Übung 7

Prof. Dr. Anette Frank
Tutorium: Thea Bartmann & Jakob Forstmann
Formale Semantik WiSe 2025/26
Frist: 7.1.2026, 23:59 Uhr

Ausgabe: 13.12.2025

1 Aufgabe: ELMo

In Übungsblatt 5 haben wir uns mit nicht-kontextualisierten Word2Vec Embeddings befasst. Obwohl sie mit der SkipGram bzw. BOW Methode trainiert werden, waren diese nicht in der Lage, zwischen verschiedenen Senses zu unterscheiden. In diesem Übungsblatt widmen wir uns den kontextualisierten Word Embeddings, hier in Form von ELMo. Diese Art von Embeddings erfordert einen Kontext und kann dabei zwischen Senses unterscheiden. Beispielsweise wäre der Vektor für *Bank* in *Ich ging zur Bank um Geld abzuheben* ein anderer als in *Auf der Bank war kein Platz mehr frei.*

Für dieses Blatt verwenden wir ELMo über das Python Package simple-elmo. Zusätzlich müssen Sie tensorflow <= 2.15.0 installieren. Dies ist nur mit Python 3.11 oder älter kompatibel. Dazu erstellen Sie am besten ein Virtual Environment. Sie finden dazu auch eine **Anleitung auf den Tutoriumsfolien**. Danach müssen Sie die options.json und die Gewichte für ELMo herunterladen und in einen beliebigen Ordner legen. Sie finden zu simple-elmo in Moodle ein minimal working example. Beim Ausführen tauchen diverse Warnungen auf, diese können ignoriert werden.

Nur als Erinnerung: Für die Projekte benötigt ihr einen CLuster Zugang. Wer diesen noch nicht hat kann eine E-Mail an die Gruppe Technik schreiben um sich freischalten zu lassen.

1.1 Ähnlichkeit von Synsets/Senses - 15 Punkte

Abgabe: Code (py), Output (txt)

In dieser Aufgabe wollen wir die Ähnlichkeit von Senses des Wortes *match* auf verschiedene Arten untersuchen. Einerseits anhand der Ähnlichkeit von Zentroiden, die wir

auf Basis aller uns zur Verfügung stehenden Sätze für einen Sense berechnen, und andererseits anhand von WordNet Pfaden. Im Folgenden finden Sie eine Auflistung aller WordNet-Senses des Wortes *match*, für die Ihnen Beispiele zur Verfügung gestellt werden:

- match.n.01 lighter consisting of a thin piece of wood or cardboard tipped with combustible chemical; ignites with friction
- match.n.02 a formal contest in which two or more persons or teams compete
- match.n.03 a burning piece of wood or cardboard
- match.n.04 an exact duplicate
- match.n.05 the score needed to win a match
- match.v.01 be compatible, similar or consistent; coincide in their characteristics
- match.v.02 provide funds complementary to
- match.v.03 bring two objects, ideas, or people together
- match.v.05 make correspond or harmonize
- match.v.07 give or join in marriage

Auf Moodle finden Sie in dem Ordner *Materialen für Übungsblatt 7* eine Datei (`match_sense_examples.txt`), die mehrere tokenisierte Sätze pro Synset/Sense enthält.

Eine Beispielzeile aus dieser Datei:
`match.n.0_4 4 curt snuffed out the match`
`match.n.01_4` ist die ID; `match.n.01` ist der Sense/das Synset gemäß WordNet (Teil der ID); `4` ist die Position von *match* im Satz; zuletzt folgt der bereits tokenisierte Satz. Alle Elemente sind durch 4 Leerzeichen getrennt.

Zunächst gilt es, für **jeden** Satz ein kontextualisiertes ELMo-basiertes Word Embedding für *match* zu gewinnen. Dafür betten wir jeden Satz in ELMo ein und erhalten den Wortvektor für *match* im jeweiligen Satz mit Hilfe der ebenfalls angegebenen Position. Sobald Sie einen kontextualisierten ELMo Vektor für *match* aus jedem Satz erhalten haben, berechnen Sie auf Basis aller zu einem Sense gehörigen Vektoren einen Zentroiden für diesen Sense.

Berechnen Sie dann die Kosinusähnlichkeit zwischen den Zentroiden und ordnen Sie die Paare absteigend von höchster Kosinusähnlichkeit zu niedrigster Kosinusähnlichkeit. (fiktives Bsp.: CosSim bank.n.01 und bank.n.02 = 0.7, CosSim match.n.01 und match.n.03 = 0.3, etc.). Vermeiden Sie dabei, Paare doppelt zu erwähnen (die Kosinusähnlichkeit von match.n.01 und match.n.03 ist die gleiche wie von match.n.03 und match.n.01) ebenso wie triviale Berechnungen (CosSim zwischen match.n.01 und match.n.01).

Ermitteln Sie nun alternativ ebenfalls für jedes Paar von Synsets die Ähnlichkeit mit Hilfe der WordNet `path_similarity`.

Kosinusähnlichkeit und WordNet path similarity können beide Werte zwischen 0 und 1 annehmen. Die Ähnlichkeit von Senses/Synsets ist hoch, wenn die Kosinusähnlichkeit (der Zentroiden) hoch ist und/oder wenn die WordNet path similarity hoch ist.

Fügen Sie die Information der path similarity jedem Paar von Synsets in Ihrem Ranking bei. Die Paare bleiben dabei weiterhin nach Kosinusähnlichkeit der Zentroiden geordnet.

Beantworten Sie abschließend die folgenden Fragen:

1. Welcher Sense weist durchweg die geringste Kosinusähnlichkeit zu allen anderen auf und was könnten Gründe dafür sein?

2. Verhalten sich Kosinusähnlichkeit und WordNet path similarity für die Beispiele gleich? Geht eine hohe Kosinusähnlichkeit mit einer hohen path similarity einher? Wählen und nennen Sie ein Positiv- und ein Negativbeispiel.

1.2 Plotting - 10 Punkte

Abgabe: Code (py), Plot (jpg, png, pdf)

Wie schon bei Übungsblatt 5 bietet es sich auch hier an, die Vektoren durch einen Plot zu visualisieren und so eine bessere Vorstellung für die zuvor erhaltenen Werte zu bekommen.

Nutzen Sie wie gewohnt **PCA**, um zunächst die 1024-dimensionalen Vektoren auf 2 Dimensionen zu reduzieren. Übergeben Sie dabei alle Vektoren auf einmal. Vergessen Sie auch nicht die Zentroiden.

Visualisieren Sie dann alle Vektoren in einem Plot, wobei Vektoren, die zum gleichen Sense gehören, die gleiche Farbe erhalten sollen. Beschriften Sie außerdem den Zentroiden mit dem Namen des Synsets (z.B. *match.n.01*).

Welcher Sense/welches Synset sticht besonders heraus und grenzt sich deutlich von allen anderen ab? Gibt es Instanzen, die stark von ihrem jeweiligen Zentroiden abweichen? Wenn ja, nennen Sie die entsprechenden Synsets.