DHBW
Duale Hochschule
Baden-Württemberg
Karlsruhe

Fakultät Wirtschaft

Studiengang Wirtschaftsinformatik

Test

2. Projektarbeit

Im Rahmen der Prüfung zum Bachelor of Science (B. Sc.)

Sperrvermerk

31. August 2020

VerfasserIn:	Test
Kurs:	WWI22B5
Dualer Partner:	Musterfrau AG, Karlsruhe
Betreuer der Ausbildungsfirma:	Leonie Musterfrau
Wissenschaftlicher BetreuerIn:	Prof. Dr. Tina Mustermann
Abgabedatum:	31. August 2020

Selbstständigkeitserklärung

Ich versichere hiermit, dass ich die vorliegende 2. Projektarbeitmit dem Thema:

Test

selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Ich versichere zudem, dass die eingereichte elektronische Fassung mit der gedruckten Fassung übereinstimmt.

Karlsruhe, 31. August 2020, _____

Test

Sperrvermerk

Der Inhalt dieser Arbeit darf weder als Ganzes noch in Auszügen Personen außerhalb des Prüfungsprozesses und des Evaluationsverfahrens zugänglich gemacht werden, sofern keine anders lautende Genehmigung der Dualen Partners vorliegt.

Kurzfassung

Hier beginnt die Kurzfassung ihrer wissenschaftlichen Arbeit...

Inhaltsverzeichnis

Abkürzungsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

1.1 Kontext und Relevanz des Themas

Keine Entwicklung der Welt der IT ist aktuell so viel besprochen wie die der künstlichen Intelligenz. Speziell durch den Aufstieg von generativer KI hat sich das Thema zu einer geradezu gesamtgesellschaftlich relevanten Entwicklung herangebildet. Maßgeblich angestoßen durch die Veröffentlichung von OpenAI's GPT-3 Modell, welches in der Lage ist, Texte zu generieren, die von menschlichen Texten nur noch schwer zu unterscheiden sind, hat sich die öffentliche Aufmerksamkeit auf die Möglichkeiten von generativer KI gerichtet. Die Konzepte und Technologien, die hinter diesen Entwicklungen stehen, sind dabei nicht unbedingt neu, eine breitere Verfügbarkeit von Rechenleistung und Trainingsdaten haben jedoch den entscheidenen Anstoß für die neue Leistungsfähigkeit dieser Modelle gegeben. KI ist also das das Thema der Stunde und als strategisch relevantes Thema für Unternehmen und Organisationen nicht mehr wegzudenken.

1.2 Ziel der Arbeit

Die vorliegende Arbeit beschäftigt sich mit einem konkreten Anwendungsfall von KI in der Praxis. Untersucht wird ein Beispiel, in dem eine semantische Datenbanksuche auf einem Materialstammdatensatz durchgeführt wird. Der Mehrwert dieses Systems liegt dabei in der Möglichkeit für Anwender, die Datenbank auf natürlichsprachliche Weise zu durchsuchen, ohne dabei auf die spezifischen Suchbegriffe und -syntaxen achten zu müssen, die in traditionellen Datenbanksystemen notwendig sind und gleichzeit von einem gewissen semantischen Verständnis des Suchsystems profitieren zu können. Die Technik aus dem Feld der KI, die für dieses System zum Tragen kommt sind sogenannte *word embeddings*, die es ermöglichen, Worte in einem Vektorraum abzubilden und so semantische Ähnlichkeiten zwischen Wörtern zu berechnen. Dieses Konzept wird in der Arbeit genauer erläutert die Effektivität verschiedener Techniken zur Erstellung von embeddings im konkreten Anwendungs-

fall beleuchtet. Das Ziel der Arbeit ist es, die Technik im Anwedungsfall zu erläutern, verschiedene Methoden und Modelle zu beleuchten und eine datengestützte Entscheidungsgrundlage für die Bewertung von word embeddings in semantischen Suchsystemen zu schaffen.

2 Theoretischer Hintergrund

2.1 Maschinelles Lernen

Maschinelles Lernen beschreibt das Konzept, auf Basis von einer großen Menge von Daten Algorithmen zu approximieren, die auf anderem Wege nicht erschlossen werden können. Man nehme beispielsweise die klassische Aufgabe, ein Programm zu schreiben, das in der Lage ist, Bilder von Hunden und Katzen zu unterscheiden. Wenn wir als Menschen uns dieser Aufgabe stellen, müssen wir nicht lange überlegen, wir lösen sie intuitiv. Wenn wir uns aber fragen, nach welchen Regeln wir diese Entscheidung treffen, wird es schon schwieriger. Wir könnten uns auf die Form der Ohren, die Farbe des Fells oder die Größe des Tieres konzentrieren. Aber wie genau wir Regelmäßigkeiten definieren, ist nicht so einfach. Maschinelles Lernen verfolgt den Ansatz, genau solche Regeln nicht mehr fest zu definieren, sondern sie anhand von einer großen Menge von Daten zu lernen.

2.2 Neuronale Netze

Ein Mittel der Wahl um das Konzept des maschinellen Lernens umzusetzen, sind sogenannte künstliche Neuronale Netze. Neuronale Netze, lose inspiriert von der Struktur des menschlichen Gehirns, bestehen aus einer Vielzahl von einfacher Einheiten, sogenannte Knoten, die in Schichten angeordnet sind und über unterschiedlich gewichtete Verbindungen verknüpft sind. Diese Struktur ermöglicht es, komplexe statistische Zusammenhänge in einem Datensatz zu modellieren, indem für einen gegebenen Datensatz mithilfe von Techniken des maschinellen Lernens die Parameter, also beispielsweise die Gewichte der Verbindungen des Netzes, so angepasst werden, dass sie die gegeben Daten möglichst genau abbilden. Wurde dieser Prozess erfolgreich durchlaufen, so kann das Modell im Anschluss dazu genutzt werden, Aussagen über Daten, die es im Lernprozess noch nie gesehen hat, zu treffen oder Vorhersagen abzugeben. Das interessante an diesem Ansatz ist es, dass durch diesen Ansatz, gerade bei großen neuronalen Netzen auch nicht triviale, subtile Muster im Daten-

satz erkannt werden können und so, wie oben bereits angedeutet, Approximationen für Probleme getroffen werden können, die formal nur schwer beschrieben werden können.

2.3 Word Embeddings

Eine weitere, für diese Arbeit relevante Entwicklung der jüngeren Forschung sind die Fortschritte der Computerlinguistik. Ein Kernproblem dieses Feldes ist die Forschung an der Repräsentationen von Sprache. Hierbei geht es nicht einfach darum, einzelne Wörter in ihrer Schriftform zu speichern, sondern vielmehr den Wort*sinn* festzuhalten. Man betrachte zum Beispiel die Wörter *Couch* und *Sofa*, die in Schriftform, mit Ausnahme des zweiten Buchstabens vollkommen unterschiedlich sind, in ihrer Bedeutung aber nahezu Synonym verwendet werden. Weiterhin möchten wir Aussagen über die Beziehung von Wörtern treffen können. *Heiß* und *kalt* haben in ihrer Wortbedeutung einen klaren Zusammenhang (Es handelt sich um Gegensätze), den wir eventuell darstellen möchten, genauso wie *Replika* und *Fälschung* im Grunde dasselbe meinen, aber einen klaren Unterschied in ihrer Konnotation aufweisen. Eine Form der semantisch reichen Repräsentationen zu finden, die diesen Anforderungen genügt ist nicht trivial, es handelt sich aber wieder um ein solches Problem, das, wie oben beschrieben, intuitiv einfach zu lösen, formal jedoch schwer zu beschreiben ist. Und genau wie oben beschrieben, können die Techniken aus dem Feld des maschinellen Lernens auf dieses Problem angewandt werden, um es zu lösen.

Die Grundlage für die nun folgenden Überlegungen bildet die 1950 erstmals formulierte Verteilungshypothese der Linguistik. Im Grunde besagt sie, dass Wörter, die in ähnlichen Kontexten auftauchen, eine ähnliche Bedeutung haben. Wenn die Wörter *Pizza* und *Burger* beispielsweise beide häufig im Zusammenhang mit den Wörtern *Essen* und *geniessen* auftauchen, kann daraus geschlossen werden, dass sie ihr Wortsinn eine Ähnlichkeit hat, in diesem Fall, dass es sich bei beiden Wörtern um Essen handelt.

	Essen	italienisch	Auto
Pizza	150	122	11
Burger	136	3	13
Porsche	0	6	350
Ferrari	1	199	475

Tabelle 1: Wort-Wort-Matrix auf Basis des Wikipedia Corpus und ausgewählten Worten

Auf Basis dieser Erkenntnis kann eine erste simple Repräsentationen des Wortsinns gefunden werden. Gegeben sei ein Corpus C auf Basis dessen wir einen Wortsinn für jedes Wort im Vokabular V des Corpus finden wollen. Auf Basis der Verteilungshypothese kann nun eine Wort-Wort-Matrix aufgestellt werden, die abbildet, wie häufig Worte im Kontext anderer Worte auftauchen. Dafür muss ein Kontext definiert werden, häufig ist dieser Kontext ein Bereich um das Wort, kann aber auch beliebig definiert werden, beispielsweise als eine Menge Dokumente im Corpus. Als Ergebnis erhält man eine Matrix mit der Dimension $|V| \times |V|$, beziehungsweise einen $|V|$-dimensionalen Spaltenvektor für jedes Wort.

Die so gewonnen Vektoren haben bereits einige der gewünschte Eigenschaften um den Wortsinn zu encodieren. Die in Tabelle 2.3 dargestellte Wort-Wort-Matrix basiert auf dem Corpus der englischsprachigen Wikipedia und illustriert eine Eigenschaft, die auch noch bei der Reduktion auf wenige Dimensionen sichtbar wird: Wörter, die sich ähnlich sind, tauchen in ähnlichen Kontexten auf. Die Vektoren für Automarken tauchen beide ähnlich häufig im Kontext des Wortes Auto auf, genauso wie Essbares ähnlich häufig im Kontext von dem Wort Essen auftauchen. Auch eine 19xx von den Linguisten ursprünglich als IQ-Test entwickelte Methode lässt sich hier nachstellen. Die Idee dieses Testes ist es Fragen nach folgendem Schema zu stellen. *Deutschland gehört zu Berlin. Was gehört zu Paris?* Auch bei diesem sehr simplifizierten Beispiel lässt sich ein sogenanntes Sinn-Parallelogram aufstellen. Anhand der in Abbildung 1 getroffenen Visualisierung des Beispiels lässt sich leicht erkennen, dass dieses Problem durch einfach Vektorrechnung lösen lässt. *Pizza gehört zu Burger. Was gehört zu Porsche?*. Die Antwort innerhalb dieses Beispiels

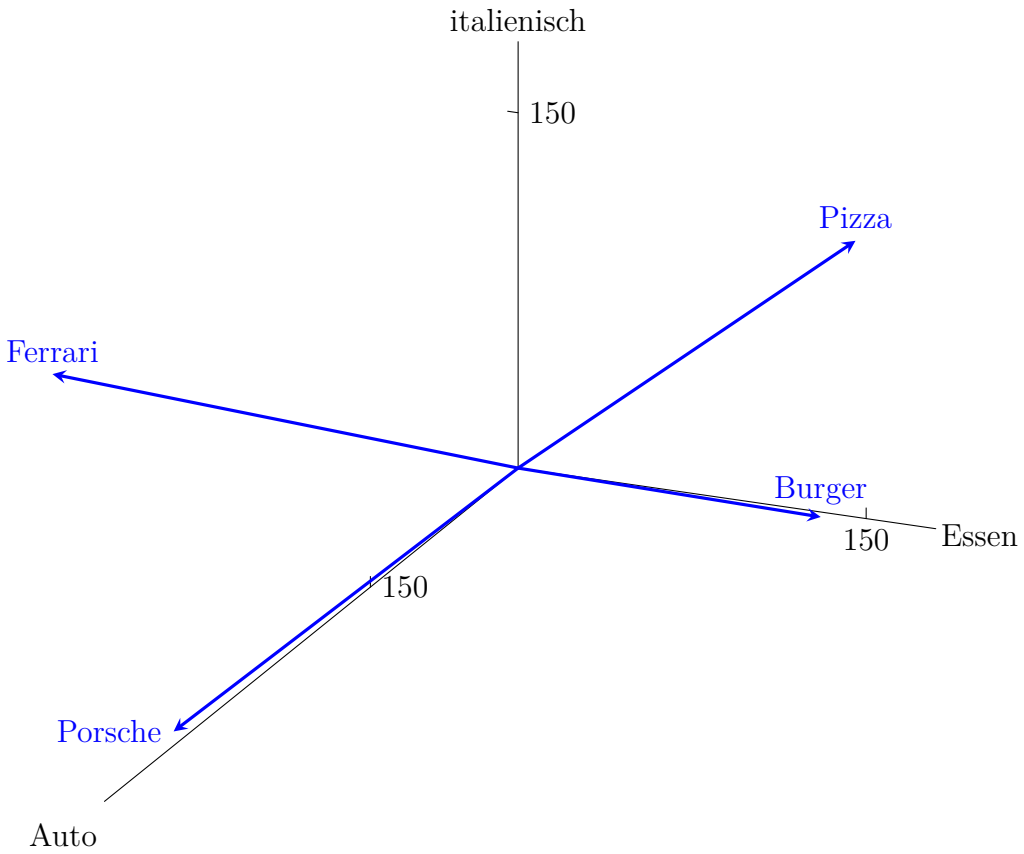

Abbildung 1: Visualisierung der in Tabelle 2.3 berechneten Vektoren in dreidimsionalem Raum

ist Ferrari und lässt sich bestimmen, indem der Vektor für Pizza von Burger subtrahiert wird und das Ergebnis dieser Operation auf den Vektor für Porsche aufaddiert wird. Das Ergebnis ist ein Vektor der in die Nähe von Ferrari zeigt. Dieses Beispiel ist selbstverständlich enorm simplifiziert, es lässt sich aber genau diese Encodierung von Wortsinn auch bei größeren Vokabularien feststellen.

Vorgestellt wurde die hiermit die einfachste Form von statischen Embeddings, mittlerweile gibt es eine Vielzahl von Techniken um dieses stumpfe Zählen von Worten mit verschiedenen Verfahren zu optimieren und schlussendlich bessere und sinnhaltigere Ergebnisse zu erzielen. Ein Problem dieser Methode ist beispielsweise, dass die so gewonnen Vektoren eine sehr hohe Dimensionalität haben ($|V|$), dafür aber größtenteils leer sind, also 0 enthalten. Die Empirik hat gezeigt, dass sich wesentlich bessere Ergebnisse erzielen lassen, wenn Wörter als Vektoren mit niedrigerer Dimensionalität dargestellt werden, die Intuition hinter dieser Erkenntnis ist es, dass dadurch eine gewisse "Abstraktion"des Wortsinnes stattfindet und so konzeptuelle Zusammenhänge besser abgebildet werden können. Eine Methode, diese Reduktion der Dimensionalität mithilfe der oben besprochenen Konzepte des maschinellen Lernens zu erreichen, soll im Folgenden dargestellt werden.

3 Ist- und Problemanalyse

4 Optimierung des Systems

5 Fazit

Quellenverzeichnis

Anhang

1. Digitale Version der Arbeit
2. Interviews
    2.1. Expertmann 2018