Vortrag: „Phonologically-Enhanced Character Embeddings“ Tobias Ramoser Betreuer : Martin Schmitt

Die Motivation der Arbeit besteht darin, dass die Anwendungen der maschinellen Sprachverarbeitung nicht mehr aus dem Alltag wegzudenken sind: maschinelle Übersetzung, Spracherkennung, etc. Die meisten Ansätze beschäftigen sich mit den Wörtern. Die Beziehungen und Zusammenhänge zwischen Buchstaben werden im Vektorraum dargestellt. In dieser Arbeit geht es darum zu versuchen phonologische Vektorraumpräsentationen von Buchstaben zu kreieren.

Das Ziel der Arbeit ist die Erstellung von verschiedenen Vektorenrepräsentationen von Buchstaben mit phonologischen Features und ihren Vergleich mit Zufallsvektoren bei der Transkription in SAMPA (ASCII-basiertes phonetisches Alphabet, mit welchem die Aussprache der Laute dargestellt wird. Dieses Alphabet wurde zwischen 1987 und 1989 entwickelt, um phonemische Transkriptionen der offiziellen Sprachen der damaligen Europäischen Gemeinschaft übermitteln und verarbeiten zu können.)

Weiterhin ist Tobias auf die Definition und Besonderheiten von Phonologie eingegangen. Phonologie beschreibt die Systematik der Laute innerhalb einer Sprache . Die kleinste Einheiten, die die Phonologie untersucht sind die Phonemen, die sich nach Artikulationsart, -ort und –Stimmhaftigkeit unterscheiden.

In dieser Arbeit wurde Word2Vec benutzt (ein Programm zur automatischen Vektorerstellung von Wörtern). Als Input kriegt sie einen Text mit vielen Wörtern und als Output werden Wortvektoren mit Distanzangaben erstellt . Die Verarbeitung wird durch neuranales Netzt durchgeführt, das aus Architektur und Lernalgorithmus besteht, wobei ähnliche Wörter ähnlihce Vektoren enthalten.

Es wurden insgesammt vier Experimente durchgeführt: Zwei Implementierungen für Vektorenerstellung plus Zufallsvektoren; Word2Vec Vektoren; Transkription in SAMPA.

Seine eigene Implementierung wird „Char-Vectors“ genannt. Er verwendet die bereits definierte phonolodische Features und die Unterkategorien. Der Vektor wird mit (0,0,0,0,0) initialisiert und danach berechnet. Immer wenn Eigenschaft auf Phonem zutrifft enthällt es den dementsprechende3n Vektor. Die Berechnung des Buchstabenvectors passiert mittels Durchschnitt aller entsprechenden Phonemvektoren.

Die zweite Implementierung heißt „One-not Vektoren“. Hier werden die gleiche Features und Unterkategorien wie im Char-Vectors verwendet. Allerdings wird hier eine binäre Klassifizierung durchgeführt. Die Vektorgröße wächst somit auf 22 Dimensionen.

Die Implementierung mit Zufallsvektoren wird „Randomized Vectors“ genannt. Es werden 2 Arten von Buchstabenvektoren generiert: 15 Dimensionen und 100 Dimensionen. Dafür wird das „random“-numpy-Moduls verwendet.

Im Falle vom Experiment Word2Vec Vektoren bestehen die Trainingsdaten aus Buchstaben anstatt aus ganzen Wörtern. Phonologisch ähnliche Buchstaben werden aufeinander trainiert.

Die beste Ergebnisse zeigt unerwartet random Experimente mit der 100-Version. Um diese Ergebnisse zu bestätigen, wird eine Fehlerquote mittels Levenshtein-Distanz berechnet .Es besagt, wie viele Korrekturen durchschnittlich gemacht werden mussten. Das bestätigt die bereits erhaltene Ergebnisse der quantitativen Auswertung

Vortrag: „Predicting New Domain Senses in English Medical Texts“ Jakob Sharab Betreuer : Fabienne Braune

Einige Wörter haben in verschiedenen Domänen eine andere Bedeutung. Z.B.: das Wort „administration“: allgemein: „Verwaltung“ und in Medizin: „Verabreichung“ (eines Medikamentes). Das führt zu Fehlern bei Statistical Machine Translation Systemen (SMT).

Deshalb Definition eines neues Task „Sense Spotting“. Das Ziel hier ist die Features zu finden, die Bedeutungsveränderung indizieren und danach mit Hilfe dieser Features Classifiers zu trainieren. Eines dieser Features ist das „Topic Model Feature“.

Der Sinn des „Topic Modeling“ ist in großen Textkorpora darin enthaltene Topics zu finden. Dafür werden mit Hilfe von Algorithmen die einzelne Wörter in den Dokumenten analysiert. Diese Methode hat als Vorteil, dass eine vorgehende Annotation der Daten nicht nötig ist.

Das Nutzen von „Text Modeling“ liegt daran, dass es das Organisieren von großen Textarchiven möglich macht.

Es werden für diese Bachelorarbeit die Klassifikatoren verwendet, die mit generativen Modellen arbeiten, nämlich Latent Dirichlet Allocation (LDA). Es ist ein generatives Wahrscheinlichkeits-Modell, das benutzt wird, um Dokumente in einzelne Topics zu untergliedern. Die Idee dahinter ist es, dass jedes Dokument aus einer zufälligen Mischung latenter Topics besteht. Jeides Topic seinerseits besteht aus einer Verteiliung über Wörter. Diese Modell basiert auf der Annahme der „bag-of-words“ (Reihenfolge von Wörtern ist vernachlässigbar, Wörter sind austauschbar). Somit auch wenn man sogar alle Wörter in einem Text vermischt, soll das grundlegendes Thema immer noch erkennbar. Weiterführend auch die Dokumente innerhalb eines Korpus sind austauschbar.

Generative Entstehung eines Dokumentes basiert auf folgenden Schritten: Festlegung der Anzahl von Wörtern, die ein Topic enthält; Festlegung der Mischung an Topics die in einem Dokument enthalten sind z.B. aus 60 % Topic „Medikamente“ + 30% Topic „Krankheiten“; Generierung der Wörter: Auswählen des Topics aus dem das Wort stammt, das Wort selbst mit Hilfe des Topics generieren.

Durch Umkehrung des generativen Prozesses werden Dokumente in Topics unterteilt.

Das Topic Model Feature ist das Feature das im Rahmen des „Sense Spotting“ Task definiert wurde. Das nimmt die Änderung der Häufigkeit eines Wortes innerhalb eines Topics, beim Wechsel in die neue Domäne, als Indikator für eine Bedeutungsveränderung.

Jakob verwendet die „Kosinus-Ähnlichkeit“, „relative Entropie“ und die „Ähnlichkeit aufgrund der Anzahl gleicher Wörter“ zum Messen der Ähnlichkeit zwischen den Topics. Diese Bachelorarbeit hat dann als Ziel, verschiedene Ähnlichkeitsmaße miteinander zu vergleichen:

1). Kosinus-Ähnlichkeit:
- Nimmt zwei Vektoren und berechnet den zwischen diesen eingeschlossenen Winkel.
- Gibt einen Wort zwischen 0 und 1.
- Je höher der Wert, desto ähnlicher sind zwei Vektoren

2). Relative Entropie
- Berechnet den Abstand zwischen zwei Wahrscheinlichkeitsverteilungen
- Je höher der Wert, desto weiter auseinander sind zwei Verteilungen.

3). Ähnlichkeit aufgrund der Anzahl gleicher Wörter:

- Idee: Je mehr gleiche Wörter unter den Top ? Wörtern zweier Topics sind, desto ähnlicher sind sie sich
- Problem: Die Zahl der ähnlichen Wörtern zwischen einem Topic aus der alten Domäne und denen der neuen Domäne war sich immer sehr ähnlich
- Lösung: Gewichtung der Wörter nach ihren Wahrscheinlichkeiten

In dieser Bachelorarbeit wurden zwei parallele Korpora (Englisch, Deutsch) benutzt. Davon Daten für die:
- medizinische Domäne (EMEA Korpus). Enthält Prüfberichte von Medikamenten von der Pharmabehörde „European Medicines Agency“ (Über 41.000 Dokumente)
- Nachrichten Domäne: General Korpus (verwendet im WMT Shared Task 2016) Besteht aus Daten mehrerer Korpora. News-Commentary: 272.807 Sätze Europarl-Korpus: 1.920.209 Sätze Common-Crawl: 2.399.123 Sätze

Das Experiment wurde folgend durchgeführt:
- die einzelne Dokumente wurden tokanisiert und die Stoppwörter wurden entfernt.
- die Dokumente wurden in 100 Topics mit Hilfe der LDA unterteilt (Verwendete Systeme: Vowpal Wabbit und Gensim)
- Auswählen eines Wortes, das hohe Wahrscheinlichkeit in einem Topic aus der alten Domäne hat und dessen Bedeutung sich verändert
- für die Kosinus-Ähnlichkeit und Entropie: Herausfiltern der Wahrscheinlichkeiten aller gleichen Wörter zwischen dem Topic aus der alten Domäne und allen anderen Topics aus der neuen Domäne. Für die Ähnlichkeit aufgrund der Anzahl gleicher Wörter - Überprüfen der Anzahl der gleichen Wörter unter den Top 2500 Wörtern .
- Anwenden der Formel für das Topic Model Feature für alle drei Ähnlichkeiten
- Berechnung von Werten auch für Wörter, die keine Bedeutungsveränderung erfahren haben
- Um einen Vergleichswert zu haben wurde hieraus der Mittelwert abgebildet

Jakob hat im Laufe seiner Arbeit auf einige Probleme gestoßen. Es war nicht einfach, viele Beispiele für Wörter zu finden, deren Bedeutung sich in der neuen Domäne verändert, da diese Wörter immer die Bedingung erfüllen müssten, in einem Topic eine hohe Wahrscheinlichkeit zu haben. Da es schwierig ist, ohne einen Klassifikator eine Decision Boundary zu finden, konnten die Ergebnisse nur quantitativ miteinander verglichen werden. Dabei haben die Relative Entropie und das Maß aufgrund der gleichen Wörter die besten Resultate erzielt.