waren hier vor allem die für seine Arbeit relevanten Architekturen die er vorstellte, namentlich das Continuous Bag-of-Words Modell und das Skip-Gram Modell und stellte deren Funktionsweise vor. Außerdem nannte er noch die drei Lernalgorithmen die er benutzt. Diese sind Hirarchical Softmax, Negative Sampling und Downsampling of frequent words. Abschließend für die Hintergründe seiner Arbeit stellte er noch das SAMPA-Alphabet vor. Hier handelt es sich um ein auf ASCII basiertes phonetisches Alphabet das maschinell lesbar ist und zur Darstellung der Aussprache dient.
Der letzte Teil von Tobias' Vortrag widmete sich den Experimenten für seine Arbeit und deren Auswertung. In seiner Arbeit führte Tobias insgesamt vier Experimente durch. Diese bestanden aus zwei Implementierungen für Vektorerstellung, einer für Zufallsvektoren und einem Experiment mit Word2Vec Vektoren. Daraufhin ging er im Detail auf die Experimente ein. Das erste führte er mit Char-Vektoren durch. Hierbei handelt es sich um eine von Tobias selbst durchgeführte Implementierung. Phonologische Features und Unterkategorien wurden vorher definiert und ein ein Vektor mit den Werten (0,0,0,0,0) wurde initialisiert. Dann werden Phonemvektoren berechnet und daraufhin Buchstabenvektoren mit dem Durchschnitt der entsprechenden Phonemvektoren berechnet. Das nächste Experiment benutzte One-hot Vektoren die die gleichen Features und Unterkategorien wie die Char Vektoren verwenden, allerdings eine binäre Klassifizierung benutzen. Für die zufällig generierten Vektoren benutzte er zwei unterschiedliche Arten, eine mit 15 Dimensionen und eine mit 100 unter Verwendung des „random“-numpy-Moduls. Für die Word2Vec Vektoren wurden die selben Dimensionen wie bei den zufällig generierten Vektoren benutzt. Als Trainingsdaten dienten Buchstaben anstatt Wörter, wobei diese nach phonologischer ähnlichkeit aufeinander trainiert wurden.
In der quantitativen Analyse schnitten die 100-dimensionalen Zufallsvektoren am besten ab, gefolgt von den One-hot Vektoren. Am schlechtesten schnitten die 15-dimensionalen Zufalls- und Word2Vec-Vektoren ab. In der qualitativen Analyse berechnete er eine Fehlerquote mithilfe der

Levenshtein-Distanz. Die qualitative Analyse unterstützte die Ergebnisse der quantitativen Analyse.

2. Vortrag

Jakob Sharab, BA-Betreuerin: Fabienne Braune
Predicting New Domain Senses in English Medical Texts

Auch Jakobs Vortrag begann mit der Motivation für seine Arbeit. Es geht darum dass Wörter je nachdem in welcher Domäne sie verwendet werden mit unterschiedlicher Bedeutung übersetzt werden können. Dabei führte er das Beispiel von „administration“ heran das je nach Kontext sowohl als Verwaltung wie auch als Verabreichung übersetzt werden kann. Dies führt zu Fehlern bei Statistical Machine Translation Systemen. Deshalb wurde das Task „Sense Spotting“ definiert. Hierbei geht es darum Features zu finden die auf eine Bedeutungsveränderung hinweisen und mithilfe dieser einen Cassifier zu trainieren.
Als nächstes sprach Jakob über Topic Modeling. Dies ist eines der zuvor erwähnten Features zum entdecken von Bedeutungsveränderung. Das Ziel von Topic Modeling ist in großen Textkorpora die darin enthaltenen Topics zu finden. Dabei werden mithilfe von Algorithmen die einzelnen Wörter in Dokumenten Analysiert. Der Vorteil dieses Verfahrens ist dass es ohne vorhergehende Annotation auskommt. Danach ging Jakob mehr auf das Modell ein dass er für das Topic Modeling benutzt hat. Zuerst gab er eine kurze Erklärung für Generative Modelle. Hierbei geht es darum Modelle zu bauen die analysieren wie das zu unterscheidende Objekt auszusehen hat, dann werden die Elemente an alle Modelle übergeben und es wird eine Wahrscheinlichkeit für jede Klasse berechnet.
Jakob benutzte die Latent Dirichlet Allocation für seine Arbeit welche ein generatives Modell ist. Es dient dazu Dokumente in einzelne Topics zu untergliedern. Es basiert auf dem „bag-of-words“-Prinzip. Durch eine Umkehrung des generativen Prozesses werden hier die Dokumente in Topics unterteilt.
Im dritten Teil seines Vortrags ging Jakob noch einmal genauer auf das Topic Model Feature ein. Dieses Modell erkennt ob sich die Häufigkeit eines Wortes innerhalb eines Topics ändert wenn es in eine andere Domäne wechselt und erkennt daran ob eine Bedeutungsveränderung vorliegt. Jakob erklärte auch noch genauer wie man dieses Modell als Formel definiert.
Als nächstes erklärte Jakob die Ähnlichkeitsmaße die für seine Arbeit relevant sind. Diese sind die Kosinus-Ähnlickeit, die relative Entropie und die Ähnlichkeit aufgrund der Anzahl gleicher Wörter welche alle jeweils in der Formel des Modells verwendet werden. Je nachdem ob die Werte dann hoch oder niedrig sind entscheidet ob sich die Bedeutung verändert hat.
Jakob gab danach eine Übersicht über die Daten die er verwendet hat. Es handelt sich hier um Parallel-Korpora in Englisch und Deutsch. Aus der medizinischen Domäne verwendet er das EMEA Korpus das Prüfberichte von Medikamenten enthält und aus der Nachrichten Domäne verwendet er das General Korpus das im WMT Shared Task von 2016 verwendet wurde. Letzteres besteht aus den Daten mehrerer Korpora.
Der sechste Punkt von Jakobs Präsentation beschäftigte sich mit seinem Experiment. Hier ging er sehr genau auf seine Vorgehensweise ein. Zuerst wandte er Tokenisierung auf die einzelnen Dokumente an und entfernte die Stopwörter. Danach unterteilte er die Dokumente in jeweils 100 Topics mithilfe der Latent Dirichlet Allocation. Aus diesen extrahierte er dann Wörter die hohe Wahrscheinlichkeiten in einem Topic der alten Domäne aufwiesen und deren Bedeutung sich veränderte. Für diese Berechnete er dann die Ähnlichkeiten auf welche er dann die Formel für das Tpoic Model Feature anwandte. Um einen Vergleichswert zu haben berechnete er diese Werte auch für Wörter die keine Bedeutungsveränderung hatten und bildete daraus den Mittelwert.
Zum Schluss sprach Jakob noch über Probleme und Ergebnisse während seiner Arbeit. Ein Problem war dass er Probleme damit hatte Beispiele zu finden in denen sich die Bedeutung in der neuen Domäne veränderte. Außerdem konnte er ohne Cassifier die Ergebnisse nur quantitativ vergleichen. Als Ergebnis konnte er präsentieren dass die Ähnlichkeitsmaße der relativen Entropie und das Maß aufgrund der gleichen Wörter die besten Resultate erzielten.

3. Vortrag

Mai Linh Pham BA-Betreuer: Yadollah Yaghoobzadeh
Thema: Analysis of NIL Results in an Entity Linking System

At the beginning of her report Mai Linh gave some definitions for the terms Entity Linking, NIL Results and Fine-Grained Entity Annotation. She went on with giving an explanation of the motivation for her work. Entity Linking systems can not link all entities because some entities are missing from knowledge bases. That is why there is a need to improve Entity Linking systems by analysing NIL results. This can be done by clustering NIL mentions for analysis. The goal is to examine wether fine-grained types are useful for clustering and analyzing NIL mentions.
After that she gave an overview of the individual steps of her task. The first step is to combin the outputs of an entity annotation tool and an entity linking system, then extract and afterwards cluster the NIL-output and to use fine grained types for the clustering task. She also gave a visual representation of the entire process.
In the next part she gave a more detailed description of the tools she used, namely FIGER, a fine-grained entity annotation system and WAT, an entity linking system. FIGER uses 112 tags and standard BIO-encoded tags. WAT uses the JSON-format. It is important that the output of WAT has the same format as the FIGER output.
She went on to explain the extraction of NIL mentions in detail. Here only the mentions of unlinked entities that are in a list of entity names that was created in the knowledge base were regarded.
Next came a closer look at the different clustering approaches. Those were a coarse-grained, fine-grained and top-level type. For coarse-grained clustering she only gave a broad overview. Fine-grained clustering works by dividing types into multi-level and single-level types and then clustering them semantically to map single types to multi-level types to then cluster the muti-level types. In top-level clustering many first-level tags in multi-level types are identical. These get grouped up to reduce the total number of clusters. Types that are too specific get grouped into a cluster that is titled undefined. Then new domains are created to reduce the undefined types.
In the last part of her presentation Mai Linh presented her conclusions. She found that fine-grained entity types can be used for clustering semantically related NIL mentions and take into account the lexical and contextual properties of an entity. She also concluded that they are more informative than coarse-grained types.