Woche 12: Wie war das nochmal? – Überblick und Ausblick

Skript

Erarbeitet von
Dr. Maike Mayer

Lernziele

- Erinnern wichtiger Inhalte zu der Textvorverarbeitung (wie Stemming oder Stoppwortentfernung)
- Nachvollziehen der Zusammenhänge zwischen den Inhalten

Inhalt

Einstieg

Sprache ist für uns eine wichtige Kommunikationsform und wir sind sehr gut darin, Informationen, die über Sprache vermittelt werden, in Echtzeit zu verarbeiten und zu verstehen. Gut, es kann trotzdem mal zu Missverständnissen oder dem berühmt-berüchtigten „aneinander Vorbeireden" kommen, aber ganz grundlegend bekommen wir das schon ziemlich gut hin. Wir sind beispielsweise auch in der Lage, den Kontext von

Gesprochenem zu nutzen, um Emotionen und Stimmungen zu erfassen oder auch, um uns die Bedeutung von uns unbekannten Wörtern herzuleiten.

Einblendung Illustrationen

Wir denken im Alltag aber oft gar nicht darüber nach, dass wir bei Sprache Informationen verarbeiten oder wie wir das machen. Manchmal wird man sich der Komplexität dahinter vielleicht bewusst, wenn man beispielsweise einem Kind eine Wortbedeutung erklären soll – gerne von ganz basalen Wörtern. Da wird es besonders kompliziert.

Nun stell dir aber mal vor, du willst nicht einem Kind, sondern einem Computer beibringen, Sprache zu verstehen bzw. mit Sprache umzugehen. Genau das ist das Ziel der Sprachverarbeitung oder auch Natural Language Processing, einem Feld, das sich mit der Verarbeitung von Sprache beschäftigt. Ein zentrales Konzept dafür ist die Textrepräsentation, also wie wir Texte in eine Form bringen, die Maschinen verstehen können. Und genau damit beschäftigen wir uns in dieser Woche näher …

Einblendung Illustrationen/Stichwort

Dem Computer Texte beibringen

In Woche 2 dieses Kurses haben wir schon mal darüber gesprochen, wie wir es schaffen, dass Maschinen überhaupt mit Texten umgehen können. Texte liegen im Computer als Zeichenketten – also als sogenannte Strings – vor, die als Zahlen kodiert gespeichert werden. Ein Zeichen wird dabei fest einer Zahl zugeordnet. Das ist die Basis für unsere Textvorverarbeitung, also dafür, den Text erstmal in eine Form zu bringen, sodass der Computer damit arbeiten kann.

Einblendung Illustrationen (ASCII-Kodierung für „A“: 65)/Stichwort

Die Vorverarbeitung von Texten, das weißt du bereits, hängt stark davon ab, für welche spätere Anwendung das Modell benutzt werden sollen und auch die Sprache ist für die Vorverarbeitungsschritte relevant. Da Woche 2 allerdings schon eine ganze Weile zurückliegt, wiederholen wir nochmal kurz die Vorverarbeitungsschritte, die du schon kennst.

Einblendung Illustration

Da wäre die Tokenisierung, denn zunächst müssen wir den Text in Einzelteile – auch Tokens genannt – unterteilen. Hier gibt es verschiedene Möglichkeiten, aber oft wird der Text erst in Einzelsätze und dann in Einzelwörter aufgeteilt. Dann folgt oft eine Normalisierung. Hierbei werden häufig alle Tokens durchgegangen und Großbuchstaben in Kleinbuchstaben umgewandelt, denn „aber“ mit kleinem „a“ und „Aber“ mit großem „A“ sind für den

Computer verschiedene Wörter. Gerade hierbei ist die Sprache des Textes relevant, denn auf Deutsch kann Groß- und Kleinschreibung durchaus die Wortbedeutung beeinflussen, was sprachspezifische Vorverarbeitungsschritte erforderlich macht.

Einblendung Illustrationen/Stichwörter

Dann wäre da noch die Stoppwortentfernung. Dabei werden Stoppwörter – also Wörter, die für die Bedeutung des Textes irrelevant sind, wie beispielsweise die Artikel „der“, „die“ oder „das“ – aus dem Text gestrichen. Nun können wir die Wörter noch in ihre Grundform umwandeln, also „ging“ beispielsweise zu „gehen“ machen – das nennt man Lemmatisierung – oder auf ihren Stamm zurückführen, also auf „geh“ für „gehen“, – das nennt man dann Stemming.

Einblendung Illustrationen/Stichwörter

Gut, jetzt haben wir den Text so vorverarbeitet, dass er in eine vernünftige Repräsentation gebracht werden kann. Aber wie kriegen wir die Maschine jetzt dazu, den Text zu verarbeiten oder gar zu verstehen?

Jetzt aber

Machine-Learning-Algorithmen benötigen Eingaben in numerischer Form, um damit arbeiten zu können. Vielleicht denkst du jetzt: Aber die Zeichen liegen doch schon in numerischer Form vor. Ja, aber wenn die Maschine den Inhalt des Textes verstehen soll, müssen wir den Text in eine entsprechende numerische Form bringen und nicht die einzelnen Zeichen der Wörter. Das Umwandeln eines Textes in numerische Form bezeichnet man als „feature extraction“.

Einblendung Illustrationen/Stichwort

Zunächst kann man beispielsweise ein Wörterbuch, auch Vokabular genannt, erstellen. Jedes Wort, das in unserem Text vorkommt, stellt einen Eintrag dar. Näheres zu den Wörterbüchern erfährst du in dieser Woche, genau wie zu dem sogenannten Worttaschenmodell oder auch Bag-of-Words-Modell, bei dem ein Dokument als Vektor repräsentiert wird. Dabei wird für jedes Wort in einem Wörterbuch angegeben, wie oft es in dem Dokument vorkommt oder alternativ auch, ob ein Wort des Wörterbuchs in einem Text enthalten ist oder nicht. Die Vorkommenshäufigkeit kann man dann beispielsweise auch als Maß dafür nutzen, wie relevant das Wort für den Text ist. Der Nachteil an einem Worttaschenmodell ist allerdings, dass der Kontext der Wörter verloren geht. Beispielsweise werden inhaltlich gleiche Aussagen, die aber unterschiedlich formuliert werden, nicht als ähnlich erkannt.

Einblendung Illustrationen/Stichwörter

Will man den Kontext einbeziehen, muss man die Einbettung oder auch Embedding eines Worts berechnen, also quasi die Lage des Worts im Vektorraum. Ähnliche Wörter liegen hierbei nah beieinander. Wir stellen dir in dieser Woche mit BERT ein Sprachmodell vor, das den Kontext von Wörtern erfassen kann und somit eine höhere Genauigkeit bei der Sprachverarbeitung erreicht. BERT steht für Bidirectional Encoder Representations from Transformers. BERT ist quasi eine Art neuronales Netzwerk, das die Bedeutung von Wörtern und Sätzen erfassen kann. Wir stellen dir BERT aber nicht nur vor, sondern geben dir auch einen ganz kleinen und vereinfachten Einblick in die Programmierung dahinter.

Einblendung Stichwörter

Quelle [1]

Auf geht's!

In dieser Woche dreht sich also alles darum, wie du es schaffst, dass dein Text von einem Machine-Learning-Modell verarbeitet und verstanden werden kann. Du lernst Wörterbücher, das Worttaschenmodell und Worteinbettungen kennen. Außerdem erläutern wir dir den Unterschied zwischen einfacher und fortgeschrittener Text-Repräsentation und stellen dir BERT vor. In dieses Sprachmodell bekommst du auch im Rahmen der Programmierung einen kleinen Einblick. Abschließend haben wir dann noch ein Anwendungsbeispiel für dich zum Thema Sprach- bzw. Textverarbeitung vorbereitet. Du lernst ein Projekt kennen, das sich mit Argument Mining auf Twitter beschäftigt. Dabei sollen mit Hilfe von Sprachverarbeitung automatisch Argumente und Argumentationsstrukturen in Texten auf Twitter erkannt werden, um so zwischen Argumenten und Nicht-Argumenten zu unterscheiden und die Qualität der Diskussionen und Debatten zu verbessern.

Einblendung Videotitel

Also sprechen wir in dieser Woche im übertragenen Sinne mit Computern, statt nur über sie. Viel Spaß dabei und in Woche 12!

Quellen

Quelle [1] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Disclaimer

Transkript zu dem Video „Woche 12: Wie war das nochmal? – Überblick und Ausblick“, Dr. Maike Mayer.