Protokolle von 12.06.2017

Präsentation von Tobias Ramoser

Das Thema seiner Bachelorarbeit ist “Phonologically-Enhanced Character Embeddings”. Die Arbeit war unter der Betreuung von M.Sc. Martin Schmidt gemacht. Vor allem Tobias hat einen Überblick zu seine Präsentation gegeben, die Hauptpunkte davon sind Motivation, Ziel, theoretisches Basis, durchgeführte Experimente und ihre Ergebnisse and Auswertung sowie die benutzte Quellen. Zu Motivation hat Tobias gemerkt, daß man die Anwendungen der maschinellen Sprachverarbeitung nicht mehr aus unserem Alltag wegzudenken soll, z.B. Spracherkennung, Maschinelle Übersetzung etc. In seiner Arbeit hat er die Beziehungen und Zusammenhänge zwischen Buchstaben im Vektorraum recherchiert, das Feld, welche noch die Schwäche hat und verbessert werden soll ist phonologische Vekorraumpräsentationen von Buchstaben. Daraus kommt ein Ziel: Erstellung von verschiedenen Vektorrepräsentationen von Buchstaben mit phonologischen Features und der Vergleich mit Zufallsvektoren bei der Transkription in SAMPA.

Thematischer Hintergrund erhielt das Folgendes - die Grundlagen von Phonetik und Phonologie, Word2Vec und SAMSA-Alphabet. Die Hauptmerkmale der Artikulation und Stimmhaftigkeit sind die Artikulationsart (plosiv, frikativ, nasal, lateral etc), Artikulationsort (bilabial, labiodental, alveolar, dental etc) sowie die Stimmhaftigkeit (stimmhaft, stimmlos). Dann gab er die Definition von Phonologie and die Merkmale von Phonemen.

Als Nächstes, hat Tobias Word2Vec Programm präsentiert. Es wurde von Mikolov entwickelt und repräsentiert das Programm zur automatische Vektorerstellung von Wörtern, als Input wurden die Trainingsdaten gegeben (Text mit vielen Wörtern), Output - Wortvektoren und Distanz zu einem Wort. Verarbeitung wurde mit der Hilfe von neuronales Netz durchgeführt, sie besteht aus Architektur und Lernalgorithmus und ähnliche Wörter erhalten ähnlichen Vektor. Wiederum, Architektur besteht aus zwei Modellen: Continuous Bag-of-Words Modell (es ist dreischichtiges neuronales Netz und sagt aus einem Kontext ein gewisses Wort voraus). Das zweite Modell ist dreischichtiges neuronales Netz und sagt den Kontext aus gegebenen Wort voraus.

Lernalgorithmen bestehen aus hierarchical Softmax, negative Sampling und Downsampling der negativen Wörtern. Sampa-Alphabet kommt aus ASCII-basiertes, maschienen-lesbares, phonetisches Alphabet, mit welchem die Aussprache der Laute dargestellt wird. Es wurde von 1987-1989 entwickelt, um phonetische Transkriptionen der offiziellen Sprachen der damaligen Europäischen Gemeinschaft übermitteln und verarbeiten zu können.

Was die gemachten Experimente betrifft, es war insgesamt vier, die er gemacht hat, dazu zwei Implementierungen für Vektorenerstellung und Zufallsvektoren und Word2Vec Vektoren sowie Transkription in SAMPA. Tobias hat eine Implementierung von Char-Vectors präsentiert: hier werden phonologische Features und Unterkategorien wie vorher definiert, Initialisierung des Vektoren, dann kommen die Berechnung der Phonemvektoren und immer wenn eine Eigenschaft auf Phonem zutrifft, erhält es den dementsprechenden Vektor und die Berechnung der Buchstabenvektoren mittels Durchschnitt aller entsprechenden Phonemvektroen wird gemacht. Auch hat er Information über One-hot Vektoren und Randomized Vektoren gegeben, die haben gleiche Features und Unterkategorien wie bei Char-Vektoren und sie haben auch binäre Klassifizierung. Außerdem Vektorgröße “wächst” somit auf 22 Dimensionen. Randomized Vektoren repräsentieren zufall generierte Vektoren.

Es entsteht eine Generierung von zwei Arten von Buchstabenvektoren - 15 und 100 Dimensionen und Verwendung der “random” - numpy-Moduls.
Zu den Ergebnissen wurde es qualitative und quantitative Analyse gemacht. Quantitative Analyse hat gezeigt dass die Accuracy bei der word2vec12 ist am geringsten ist, nämlich 0.0014 % und bei randim100 ist am größten - 0,75%. Vor allem eine Berechnung der Fehlerquote mittels Levenshtein-Distanz hat das Ziel um zu zeigen wie viel Korrekturen durchschnittlich gemacht werden müssen und das bestätigt auch die Ergebnisse der quantitativen Auswertung. Dazu wurde es als ein Beispiel Fehlerdatei gezeigt und die benutzte Quellen gegeben.

Präsentation von Jacob Sharab:

Das Thema der Bachelorarbeit ist “Predicting New Domain Senses in English Medical Texts”, die Betreuerin ist Fabienne Braune. Als erstens hat Jakob einen Überblick zu seiner Präsentation gegeben, was enthält die Motivation, Topic Modeling und Topic Model Feature, Ähnlichkeitsmaße zwischen Topics, verwendete Daten und Experimentdurchführung. sowie die entstehende Probleme und Ergebnisse.
Generell haben die Wörter verschieden Bedeutungen bei den unterschiedlichen Domänen (z.B. das Wort “administration” hat allgemeine Bedeutung “Verwaltung und spezielle medizinische „Verabreichung“ (eines Medikamentes). Daher kommt einen Fehler bei Statistical Machine Translation System (SMT): Bilden von Wortpaaren mit „administration“ [Wort aus der ursprünglichen Sprache] + „Verwaltung“ [dessen Übersetzung] werden in neuer Domäne nicht mehr korrekt. Deshalb Definition eines neues Task „Sense Spotting“ ist nötig für eine Suche von Features die Bedeutungsveränderung indizieren und dann Training eines Classifiers mit Hilfe dieser Features. Eine von den Features stellt das Topic Model Feature dar. Allgemein Topic Modeling hat das Ziel in großen Textkorpora darin enthaltene Topics zu finden mit der Hilfe von Algorithmen, die einzelne Wörter in den Dokumenten analysieren. Der Vorteil davon liegt daran daß keine vorhergehende Annotation der Daten nötig ist. Das hat auch praktisches Nutzen da die Menge von Daten heutzutage menschliche Kapazitäten übersteigt und es wird dann nötig für Organisierung von großen Textarchiven.
Als Nächstes hat Jacob den Klassifizierungsproblem bei generativen Modellen gesprochen. Man will z.B. zwischen einem Hund und einem Elefanten unterscheiden. Dazu gibt es unterschiedliche Ansätze: diskriminativer Ansatz (Training eines Klassifikators mit Features, der eine „Linie“ (Decision Boundary) findet, die Klassen voneinander trennen und je nachdem auf welche Seite der Linie die Werte fallen wird das Tier als Hund oder Elefant klassifiziert) und generativer Ansatz (man soll zweier Modelle bauen, die analysieren wie ein Hund und wie Elefant ausschaut, übergeben des zu klassifizierenden Tieres an die Modelle und danach die Wahrscheinlichkeit für jede Klasse berechnen). Das Thema Latent Dirichlet Allocation (LDA) repräsentiert generatives Wahrscheinlichkeits-Modell und wird in der Arbeit dafür genutzt um Dokumente in einzelne Topics zu untergliedern. Grundidee liegt daran daß jedes Dokument aus einer zufälligen Mischung latenter Topics besteht aus einer Verteilung über Wörter. Das Prinzip basiert auf der Annahme der „bag-of-words“ und generative Entstehung eines Dokumentes. Damit durch Umkehrung des generativen Prozesses werden Dokumente in Topics unterteilt.
Das Thema “Topic Model Feature” war detailliert präsentiert weil Feature, welche im Rahmen des „Sense Spotting“ Task definiert wurde, nimmt die Änderung der Häufigkeit

eines Wortes innerhalb eines Topics beim Wechsel in die neue Domäne als Indikator für eine Bedeutungsveränderung. Dabei wurde die Ähnlichkeitsmaße zwischen den Topics berechnet und miteinander verglichen. Es gibt verschiedene Ähnlichkeitsmaße: Kosinus-Ähnlichkeit, Relative Entropie und die Ähnlichkeit aufgrund der Anzahl gleicher Wörter.
Als Daten wurden parallele Korpora (Englisch, Deutsch) verwendet. Daten für die medizinische Domäne erhalten EMEA Korpus (über 41.000 Dokumente). Daten für die Nachrichten Domäne bestehen aus Daten mehrerer Korpora und haben auch ein General Korpus (verwendet im WMT Shared Task 2016). Während der Arbeit sind folgende Fehler aufgetreten: es war nicht einfach viele Beispiele für Wörter zu finden, deren Bedeutung sich in der neue Domäne verändert weil die Wörter die Bedingung erfüllen müssen, in einem Topic eine hohe Wahrscheinlichkeit zu haben. Da es schwierig ist ohne einen Classifier eine Decision Boundary zu finden, die Ergebnisse konnten nur quantitativ miteinander verglichen werden und der Vergleich ist nach dem Prinzip wie weit die Werte für Wörter, deren Bedeutung sich in einer neuen Domäne ändert, unterhalb des Durchschnittswertes für Wörter liegt, deren Bedeutung sich nicht ändert.
Nach den Experimenten haben die Relative Entropie und das Maß aufgrund der gleichen Wörter die besten Resultate erzielt.

Presentation by Mai Linh Pham:

The topic of Mai Linh's thesis is "Analysis of NIL results in an entity Linking System", it was from Yadollah Yaghoobzadeh supervised. First of all Mail Linh gave some definition to her thesis, there are entity linking (it's a process of linking mentions from text to corresponding entity in knowledge base), NIL results are the entities that are not linked to the KB and Fine-grained entity annotation - fine-grained tagset has more tags than standard NER tags as LOC, ORG, PER.
The motivation to her work is that entity linking systems can't link all entities that's why some entities are missing from KB and improvement from EL systems by analyzing the NIL results and then cluster NIL methods for analysis by introducing new method for clustering. The goal of the thesis is defined like this: examine whether fine-grained types are useful for clustering and analyzing NIL mentions.
Task description includes the combination of an entity annotation tool and an entity linking system then the extraction of NIL output, its clustering and usage of fine-grained types for clustering task.
Than Mai Linh represented and described the whole process that includes El system WAT, NER System FIGER, KB, NIl etc. Tools that were used are: FIGER (fine-grained entity annotation system) that has 112 tags und allows an overlapping types. For instance, The New York Times - company, written work, news agency. This tool is generally better in recognising uncommon entities ("long tail"). Another tool is WAT - it is entity linking system that has next components: spotter (it scans input text for mentions, retrieves list of candidate entities) disambiguator (it ranks candidate entities with different disambiguation algorithms) and pruner (removes useless annotations and aims at increasing the precision). WAT Output has JSON-format und needs to be in same format as FIGER output.
For extracting NIL mentions it is necessary to create a list of all names in the KB that includes all Wikipedia titles and redirect links with the regard of only unlinked mentions that

have a corresponding KB entry. Unliked mentions are those that annotated by FIGER but not linked by WAT.

By the clustering of NIL Mentions there are three clustering approaches: coarse-grained type, fine-grained type and top-level type. Fine-grained clustering divides types into multi-level and single-level types, clusters types semantically, then maps single types to multi-level types and clusters multi-level types. Top-level clustering groups first-level tags to reduce number of clusters, then groups too specific types to cluster “undefined” and creates new domains to reduce undefined types.

At the end of presentation Main Linh concluded that Fine-Grained types can be used for clustering semantically related NIL mentions and are more informative than coarse-grained types. The information anchord in tags can be used for analysis.