Phonologically Enhanced Character Embeddings

Vortragender: Tobias Ramoser Betreuer: Martin Schmitt

Tobias hat in seiner Bachelorarbeit untersucht, ob man word embddings nicht nur für Wörter machen kann, sondern auch phonologisch berechnen kann, also Beziehungen und Zusammenhänge zwischen Buchstaben im Vektorraum darstellen kann, in dem man deren phonologische Eigenschaften mit Vektoren modelliert.

Um dies zu untersuchen, hat er verschiedene Vektorrepräsentationen von Buchstaben mit phonologischen Features berechnet und ihre Performance mit der von Zufallsvektoren verglichen.

Er erklärt zuerst einige zugrundeliegende Begriffe. Die Phonetik ist die Lehre von der Artikulation von Lauten. Hier wird jeder Laut in drei Kategorien bewertet: Artikulationsart, Artikulationsort und Stimmhaftigkeit. Die Phonologie dagegen beschreibt die Systematik der Laute innerhalb einer Sprache. Hier werden Phoneme definiert: diese müssen wortunterscheidende sein und sich in mindestens einer phonetischen Eigenschaft unterscheiden.

Um die Vektoren zu erstellen, benutzt er word2vec. Dies ist ein Programm zur automatischen Vektorerstellung zu Wörtern. Als Input werden Trainigsdaten eingegeben und dann lernt ein Neural Network, welches aus einer Architektur und einem Lernalgorithmus besteht, Vektoren, wobei ähnliche Wörter ähnliche Vektoren erhalten. Er kontrastiert zwei unterschiedliche Architekturen für NNs: das continuous bag of words Modell ist ein zweischichtiges NN welches aus einem Kontext ein Wort vorhersagt, und ein Skip-Gram-Modell sagt aus dem Wort den Kontext voraus. Verschiedene Lernalgorithmen sind hierarchical softmax, negative sampling und downsampling of frequent words.

Für den phonologischen Input benutzt er das SAMPA Alphabet. Das ist ein ASCII-basiertes, maschinenlesbares phonetisches Alphabet mit dem die Aussprache von Lauten dargestellt wird. Es wurde entwickelt um phonemischen Transkriptionen der offiziellen Sprachen der damaligen Europäischen Gemeinschaft übermitteln.

Seine erste Idee sind char-vectors, die er selbst implementiert hat. Hier werden die phonologischen Features, die er beschrieben hatte, definiert und daraus Phonemvektoren berechnet. Immer wenn eine Eigenschaft auf ein Phonem zutrifft erhält es den dementsprechenden Vektor einen bestimmten Wert. Für Vokale werden zusätzlich der Grad der Öffnung des Mundraums und die Rundung der Lippen erfasst. Ein anderer Ansatz sind one-hot Vektoren. Hier werden die gleichen Features benutzt, die Klassifizierung ist aber binär, was dazu führt dass die Vektoren 22 Dimensionen groß werden. Als Baseline zu dem trainierten Vektoren hat er außerdem mit numpy zufällige Vektoren erstellt. Der genaue Aufbau seines Versuchs wird leider nicht ganz klar.

Die quantitative Analyse mit Accuracy hat ergeben, dass die Zufallsvektoren besser waren als die phonologischen Vektoren, Auch die Berechnung einer Fehlerquote mittels der Levenshteindistanz, die aussagt wieviele Korrekturen durchschnittlich gemacht werden müssen um das richtige Ergebnis zu erhalten, bestätigte diese Ergebnisse.

Predicting New Domain Sense in English Medical Texts

Vortragender: Jakob Sharab Betreuerin: Fabienne Braune

Die Motivation der Arbeit ist, dass Wörter in verschiedenen Domänen unterschiedliche Bedeutungen haben, z.B. heißt administration im allgemeinen Verwaltung, aber in der Medizin die Verabreichung

eines Medikaments. Dies führt zu Fehlern bei SMT Systemen weil die Wortpaare nur in der ursprünglichen Domäne anwendbar sind. Deshalb wurde eine neue Task namens Sense Spotting erfunden, bei der features gefunden werden sollen die Bedeutungsveränderung indizieren. Eines dieser features ist das Topic Model Feature.

Topic Modeling bedeutet, in Textkorpora enthaltene Topics zu finden, indem mithilfe von Algorithmen die einzelnen Wörter in Dokumenten analysiert werden. Ein Vorteil bei dieser Methode ist, dass keine vorhergehende Annotation der Daten nötig ist. Topic Modeling ist sehr nützlich, weil man damit große Datenarchive organisieren kann.

An dieser Stelle wird der Unterschied zwischen diskriminativen und generativen Modellen erklärt. Bei einem diskriminativen Ansatz wird ein Klassifikator trainert, der eine decision boundary findet die die gewünschten Klassen voneinander trennt. Bei einem generativen Ansatz wird für jede Klasse ein Modell gebaut und das zu klassifizierende Objekt an die Modelle übergeben. Jedes Modell berechnet dann eine Wahrscheinlichkeit, dass das Objekt zu dieser Klasse gehört und das Objekt wird der Klasse mit der höchsten Wahrscheinlichkeit zugeordnet.

Das Modell das in dieser Arbeit genutzt wird, Latent Dirichlet Allocation, ist ein generatives Modell. Es wird speziell genutzt um Dokumente in verschiedene Topics zu untergliedern. Die Grundeidee ist, dass jedes Dokument aus einer zufälligen Mischung latenter topics besteht, wobei jedes topics wiederum eine Verteilung über Wörter ist. Dieses Modell ist basiert auf der bag of words Annahme, die Reihenfolge der Wörter wird als vernachlässigt und die Reihenfolge der Dokumente innerhalb eines Korpus ist austauschbar. Die generative Entstehung eines Dokumentes funktioniert so: Die Anzahl von Wörtern die ein topic enthält wird festgelegt, abhängig von der Mischung an Topics die in einem Dokument enthalten sind. Dann werden die Wörter generiert, in dem das Topics ausgewählt wird aus dem das Wort stammt.

Das Topic Model feature nimmt die Änderung der Häufigkeit innerhalb eines Topics beim Wechsel in die neue Domäne als Indikator für eine Bedeutungsveränderung. Beispielsweise haben bestimmte Wörter wie medicaments oder daily in der medizinischen Domäne eine hohe Wahrscheinlichkeit, aber in der allgemeinen Domäne bei ähnlichem Topic nur eine geringe Wahrscheinlichkeit.

Das Ziel der Arbeit ist jetzt, verschiedene Ähnlichkeitsmaße miteinander zu vergleichen. Die Kosinusähnlichkeit gibt einen Wert zwischen 1 und 0 aus wobei bei 1 die Vektoren identisch sind. Die Relative Entropie berechnet Abstand zwischen zwei Wahrscheinlichkeitsverteilungen. Je höher der Wert desto weiter auseinander gehen die Verteilungen. Außerdem gibt es noch die Ähnlichkeit aufgrund der Anzahl gleicher Wörter: je mehr geliche Wörter unter den top n Wörter zweier topics sind, desto ähnlicher sind sie. Der Wert wird mit der Wahrscheinlichkeit gewichtet.

Die verwendeten Daten sind für die medizinische Domäne der EMEA Korpus mit über 41000 Dokumente und für die Nachrichtendomäne den Generalkorpus der WMT Shared Task 2016. Die Korpora wurden tokenisiert und die Soppwörter entfernt. Dann wurden die Dokumente mithilfe der LDA mit Gensim in jeweils 100 topics unterteilt. Darin wurden die Wörter ausgewählt, die eine hohe Wahrscheinlichkeit in einem topic aus der alten Domäne haben und deren Bedeutung sich verändert. Auch für Wörter die keine Bedeutungsveränderung haben wurden Ergebnisse berechnet.

Ein Problem war, dass es schwierig war genug Wörter zu finden die ihre Bedeutung ändern und wahrscheinlich sind. Die besten Ergebnisse wurden mit der relativen Entropie und dem Maß mit den häufigsten Wörtern erzielt.

Analysis of NIL Results in an Entity Linking System

Vortragende: Mai Linh Pham Betreuer: Yadollah Yaghoobzadeh

Entity linking is the process of linking mentions from test to corresponding entry in a knowledge base(KB). NIL results are entities that are not linked to the KB. The motivation of the thesis is that EL systems can't link all entities which means that some entities are missing from the KB. An EL system could be improved by analysing those NIL results. NIL mentions should be clustered for analysis and this thesis wants to introduce a new method for clustering and examine whether fine grained types are useful for clustering NIL entries. The thesis there wants to combine the outputs of an entity annotation tool and an entity linking system, extract the NIL output and cluster it using fine-grained types.

The tools used are the FIGER fine grained entity annotation system which uses 112 tags and allows overlapping types which makes it better for recognizing uncommon entities. The output of the system is in standard BIO tag form. The entity linking system is the WAT system which consists of three phases: the spotter scancs input text for mentions and retrieves a list of candidate entities, the disambiguator ranks candidate entities with different disambiguation algorithms and the pruner removes useless annotations and aims at increasing precision. The output of WAT is in JSON.

The NIL mentions here are mentions that are annotated by FIGER but not linked by WAT. The three clustering approaches used are coarse grained types, fine grained types and top level types. Examples of NIL types are: organization/company, person/artist, government/political party or music. Fine grained clustering divides types into multilevel and single level types where single types are mapped to single types. For top level clustering, the number of clusters is reduced to 30. As conclusion, the presenter says that Fine grained entity types can be used for clustering semantically related NIL mentions and are more informative than coarse-grained types.